Matias Laukkanen

The association between the size of the largest disclosed VC round and IPO success

EEA/EFTA country evidence from Logit, Random Forest, and XGBoost models

School of Accounting and Finance
Master's Thesis in Finance
Master's Degree Programme in Finance

Vaasa 2026

UNIVERSITY OF VAASA
School of Accounting and Finance
Author: Matias Laukkanen
Title of the thesis: The association between the size of the largest disclosed VC round and IPO success: EEA/EFTA country evidence from Logit, Random Forest, and XGBoost models
Degree: Master of Science in Economics and Business Administration
Degree Programme: Master's Degree Programme in Finance
Supervisor: Niranjan Sapkota
Year: 2026 **Pages:** 90

ABSTRACT:

Recent trends in the European venture capital (VC) industry indicate that private companies are shifting toward raising increasingly larger VC rounds. This thesis examines the association between the size of the largest disclosed VC financing round and initial public offering (IPO) success in the European Economic Area (EEA) and European Free Trade Association (EFTA) countries during 2000–2023. Primarily, the thesis studies whether the association exists, and secondarily, its direction. The dataset includes 231 VC-backed IPO companies. IPO success is measured by *Scaled IPO proceeds* and *IPO underpricing* variables.

The methodology includes traditional logistic regression analysis, while two machine learning algorithms, random forest and XGBoost, are applied to check the robustness of the logit results. With machine learning, the association between the thesis's VC round size variable and IPO success is studied by focusing on variable importance and partial dependence plots.

This thesis seeks to fill a small part of an apparent gap in the academic European VC-backed IPO performance literature by introducing a perspective on the VC round size. To the best of the author's knowledge, no other academic article has studied the association between the VC round size and IPO success in Europe like this thesis. In this academic gap, the traditional logit model findings are enhanced by employing modern machine learning algorithms.

The results of this thesis indicate no statistically significant conditional association between the size of the largest disclosed VC round and IPO success when using traditional logistic regression. However, the results from the random forest and XGBoost machine learning models consistently show that the size of the largest VC round is a conditionally important variable, as measured by permutation importance, for the thesis's models predicting IPO success. Additionally, partial dependence plots show that the size of the largest disclosed VC equity round has a negative conditional association with *IPO underpricing* by decreasing its average predicted probability. A positive conditional association with *Scaled IPO proceeds* is not consistently supported by the evidence.

KEYWORDS: venture capital, listing (stock market), regression analysis, machine learning.

VAASAN YLIOPISTO
Laskentatoimen ja rahoituksen akateeminen yksikkö
Tekijä: Matias Laukkanen
Tutkielman nimi: The association between the size of the largest disclosed VC round and IPO success: EEA/EFTA country evidence from Logit, Random Forest, and XGBoost models
Tutkinto: Kauppatieteiden maisteri
Oppiaine: Master's Degree Programme in Finance
Työn ohjaaja: Niranjan Sapkota
Valmistumisvuosi: 2026 **Sivumäärä:** 90

TIIVISTELMÄ:

Viimeaikaiset trendit Euroopan varhaisen vaiheen pääomasijoittamisen (VC) toimialalla osoittavat, että yksityiset yritykset nostavat yhä suurempia VC-rahoituskierroksia. Tässä tutkielmassa selvitetään suurimman julkistetun VC-rahoituskierroksen koon ja listautumisantimenestyksen välistä yhteyttä Euroopan talousalueella (ETA) ja Euroopan vapaakauppaliiton (EFTA) jäsenvaltioissa vuosina 2000–2023. Tutkielmassa selvitetään ensisijaisesti onko edellä mainittua yhteyttä olemassa, ja toissijaisesti sen tarkempaa luonnetta. Tutkielman data-aineisto koostuu 231 listautujayrityksestä, jotka ovat saaneet VC-rahoitusta ennen pörssiin listautumista, ja joita VC-sijoittajat ovat tukeneet listautumisannissa. Listautumisantimenestystä mitataan kahdella muuttujalla: *Scaled IPO proceeds* ja *IPO underpricing*.

Tutkimusmenetelmänä käytetään perinteistä logistista regressioanalyysia, josta saatavien tulosten kestävyyttä ja luotettavuutta arvioidaan myös kahdella koneoppimisen algoritmilla. Tutkielmassa käytetään random forest ja XGBoost-algoritmeja. VC-rahoituskierroksen koon ja listautumisantimenestymisen välistä yhteyttä tutkitaan koneoppimisen avulla keskittymällä siihen, kuinka tärkeä VC-rahoituskierrosmuuttuja on tutkielman malleissa ja tulkitsemalla osittaisriippuvuuskuvaajia.

VC-sijoittajien tukemien eurooppalaisten listautumisantien suorituskykyä käsittelevässä akateemisessa kirjallisuudessa on VC-rahoituskierrosten kokoon keskittyvässä tutkimuksessa aukko, jota tällä tutkielmalla pyritään täyttämään. Tutkielman tekijän parhaan ymmärryksen mukaan yksikään akateeminen artikkeli ei ole keskittynyt VC-rahoituskierrosten koon ja listautumisantimenestyksen väliseen yhteyteen Euroopassa kuten tämä tutkielma. Tutkielmassa parannetaan perinteisten logit-mallien tuloksia edellä mainitussa akateemisen kirjallisuuden aukossa käyttämällä moderneja koneoppimisen algoritmeja.

Tässä tutkielmassa esitetyillä tuloksilla ei kyetä osoittamaan tilastollisesti merkittävää ehdollista yhteyttä suurimman julkistetun VC-rahoituskierroksen koon ja listautumisantimenestymisen välillä perinteisen logistisen regressioanalyysin avulla. Random forest ja XGBoost-algoritmit kuitenkin osoittavat johdonmukaisesti, että VC-rahoituskierroksen kokoa kuvaava muuttuja on tärkeä listautumisantimenestyksen ennustamisessa. Tämän lisäksi osittaisriippuvuuskuvaajat osoittavat, että suurimman julkistetun VC-rahoituskierroksen koolla on ehdollinen negatiivinen yhteys *IPO underpricing* -muuttujaan pienentäen keskimääräistä todennäköisyyttä sen ennustamiselle. Tulokset eivät johdonmukaisesti osoita positiivista ehdollista yhteyttä *Scaled IPO proceeds* -muuttujaan.

AVAINSANAT: venture capital, listing (stock market), regression analysis, machine learning.

Contents

1 Introduction 8
1.1 VC industry developments during 2013–2023 9
1.2 Purpose and contribution of the thesis 11
1.3 Hypotheses of the thesis 12
1.4 Structure of the thesis 14
2 Venture Capital (VC) 15
2.1 Overview of VC 15
2.2 VC round stages 16
3 Initial Public Offering (IPO) 18
3.1 Overview of IPOs 18
3.2 IPO process 18
3.2.1 Documentation phase 19
3.2.2 Marketing phase 20
4 Theoretical foundation and the links to the VC round size 22
4.1 Information asymmetry 22
4.2 Signaling theory 22
4.3 VC certification and monitoring 23
4.4 Staged financing 24
5 Literature review 26
5.1 IPO underpricing 26
5.1.1 Hot and cold IPO markets and evidence from Europe 28
5.2 VC-backed IPO performance in the U.S. 29
5.3 VC-backed IPO performance in Europe 30
5.4 VC syndication in U.S. and European IPOs 32
6 Data 33
6.1 Dataset preparation 33
6.1.1 IPO data 33
6.1.2 VC transaction data 34

6.1.3 Exchange rate data 35
6.2 Dataset analysis 35
6.3 Variables in the thesis 37
6.3.1 Descriptive statistics (full sample) 41
6.3.2 Descriptive statistics (Scaled IPO proceeds) 42
6.3.3 Descriptive statistics (IPO underpricing) 44
7 Methodology 46
7.1 Model variables 46
7.2 Logistic regression (logit) models 48
7.2.1 Logit (Scaled IPO proceeds) 49
7.2.2 Logit (IPO underpricing) 49
7.3 Machine learning algorithms 50
7.3.1 Random Forest (RF) 50
7.3.2 eXtreme Gradient Boosting (XGBoost) 51
8 Results and discussions 52
8.1 Logit model coefficients (Scaled IPO proceeds) 52
8.2 Logit model coefficients (IPO underpricing) 54
8.3 Random Forest model performance 57
8.4 XGBoost model performance 59
8.5 Random Forest and XGBoost ROC curves (Scaled IPO proceeds) 61
8.6 Random Forest and XGBoost ROC curves (IPO underpricing) 62
8.7 Random Forest models' variable importance and PDPs 64
8.7.1 Random Forest PDPs 65
8.8 XGBoost models' variable importance and PDPs 66
8.8.1 XGBoost PDPs 67
8.9 Research question reformulation and 5-fold cross-validation tests 68
9 Conclusions 71
References 73
Appendices 79

Appendix 1. Variable descriptions 79
Appendix 2. Pearson and point-biserial correlation matrix 80
Appendix 3. Performance metric descriptions 81
Appendix 4. Logit confusion matrix (Scaled IPO proceeds) 82
Appendix 5. Logit confusion matrix (IPO underpricing) 82
Appendix 6. Random Forest confusion matrices 83
Appendix 7. XGBoost confusion matrices 84
Appendix 8. Random Forest variable importance 85
Appendix 9. XGBoost variable importance 86
Appendix 10. PDPs (Random Forest, Scaled IPO proceeds) 87
Appendix 11. PDPs (Random Forest, IPO underpricing) 88
Appendix 12. PDPs (XGBoost, Scaled IPO proceeds) 89
Appendix 13. PDPs (XGBoost, IPO underpricing) 90

Figures

Figure 1. Total VC deal value and VC deal count in Europe during 2013–2023. 9
Figure 2. Share of total VC deal value by size in Europe during 2013–2023. 10
Figure 3. Total VC exit and VC-backed IPO count in Europe during 2013–2023. 11
Figure 4. Number of VC-backed IPOs in the EEA and EFTA countries during 2000–2023. 36
Figure 5. Distribution of TRBC economic sectors of the VC-backed IPO companies. 37
Figure 6. ROC curves for the Scaled IPO proceeds random forest and XGBoost models. 61
Figure 7. ROC curves for the IPO underpricing random forest and XGBoost models. 63
Figure 8. Variable importance plot for the Scaled IPO proceeds random forest models. 64
Figure 9. Variable importance plot for the IPO underpricing random forest models. 65
Figure 10. Variable importance plot for the Scaled IPO proceeds XGBoost models. 66
Figure 11. Variable importance plot for the IPO underpricing XGBoost models. 67

Tables

Table 1. Descriptive statistics. All VC-backed IPOs. 41
Table 2. Descriptive statistics. Successful VC-backed IPOs (Scaled IPO proceeds = 1). 42
Table 3. Descriptive statistics. Unsuccessful VC-backed IPOs (Scaled IPO proceed*s* = 0). 43
Table 4. Descriptive statistics. Successful VC-backed IPOs (IPO underpricing = 1). 44
Table 5. Descriptive statistics. Unsuccessful VC-backed IPOs (IPO underpricin*g* = 0). 45
Table 6. Variable composition of the binary classification models. 46
Table 7. Pearson and point-biserial correlation matrix. 47
Table 8. Variance inflator factors (VIFs). 48
Table 9. Scaled IPO proceeds logit coefficient estimates. 53
Table 10. IPO underpricing logit coefficient estimates. 56
Table 11. Performance metrics for the logit and random forest models. 57
Table 12. Performance metrics for the logit and XGBoost models. 59
Table 13. 5-fold cross-validation test results. 69
Table 14. Summary of the thesis's results. 70

1 Introduction

Flickinger (2023) states that "*since the earliest recorded history, innovation has driven economic momentum (consider the wheel, radio, or agricultural tools)*". Entrepreneurs have searched funds for their risky ideas for centuries, but since 1946, the modern institutional venture capital (VC) industry has funded innovations to bring them to market (Lerner and Nanda, 2020, pp. 238–239). VC is not only important but also necessary for many innovations since any great idea could turn to insignificant without a proper funding.

In May 2020, seven of the eight largest public companies in the world by market capitalization – Alphabet, Alibaba, Amazon, Apple, Facebook, Microsoft, and Tencent – were all backed by VC before their introduction to the public markets (Lerner and Nanda, 2020, p. 237). As of early April 2026, among the eight largest companies in the world, Alphabet, Amazon, Apple, Broadcom, Microsoft, and NVIDIA all were VC-backed before public listing (CompaniesMarketcap.com, n.d.; PitchBook, n.d. -a; PitchBook, n.d. -b). These statistics highlight the persistent importance of VC and suggest that VC backing could be a factor driving success among publicly traded companies.

Private VC companies eyeing the public markets need to go through an initial public offering (IPO) process, which enables them to raise new capital from public investors, while creating an opportunity for current investors to exit their investments (Zuniga, 2024). IPOs can be the ultimate events to test the effects of VC backing since the process is targeted toward the general public, potentially representing an unbiased judge. By contrast, private companies are not accessible to all investors, and investments in those companies are usually made by, for example, institutional investors (PitchBook, 2023a).

This thesis focuses on data covering VC-backed IPO companies in the European Economic Area (EEA) and European Free Trade Association (EFTA) countries during 2000–2023. Accordingly, the next section summarizes trends from the last 10 years of the data period.

1.1 VC industry developments during 2013–2023

Figure 1 illustrates the year-by-year developments of total VC deal activity in Europe during 2013–2023. Across this period, the Compound Annual Growth Rate (CAGR) of total VC deal value is 18.4%, while CAGR of total VC deal count is 7.2%. These two rates imply that the average size of a single VC deal has risen in Europe over 2013–2023, as the growth rate of deal value outpaces the growth rate of total deal count. In Figure 1, the average size of a single VC deal has grown at a CAGR of 10.5% per year.

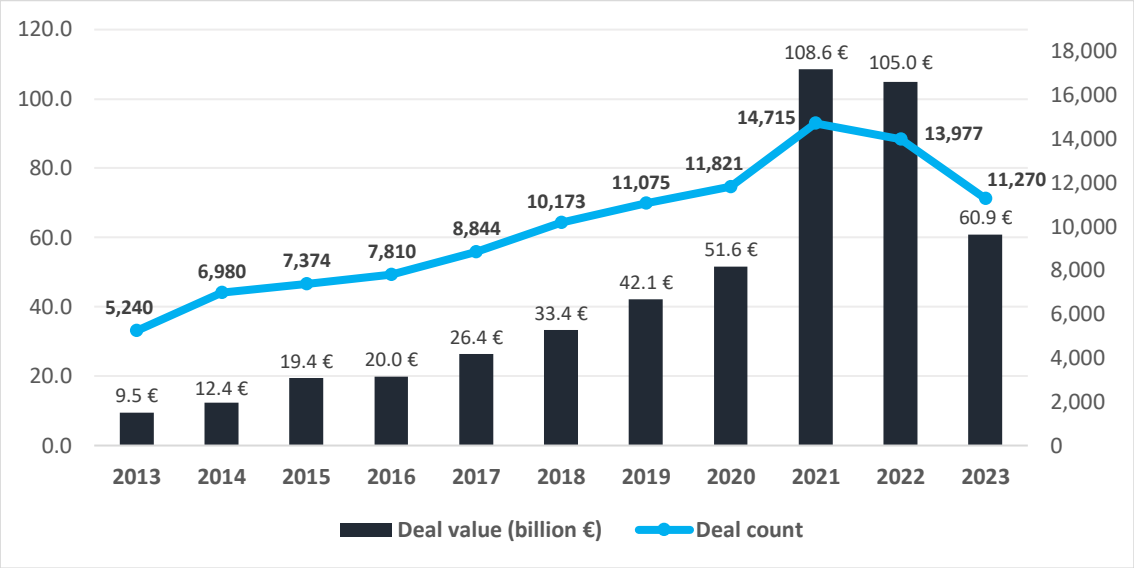


Figure 1. Total VC deal value and total VC deal count in Europe during 2013–2023 (PitchBook, 2023b; 2024).

As shown in Figure 1, VC deal activity boomed in 2021 and 2022, but 2023 marked sharp year-over-year declines of 42.0% and 19.4% in total deal value and total deal count, respectively. The boom years of 2021 and 2022 were especially driven by attractive valuations and excess liquidity in a still-low interest rate environment, ahead of uncertainty caused by the COVID-19 pandemic, high inflation, rising interest rates, a slowing economy, and geopolitical tensions (Patel, 2022, pp. 2–4; 2023, pp. 3–5). In 2023, deal activity declined from the prior two years' abnormal levels together with harsher macroeconomic conditions (Rajan, 2024, p. 4).

Figure 2 illustrates the year-by-year distribution of deal values during 2013–2023. The chart shows that the VC deals of over 25 million EUR account for the largest share of the total deal value in Europe across all years. In 2013, the share of over 25 million EUR deals represents 33.7% of total deal value, rising to 59.5% in 2023. The data reinforces the observation from Figure 1 that, between 2013 and 2023, companies have shifted toward closing larger VC deals.

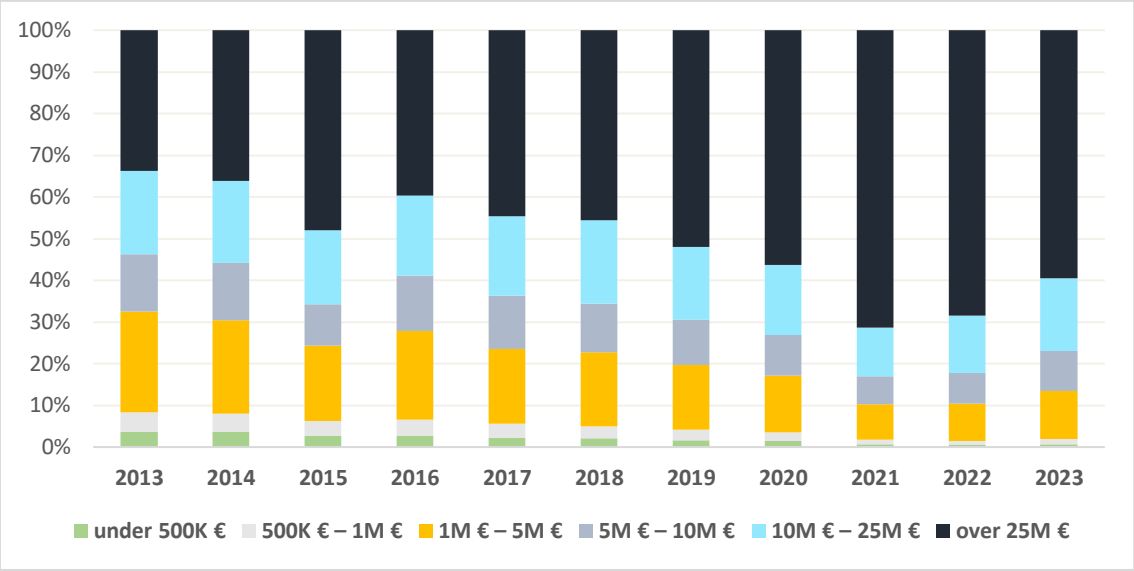


Figure 2. Share of total VC deal value by size in Europe during 2013–2023 (PitchBook, 2023b; 2024).

Figure 3 illustrates the total count of VC exits and VC-backed IPOs in Europe during 2013–2023. The number of completed IPOs accounts for the smallest share of the total number of VC exits across the years. For example, in 2023, the share of VC-backed IPOs is only 3.0%. By contrast, in the peak year of VC-backed IPOs (2021), the corresponding share is 15.8%. The CAGR of the number of VC-backed IPOs over 2013–2023 is –1.9%, indicating a slight downward trend in IPO activity over a long period. At the same time, total VC exit activity has been growing at a CAGR of 7.7% per year.

The VC-backed IPO market boomed in 2021 driven by the same factors that propelled the VC deal activity. Attractive valuations, excess liquidity, and still-favorable market

conditions amid an uncertain future drove VC-backed companies to pursue IPOs (Patel, 2022, pp. 11–12). In 2022 and 2023, a flight from public equities, worsened exit valuations, and low visibility caused by a weakened macroeconomic environment dried up the VC IPO market (Patel, 2023, pp. 12–13; Rajan, 2024, pp. 10–11).

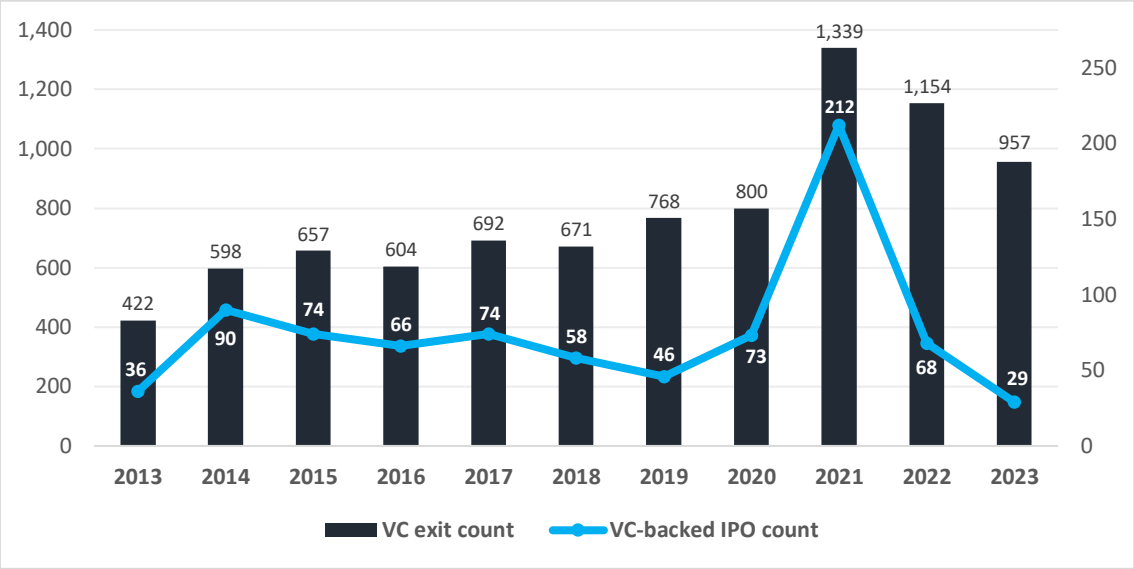


Figure 3. Total VC exit count and total VC-backed IPO count in Europe during 2013–2023 (PitchBook, 2023b; 2024).

1.2 Purpose and contribution of the thesis

Lehnertz, Plagmann, and Lutz (2022, p. 113) state that receiving a mega-deal (VC financing round with a volume of at least 100 million USD) in a less mature venture capital market, such as Europe, could yield a stronger signal compared to receiving it in the most mature market, the U.S. However, to the best of this thesis author's knowledge, no other relevant academic article published in a prestigious journal studies the associations between the VC round size and IPO performance in Europe. This observation highlights an apparent gap in the academic finance literature. Therefore, the main research question of the thesis is: *Is the size of the largest disclosed VC equity financing round associated with IPO success?* Second, the thesis seeks to find out the direction of the association, if it exists.

The idea for this thesis is inspired by Lehnertz, Plagmann, and Lutz (2022), and the applied methodology is strongly influenced by Sapkota (2024), who uses logistic regression (logit), and random forest (RF) and extreme gradient boosting (XGBoost) machine learning algorithms to study cryptocurrency exchange defaults. The thesis aligns with the broader trend of implementing machine learning algorithms in IPO research (e.g., Bastı, Kuzey, Delen, 2015; Quintana, Sáez, and Isasi, 2017; Ang, Chia, and Saghafian, 2021). However, to the best of the author's knowledge, no other academic article has applied modern machine learning algorithms to study the associations between the VC round size and IPO success like this thesis.

The contribution of this thesis is twofold. First, it seeks to fill a small part of an evident gap in the academic European VC-backed IPO performance literature by introducing a perspective on the VC round size. Second, the thesis enhances traditional logit model findings in this academic gap by employing modern random forest and XGBoost machine learning algorithms for robustness.

1.3 Hypotheses of the thesis

This thesis measures IPO success with two variables: *Scaled IPO proceeds* and *IPO underpricing*. Scaled IPO proceeds are the total gross proceeds company generates in an IPO, scaled by its total assets before the offering. IPO underpricing is the percentage difference between the closing price per share on the first trading day and the IPO offer price. The measure of interest in this thesis is the size of the largest disclosed VC equity round before company's IPO. More detailed variable descriptions are in Appendix 1.

The following hypotheses are developed to test the statistical significance of the empirical findings. The *Scaled IPO proceeds* success variable has its own set of hypotheses, and *IPO underpricing* its own. The thesis tests the findings against the null hypotheses (H_0), and the alternative hypotheses (H_1) specify the expected direction of the results.

Scaled IPO proceeds hypotheses:

H_0: *The size of the largest disclosed VC equity round is not conditionally associated with Scaled IPO proceeds.*

H_1: *The size of the largest disclosed VC equity round has a positive conditional association with Scaled IPO proceeds.*

IPO underpricing hypotheses:

H_0: *The size of the largest disclosed VC equity round is not conditionally associated with IPO underpricing.*

H_1: *The size of the largest disclosed VC equity round has a negative conditional association with IPO underpricing.*

The conditionality means that the associations are conditional on the other predictor variables and fixed effects in the thesis. The foundation for the alternative hypotheses lies in the mechanisms of signaling theory (Spence, 1973; Leland and Pyle, 1997), VC certification (Megginson and Weiss, 1991), VC monitoring (Barry, Muscarella, Peavy, and Vetsuypens, 1990), and staged financing (Gompers, 1995). Based on these, this thesis argues that the continuous size measure of the largest disclosed VC equity round contains valuable information of the strength of VC commitment and involvement in a company, which ultimately reduces information asymmetries in IPOs between company insiders and investors. Therefore, a larger VC round is theoretically a reinforcing signal of IPO quality, which encourages investors to participate in the offering with reduced risk aversion. Section 4 provides a detailed look at each of the four theoretical mechanisms.

The expected positive direction of the alternative *Scaled IPO proceeds* hypothesis (H_1) is justified by the empirical findings of Botazzi and Da Rin (2002) in Europe and by evidence

from Lehnertz, Plagmann, and Lutz (2022) in the U.S. Botazzi and Da Rin find that issuing companies backed by VC raise significantly more IPO proceeds compared non-VC-backed IPO companies. On the other hand, Lehnertz et al. show that companies that receive a mega-deal generate significantly higher IPO proceeds compared to companies that are not mega-deal recipients. Thus, it is logical to expect that a larger disclosed VC round is positively associated with *Scaled IPO proceeds*, as it signals a higher degree of VC involvement, ultimately leading to higher IPO demand and increased IPO proceeds.

The expected negative direction of the alternative *IPO underpricing* hypothesis (H_1) is justified by the empirical findings of Tanda and Manzi (2020), Pennacchio (2014), Coakley, Hadass, and Wood (2009) and Honorine and Emmanuelle (2019) in Europe. All these papers imply that VC-backing or VC syndication reduces IPO underpricing through VC certification. Therefore, this thesis suggests that a larger disclosed VC round signaling a higher degree of VC involvement (amplifying the VC certification effect) is negatively associated with underpricing.

1.4 Structure of the thesis

The remainder structure of the thesis is as follows. Sections 2 and 3 focus on introducing the fundamental concepts of venture capital and IPOs, respectively. Section 4 presents the theoretical foundation crucial to this thesis, and Section 5 reviews the empirical findings of the relevant academic finance literature. Section 6 explains the data preparation process, followed by Section 7, which presents the methodology applied in this thesis. Section 8 introduces the empirical findings with discussion, and Section 9 concludes.

2 Venture Capital (VC)

This section focuses on venture capital (VC) from an equity perspective. Section 2.1 provides an overview of VC and Section 2.2 discusses the stages of VC financing rounds.

2.1 Overview of VC

According to Cumming (2012, p. 1), venture capital (VC) is a part of the broader class of private equity, which relates to investments that are not traded on a stock exchange. Cumming specifies that VC involves investments into startup and early-stage companies that are innovative and have high growth. From another perspective, Metrick and Yasuda (2021, p. 2) characterize VC as a financial intermediary, which collects money from investors and then invests this capital in portfolio companies.

The financial intermediary vehicle refers to a limited partnership launched by a general partner (GP). The limited partnership is known in the industry as a VC fund. The GP of a VC fund is controlled by a VC firm, which is run by so-called venture capitalists. Orchestrated by the GP, a VC fund raises money from limited partners (LPs) for portfolio investments. LPs are typically institutional investors, like pension funds, university endowments, and large corporations. Angel investors are wealthy individuals who invest in startup and early-stage companies but do not meet the financial intermediary definition. Angel investors use their own capital for investments and do not raise it from LPs. (Metrick and Yasuda, 2021, pp. 1, 17)

An essential characteristic of VC is that it actively monitors and supports its portfolio companies in managing their business. This differs from solely providing capital to portfolio companies and creates an additional source for greater performance if the VC is effective in its involvement. Typically, VC takes, for example, board positions in their portfolio firms, which enables it to actively participate in business. In addition, VC can attract talent in its portfolio companies relying on their valued reputation and contacts in the industry (Metrick and Yasuda, 2021, pp. 1–2).

Cumming (2012, p. 1) implies that a typical characteristic of all VC investments is that the startup and early-stage companies cannot afford to pay dividends on equity because their cash flows are small. Thus, VCs' primary goal is to exit their portfolio investments, maximizing returns. VC funds operate for a limited period, usually around 10 years depending on the fund, and in the end the portfolio companies are sold or listed on a stock exchange through an IPO. Corporate venture capital (CVC) – where large companies invest in stakes in other companies – is its own industry and has connections to strategic investing, where the primary goal is not an exit. (Metrick and Yasuda, 2021, pp. 2, 17).

Finally, VC has the distinct characteristic of investing in the growth of new businesses, not acquiring existing companies (Metrick and Yasuda, 2021, p. 3). VC investments by portfolio company stage, or the stages of VC financing rounds that startup and early-stage companies seek, are known by several names, depending on the source. In Section 2.2, VC financing rounds are classified by stage into *pre-seed*, *seed stage*, *earlier-stage*, and *later-stage* rounds.

2.2 VC round stages

After the entrepreneurs of a startup have launched their company as far as it can go, they can consider raising VC funds to accelerate business. Raised VC funds can, for example, be used to expand company's offerings, building relevant teams, or reaching a profitability target. Companies seeking VC funding raise capital in rounds, in which one or more investors can participate. Each VC financing round can be classified by the stage the company seeking funding is in its lifecycle. (Knickerboecker, n.d., p. 4)

The first VC financing round type is called *pre-seed*. Traditionally, this financing round type has referred to funding without institutional investor involvement (such as VC fund). However, more VC funds have started to invest in pre-seed companies, making it a relevant VC financing round type. Other financing sources for pre-seed companies

include crowdfunding, friends and family, and small-business loans (SMBs). (Knickerboecker, n.d., pp. 4–5)

Seed stage VC financing rounds follow the *pre-seed*, with greater VC involvement, while capital is provided in exchange for the entrepreneurs' equity in the company. At this stage, typically small-to-medium VC funds may be involved. Angel investors may also be interested in investing. The average *Seed stage* VC financing round size ranges from 1 to 5 million USD, while the company seeking funding typically has a small team and is getting into the market by developing and testing a minimum viable product (MVP), which is an early product with just enough features to use. At this stage, the company has low or nonexistent revenues. (Knickerboecker, n.d., pp. 4, 6)

Earlier-stage VC financing rounds include Series A and Series B financing, which are formally recognized in the VC industry. In these series, large multi-stage VC funds and small-to-medium VC funds are involved. The average Series A and Series B VC financing round size range from 2 to 20 million USD. When raising a Series A round, the company is typically growing its team, finding product-market fit with an improved MVP, and setting up its distribution. By contrast, when raising a Series B round, the company is already entering new markets while continuing to develop the product. At the *earlier-stage* VC financing rounds, the company has increasing revenues but need additional growth capital. (Knickerboecker, n.d., pp. 5–6)

Later-stage VC financing rounds include Series C through Series G, each with unique characteristics defined by the company's stage and potential applications for the raised capital. Other financing sources for later-stage companies include mezzanine, bridge, and IPOs, which are outside the scope of traditional VC financing. (Knickerboecker, n.d., p. 5)

3 Initial Public Offering (IPO)

Section 3 explains the concept of initial public offering (IPO). Section 3.1 provides an overview of IPOs and Section 3.2 focuses on the IPO process in detail.

3.1 Overview of IPOs

An initial public offering (IPO) is a process in which a private company sells equity or debt securities to the public for the first time, seeking to enter a securities market where liquidity increases. In an IPO, securities are sold to many different investors, compared with the time as a private company when capital is raised from only a few investors. This thesis focuses on equity IPOs. (Ritter, 1998, p. 1)

Typical reasons for a company to start an IPO process include raising capital for growth acceleration, increasing company visibility among the general public, enabling existing investors to exit, and preparing for an inheritance event or upcoming buy-out of majority share capital (Euronext, 2025, p. 3). According to Caselli and Negri (2021, p. 326), an IPO can be arranged as an initial selling public offer or an initial subscription public offer. In a selling offer, existing investors sell their shares to the public, and in a subscription offer, the IPO company issues new shares, which are then sold to the public.

3.2 IPO process

Euronext is the largest capital market infrastructure in Europe, operating regulated stock exchanges in Amsterdam, Athens, Brussels, Dublin, Lisbon, Milan, Oslo, and Paris (Euronext, n.d.). Accordingly, this section focuses on the IPO process from a Euronext perspective, dividing the journey to a public company into the partly simultaneously running *documentation* and *marketing* phases. IPOs can take four to six months to complete (Euronext, 2025, p. 10). Additionally, Caselli and Negri (2021, pp. 323–324) explain that IPOs also require organizational changes within a company, as well as evaluating and meeting economic and financial requirements.

3.2.1 Documentation phase

Once all relevant IPO advisors have been appointed by the issuing company, their IPO journey may begin from the *documentation* phase. Relevant advisors to a company's IPO process include investment banks, legal advisors, equity advisors, auditors, listing sponsors, and public relations agencies. Investment banks serve in various roles in an IPO process, including the underwriter role, where the bank operates as an intermediary between the issuing company. All IPO advisors appointed by a company form a banking syndicate. (Euronext, 2025, pp. 18–20)

The first stage of the *documentation* phase is *preparation*, during which the IPO candidate and its advisors draw up a formal legal document called a prospectus, which will become public. This document presents a detailed look on the issuing company including its sector, business, financial information, management, and it discusses potential risk factors. In addition, the prospectus outlines offering details such as number of shares to be issued, the IPO offer price range, a timetable for the subscription period, and the use of IPO proceeds. Ongoing due diligence is essential throughout the preparation stage to ensure all relevant information is properly disclosed in the prospectus. The *preparation* stage ends with a confidential filing of the prospectus with local financial market regulators. (Euronext, 2025, pp. 9–10)

The second stage of the *documentation* phase is *review*, where the local financial market regulators carefully review the filed prospectus. During this stage, the regulators evaluate whether the prospectus includes all relevant information for the IPO investors to participate in the offering. Once the prospectus is approved by the regulators, the issuing company publishes a public an intention to float announcement. Thereby, the initiated IPO is made public, and the *documentation* phase has ended. (Euronext, 2025, pp. 9, 11)

3.2.2 Marketing phase

As the *documentation* phase kicks off the IPO process, the *marketing* phase begins running two months before the IPO date, simultaneously with prospectus preparations. *Pilot fishing* is the first stage of the *marketing* phase. Arranged by the underwriters, this stage includes confidential early-look meetings (or so-called pilot-fishing sessions) with potential key investors, which may provide secured pre-IPO demand through formal commitment subscriptions. Investors committing to subscribe are called anchor and cornerstone investors, who generate essential security and credibility for the IPO execution. In addition, during the *pilot fishing* stage, marketing documents are prepared for the upcoming IPO roadshow and meetings with investors and syndicate's equity research analysts. After management has introduced the IPO candidate for the equity analysts, they draw up an independent research report about the company's business, its competitive environment, and an indicative valuation for the IPO offer price range. (Euronext, 2025, pp. 9, 13)

The *pre-deal investor education* stage follows the *pilot fishing* while the IPO process is still hidden from the public. Using the research report prepared by the equity analysts, the analysts and the company begin to gauge key investors' impressions of the indicative valuation for the IPO offer price range. Based on the impressions and the sentiment around the valuation range, it is then adjusted up and down, leaving a discount as a reward to IPO investors bearing the risk of investing in a new public company. The *pre-deal investor education* stage ends simultaneously with the *documentation* phase, when the prospectus is approved. (Euronext, 2025, pp. 9, 14)

The next stage in the *marketing* phase is called the *placement*. Typically, around the intention to float announcement (two weeks before the IPO date), the IPO company schedules the opening of the offer period with a carefully prepared presentation to the public. This presentation is only one of many management roadshow activities in which the IPO is advertised to potential institutional and retail investors, generating demand. During the *placement* stage's offer period, the underwriters collect orders from fund

managers indicating how many shares they would like to purchase at a fixed price in the determined IPO offer price range. This process is called book building, after the tradition of recording orders in a book during meetings. (Euronext, 2025, pp. 9, 14–15)

At the end of the *placement* stage, the official IPO pricing takes place, in which the final offer price is set, and shares are allocated based on investors' orders. The final IPO offer price is set based on demand across different price levels in the range of the offering. However, as setting a higher final offer price decreases investor demand, the issuing company and underwriters must find a satisfactory balance – by adjusting the price – between the total generated IPO proceeds, expected valuation, and the desired momentum from the IPO. In other words, theoretically, a higher final offer price increases IPO proceeds and expected valuation but reduces IPO momentum, whereas a lower offer price decreases the IPO proceeds and expected valuation but results in greater IPO momentum. (Euronext, 2025, pp. 9, 15)

In a typical IPO arrangement, underwriters purchase all IPO shares from the issuing company and sell them to the public on the IPO date. Underwriters in the banking syndicate buy the shares from the issuing company at the final IPO offer price, less a gross spread that they charge for their underwriting services. This arrangement between the underwriters and the IPO company is called the firm commitment. (Bodie, Kane and Marcus, 2014, p. 61)

If overall IPO demand is weak, the IPO can be cancelled, and if demand exceeds 100% of the base deal, the company can choose to exercise extension clause or a so-called greenshoe (an over-allotment option), both of which increase the number of offered shares. With the greenshoe option, the underwriter buys back newly issued shares to stabilize the share price if it slumps below the offer price after the IPO. After IPO pricing, the *post-listing* stage, with trading on the stock exchange, takes over as settlement and share deliveries are finalized. This is the end of the IPO process. (Euronext, 2025, pp. 9, 15–16)

4 Theoretical foundation and the links to the VC round size

This section focuses on the relevant theoretical foundation and its connection to the continuous measure of interest in this thesis, which is the size of the largest disclosed VC equity round. Section 4.1 introduces the foundational problem of information asymmetry, while Sections 4.2 to 4.4 each provide a theoretical justification for studying the association between the VC round size measure and IPO success.

4.1 Information asymmetry

Akerlof (1970) develops a market theory that highlights the difficulty and foundational problem of distinguishing between good and bad quality in the business world. Akerlof explains that buyers, in general, cannot observe quality with certainty, while sellers have more knowledge about the quality of goods or services they are trying to sell. The author calls this imbalance between the buyers and sellers information asymmetry, which can create mispricing in markets.

As Akerlof's theory suggests, markets suffer from an information asymmetry problem. This motivates the need for credible quality signals that can be assessed by the market participants who are at an informational disadvantage. The analogy to the VC and IPO markets relevant to this thesis can be derived from the reasoning that investors (buyers) evaluating VC-backed companies at IPO face the same informational disadvantage relative to the listing companies (sellers) seeking financing.

4.2 Signaling theory

Following the idea of Spence (1973), under conditions of uncertainty (information asymmetry), a quality signal is credible when its cost is negatively correlated with the sender's (seller's) underlying quality. Thus, those with higher underlying quality face a lower cost of signaling quality compared to those with lower quality. Applying Spence's theory, low-quality sellers attempting to mimic the signal of high-quality sellers face

higher expected costs, making imitation unprofitable. Summarizing the relevant ideas of the paper, because sellers with higher underlying quality face a lower cost of signaling quality relative to those with lower quality, only high-quality sellers can rationally choose to pursue credible quality signals.

This signaling theory introduced by Spence (1973) provides a foundation for the significance of the largest disclosed VC equity round size in IPOs. The size of a VC round is an observable variable for all market participants evaluating an IPO. For outside observers, a larger disclosed VC round indicates a greater VC commitment to the invested company. Applying the mechanics from Leland and Pyle (1997), when venture capitalists back a high-quality company with large investments, the expected cost (loss) for the commitment is lower. By contrast, if venture capitalists were to back a low-quality company with large investments, the expected cost of the commitment is higher, making the investment likely to be unprofitable. Therefore, investors evaluating an IPO can draw theoretical conclusions about company quality by observing the size of the largest VC round and, ultimately, about IPO success.

4.3 VC certification and monitoring

Megginson and Weiss (1991) introduce the fundamental concept of VC certification. They explain that this third-party certification is particularly effective at mitigating information asymmetries between the issuing company's insiders and public investors in IPOs regarding the listing company's value, highlighting the informational disadvantage of public investors. VCs, with reputational capital at stake, can help ensure that all relevant private information is reflected in the IPO offer price.

Based on Megginson and Weiss (1991, pp. 881–882), VC certification is a credible third-party certification for the following main reasons. First, because reputational venture capitalists bring companies to the market on an ongoing basis, they have a strong incentive to build a favorable reputation with all market participants, which they can use in upcoming IPOs. Second, according to Sahlman (1990), venture capitalists achieve high

returns, which are related to the size, age, and historical performance of the VC fund. In addition, Sahlman finds that venture capitalists are effective at attracting additional capital, and investments in reputational capital further incentivize them to remain competitive in the venture capital industry and capital markets. Third, the services provided by the venture capitalists are costly and difficult for the offering company to obtain on its own. Considering this, VCs are brought into a company only if they are believed to bring significant value.

Barry, Muscarella, Peavy, and Vetsuypens (1990) introduce the fundamental concept of VC monitoring. Based on their paper, the core idea in VC monitoring is that venture capitalists' active involvement in company's governance reduces investor uncertainty. The monitoring role of venture capitalists includes taking significant equity stakes in their investee companies, participating in board activities, and maintaining their investment in the company beyond the IPO. VC monitoring is complementary to the VC certification concept of Megginson and Weiss (1991), and its foundation lies in active monitoring rather than solely in reputational endorsement.

In this thesis, the measure of the largest disclosed VC equity round size reflects information from both VC certification and monitoring because the thesis focuses solely on pure VC rounds. As the VC round size variable is continuous measure of financial commitment, it also captures valuable information of the degree of VC involvement, which determines the strength of the effects. With larger VC rounds, venture capitalists have more skin in the game, amplifying their level of monitoring and the effect of VC certification.

4.4 Staged financing

Gompers (1995) examines the foundations of staged VC financing. Gompers explains that staged financing is a strategy favored by venture capitalists, whereby they infuse capital into a company in stages, while actively monitoring its progress and preserving the option to abandon (stop financing) if the company's progress is not satisfactory.

According to the author, venture capitalists decide based on potential agency and monitoring costs how frequently periodical progress is evaluated and further capital is supplied.

Applying the presented mechanics from Gompers (1995), the measure of the largest disclosed VC equity round size includes information on renewed VC endorsement within a sequential process of actively evaluating a company's progress. In other words, each round represents a renewed assessment of firm quality, and the level of VC conviction is reflected in the size of the VC round. Thus, investors evaluating an IPO can draw theoretical conclusions about its potential success by observing the size of the largest VC round.

5 Literature review

This section reviews the academic finance literature on IPO underpricing, VC-backed IPO performance, and VC syndication. Section 5.1 focuses on IPO underpricing. Section 5.2 focuses on VC-backed IPO performance evidence from the U.S., while Section 5.3 examines the European evidence. Section 5.4 focuses on VC syndication from both the U.S. and European perspectives.

5.1 IPO underpricing

IPO underpricing is the percentage difference between the closing price per share on the first day of trading and the IPO offer price. While IPO underpricing is a common IPO success metric, there is a disagreement in the academic literature about whether underpricing should be considered a signal of IPO success or not. According to Lehnertz et al. (2022, p. 104) the media usually conclude that higher IPO underpricing is a signal of a successful IPO. Indeed, there are several reasons why IPOs are deliberately underpriced. If underpricing is an intentional process in the markets, it is reasonable to assume that companies view only positive underpricing as IPO success.

Benveniste and Spindt (1989) and Hanley (1993) provide demand-based explanation for intentional IPO underpricing, which implies that IPO underwriters tend to adjust the final IPO offer price only partially upward when overall pre-IPO demand is strong. Following Hanley, the information gathered about investors' pre-IPO demand influences the final offer price set by the underwriters, and in response to strong demand, they leave residual underpricing in the final offer price for investors who truthfully reveal their interest on the issue. In other words, the underwriters do not tend to adjust the final offer price fully to match what the pre-IPO demand suggest. Thus, high IPO underpricing may reflect deliberate partial offer price adjustment due to investors truthfully revealing strong pre-IPO demand. This is called the *market feedback hypothesis* (Ritter, 1998, p. 8).

Allen and Faulhaber (1989) and Welch (1989) argue that companies deliberately underprice at their IPOs, signaling higher quality and leaving a good impression with investors. Allen and Faulhaber find that only high-quality issuers exercise underpricing because they, for example, expect to recoup the cost of a lower offer price with a more favorable price at a subsequent offering. The underlying idea is that company insiders have an informational advantage regarding the listing company's value. Welch reinforces the signaling argument with findings that IPO underpricing is a credible signal of quality for investors because it is costly to mimic by low-quality companies. This is called the *signaling hypothesis* (Ritter, 1998, p. 9).

The *winner's curse hypothesis* suggests that IPOs are underpriced to encourage investors at an informational disadvantage to participate in the offer. The rationale is that uninformed investors have a disadvantage in IPO evaluation. When they choose to participate in better (more desired) IPOs, they receive only a fraction of the subscribed shares, whereas in worse (less desired) IPOs, they receive all the shares they subscribed for. The actual "curse" is that the uninformed investors get all shares because informed investors do not want to participate in the IPO. By underpricing an IPO, company can compensate uninformed investors for this allocation bias. (Ritter, 1998, pp. 7–8)

According to the *bandwagon hypothesis*, IPOs are underpriced to encourage the first investors to subscribe for IPO shares, so that others can follow, amplifying the bandwagon effect (Ritter, 1998, pp. 8–9). The *investment banker's monopsony power hypothesis* suggests that investment bankers exploit their superior market knowledge and relationships by focusing on offering underpriced shares to potential buy-side IPO investors (Ritter, 1998, p. 9). Based on the *ownership dispersion hypothesis*, companies deliberately underprice their IPOs to create excess demand and to achieve a diverse ownership structure with many small shareholders (Ritter, 1998, p. 10).

Considering also the other side of the IPO underpricing discussion in the academic literature, Loughran and Ritter (2004) argue that with higher IPO underpricing, the

company is leaving money on the table. In other words, the issuer generates less proceeds from the IPO relative to its potential, which one could argue is a signal of an unsuccessful IPO.

5.1.1 Hot and cold IPO markets and evidence from Europe

Ibbotson and Jaffe (1975) are among the first to study “hot issue” markets. Helwege and Liang (2004, p. 541) explain that typical characteristics of hot IPO markets include high IPO underpricing, particularly large offering volumes, frequently oversubscribed IPOs, and occasional clustering in specific industries. By contrast, cold IPO markets have low IPO underpricing, small offering volumes, and rarely oversubscribed IPOs.

Hot and cold IPO markets are usually explained by signaling mechanisms, industry-based theory, and the overly bullish behavior of irrational investors. The signaling mechanism suggest that during hot IPO markets, higher quality companies go public because information asymmetry costs (adverse selection costs) are not fully reflected in IPO offer prices. The industry-based theory implies that hot IPO markets are driven by specific industries with small and risky IPOs. (Helwege and Liang, 2004, p. 542)

Gajewski and Gresse (2006) study IPO underpricing in Europe during three market periods: *low period* (1995–1997), *hot market* (1998–2000), and *extremely cold market* (2001–2004). Based on a sample of 2104 European companies across 15 different countries, during the *low period*, the mean IPO underpricing is 15.86%. During the *hot market* and the *extremely cold market* periods, the mean IPO underpricing is 27.18% and 12.19%, respectively. This evidence implies that higher IPO underpricing follows hot IPO markets, and lower IPO underpricing follows cold IPO markets in Europe.

Assoil and Laporte (2024) show that IPO underpricing is slightly higher during bull (hot) market periods than bear (cold) market periods in Euronext Paris during 2000–2020. The authors classify a market downturn in the early 2000’s, the financial crisis during 2007–2008, a short recession in 2016, and the COVID-19 pandemic in the bear market category.

Other periods belong to the bull market category. During the bull market periods, the mean IPO underpricing is 3.904%, and 3.536% during the bear market periods.

5.2 VC-backed IPO performance in the U.S.

VC-backing in IPOs in the U.S. is a widely studied field in the academic literature of finance. Megginson and Weiss (1991) are among the first to present empirical evidence that VC-backing is a certifying signal of IPO quality, reducing investor uncertainties when going public. Barry, Muscarella, Peavy, and Vetsuypens (1990) show that venture capitalists' intensive monitoring of their investee companies signals IPO quality. Following the foundational work of both papers, many others have examined VC-backing in IPOs (e.g., Gompers, 1996; Lee and Wahal, 2004; Hochberg, Ljungqvist, and Lu, 2007; Chemmanur, Krishnan, and Nandy, 2011; Krishnan, Ivanov, Masulis, and Singh, 2011).

VC rounds, as a variable of interest, are not as widely examined in the academic literature as the effects of pure VC-backing. Barry et al. (1990) consider the size of the equity stake held by a venture capitalist as a monitor of IPO performance, and Tian (2011) finds that the number of completed VC rounds is positively associated with eventually completing an exit through an IPO, if the company is located far from the VC investor.

Lehnertz, Plagmann, and Lutz (2022) is the closest to the goal of this thesis. They examine whether receiving a so-called mega-deal (VC financing round with a volume of at least 100 million USD) is associated with improved IPO success in the U.S. Lehnertz et al. find that VC-backed companies that have secured a mega-deal generate, on average, 2.5 times higher IPO proceeds and 19.2% higher IPO underpricing. While the higher IPO underpricing may appear inconsistent with the alternative IPO underpricing hypothesis (H_1) in this thesis, Lehnertz et al. also show that a mega-deal increases the IPO offer price revision by 8.6% on average.

As Lehnertz et al. (2022) imply in their study, the simultaneous increases in average IPO underpricing and IPO offer price revision demonstrate that easily observable mega-deals

(100 million USD) create additional demand in IPOs by potentially attracting nonspecialized investors. In other words, Lehnertz et al. view the higher IPO underpricing as evidence of a demand effect related to securing a mega-deal rather than, for example, an inconsistency with the theory that a higher degree of VC commitment with a large VC deal reduces information asymmetry and IPO uncertainty through amplified VC certification.

5.3 VC-backed IPO performance in Europe

Botazzi and Da Rin (2002) imply that the European VC industry is far less mature than the well-developed U.S. VC industry. Pantea and Tkacik (2025) review that the European VC market is significantly smaller than the U.S. market, that the VC activity is concentrated in only a minority of countries, and that the government VC (GVC) and the bank VC (BVC) play a more significant role than in the U.S. Pantea and Tkacik also note that fragmented legislation across the EU countries and a preference for raising capital locally (in the home country) characterize the European VC industry. The authors conclude that Europe's fragmentation and the high significance of GVC are likely to persist in the VC market.

Lehnertz, Plagmann, and Lutz (2022, p. 113) report that it would be interesting to examine the effects of mega-deals (larger VC rounds) in less mature venture capital markets, such as Europe. Lehnertz et al. add that receiving such a larger deal in a less mature market could be a stronger signal of quality compared to receiving it in the United States. However, in studies using European samples, the academic finance literature shows an apparent gap in the consideration of VC rounds as a variable of interest when studying IPO performance. In addition, VC-backing in Europe is far less studied field in the academic finance literature than in the U.S.

Botazzi and Da Rin (2002) examine whether VC-backing is associated with higher total funds (proceeds) raised at IPO compared to IPOs without VC-backing. To study the association, the authors scale the total raised funds by total assets. They find that VC-

backed companies raise, on average, 60% more funds at IPO than non-VC-backed companies, and the results are statistically significant. Botazzi and Da Rin indirectly conclude that VC-backing is signal a of quality, as VC helps companies overcome credit constraints, explaining higher IPO proceeds.

Tanda and Manzi (2020) study the effects of VC-backing on IPO underpricing using meta-analysis approach. They show that VC-backing reduces IPO underpricing in Europe, which is in line with the theory that VC involvement reduces information asymmetry through VC certification. However, the results are not statistically significant. Pennacchio (2014) shows that VC-backing reduces IPO underpricing in Italy. They demonstrate that this effect is causal and statistically significant. According to their results, the underpricing of VC-backed IPOs is approximately half that of IPOs without VC-backing, which strongly indicates the existence of the VC certification effect.

Coakley, Hadass, and Wood (2009) demonstrate, using a sample of UK-based companies, that the strength of VC certification effect varies across market cycles. The authors show that during normal markets, VC-backing reduces IPO underpricing, which implies that VC certification does play a role in reducing information asymmetry. By contrast, they also show that during the dot-com bubble period, the VC certification role broke down, IPO proceeds where high, and IPO underpricing peaked. Coackley et al. suggest that this change occurred as venture capitalist and prestigious underwriters neglected their certification functions during the exuberant investor sentiment of the bubble years. Their empirical findings are directly related to the concept of hot and cold IPO markets discussed in Section 5.1.1.

From a different perspective, Bessler and Seim (2012) show that European VC-backed companies have high positive IPO underpricing at the first day of trading and that VC-backed companies that are larger and those that list on main markets have higher abnormal stock returns nearly three years after the IPO, compared to non-VC-backed companies.

5.4 VC syndication in U.S. and European IPOs

Following Lerner (1994, p. 16), VC syndication is a cooperation process in which two or more VCs jointly invest in a portfolio company by pooling capital in a VC round. As discussed in Section 4.3, VCs can theoretically reduce information asymmetry through VC certification on their own, but VC syndication in an investment can potentially create an even stronger signal. In syndicates, VCs combine their expertise, gain reassuring information from one another, and spread investment risk (Lerner, 1994 p. 17).

Tian (2012) studies the effects of VC syndication on IPO performance in the U.S. The author finds that VC syndicates infuse significantly more capital in all types of earlier financing rounds compared to single VCs, and syndicate-backed companies are more likely to complete a successful exit. In addition, Tian shows that VC syndication reduces IPO underpricing, and companies backed by a syndicate have higher IPO market valuations. The results remain strong even after addressing potential endogeneity problems with an instrumental variable approach.

The findings of Tastan and Falconieri (2013) suggest results contrary to Tian (2012). They show that U.S. companies backed by larger VC syndicates have higher IPO underpricing. Tastan and Falconieri conclude that this evidence may be driven by potential coordination problems and conflicts of interest within VC syndicates. The authors note that Tian does not focus on VC syndicate size, while still controlling for it, with no statistical significance. By contrast, Tastan and Falconieri focus their study specifically on differentiating VC syndicate size.

Shifting focus to Europe, Honorine and Emmanuelle (2019) study the effects of VC syndication on IPO underpricing in France. They find that VC syndication reduces IPO underpricing compared to companies not backed by VC syndicated investments. In addition, Honorine and Emmanuelle show that the reducing effect on IPO underpricing is stronger when the VC syndicate is larger. In other words, larger number of VCs in a syndicate is associated with lower IPO underpricing.

6 Data

Section 6 describes the dataset used in this thesis. Section 6.1 explains how the dataset is compiled from IPO data, VC transaction data, and exchange rate data. Section 6.2 provides a basic analysis of the dataset, while Section 6.3 describes variables in this thesis including comprehensive descriptive statistics.

6.1 Dataset preparation

All data are gathered from the LSEG Workspace database. VC-backed IPO data for the period 2000–2023 are collected using LSEG's screener for equity corporate deals. Subsequently, VC transaction data from the LSEG Private Equity Screener are manually matched with the IPO data to create a complete dataset for the empirical research in this thesis. To normalize the dataset's numerical values (different currencies), LSEG exchange rate data for the period 2000–2023 are also collected.

6.1.1 IPO data

By screening the universe of equity corporate deals in LSEG Workspace, a total of 377 VC-backed IPOs can be identified by applying filters introduced in this section. The *Venture Capital-Backed IPO Issue Flag* is set to true to include only VC-backed IPO companies. The listing date may range from 01.01.2000 to 31.12.2023, and the IPO target markets are set for the following countries: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Switzerland, and the United Kingdom.

During 2000–2023, all the countries have been members of the European Economic Area (EEA), which currently consists of 27 European Union (EU) member states and three (out of four) European Free Trade Association (EFTA) countries (Eurostat, n.d. -a). Additionally,

Switzerland and the United Kingdom are included in the IPO target market universe to ensure a greater number of observations. Switzerland is a member of EFTA but not part of the EEA (European Union, n.d. -b), and the United Kingdom withdrew from the EU on 31.01.2020, thus exiting EEA and terminating its EFTA membership at the same time (European Union, n.d. -c). To summarize, the country scope of the thesis includes all EEA and EFTA countries during 2000–2023, as well as the United Kingdom, despite Brexit in 2020. Hereafter, this thesis refers to all countries in the dataset simply as EEA or EFTA countries.

When collecting VC-backed IPO data, only primary IPO issuance tranches are considered to ensure comparability. The IPO issuer's stock must not be traded on any other market at the time of the offering. This ensures that every VC-backed IPO in the dataset represents a first introduction to the public equity markets.

Additionally, the LSEG screener is configured to identify values for the following IPO variables: *Listing Date*, *Issuer/Borrower Founded date*, "*Primary Exchange Nation of Issuer's Stock*, *Issuer/Borrower TRBC Economic Sector*, *Proceeds Amount Incl Overallotment Sold This Market*, *Offer Price*, *Stock Price at Close of Offer/First Trade*, *Net Income After Taxes Before Offering*, and *Total Assets Before Offering*. If a value is missing, the IPO observation is removed from the dataset. However, founding years are manually searched from company websites if the data are missing.

All financial data are obtained in the currency of the original IPO tranche, which usually corresponds to the currency of the target market (exchange nation). Section 6.1.3 explains how all tranche currency values are converted into USD to ensure comparability.

6.1.2 VC transaction data

After identifying the 377 VC-backed IPO companies, the LSEG Private Equity Screener is used to search for their historical VC transaction data. VC transaction data are screened not only for the period 2000–2023 but also to capture all historical VC transactions since

each company's founding. The Private Equity Screener is configured to include only investments with an VC, which the LSEG Workspace classifies as *Seed*, *Early Stage*, *Later Stage*, or *Expansion* by the stage of the investment.

The complete VC transaction dataset is compiled to include the equity portions of the VC financing rounds (*Round Equity Total*) in IPO tranche currency, and the serial numbers of the investment rounds made in the company (*Round Number*). Possible debt portions of a VC round are not included. If VC investment data are unavailable for any IPO company, the observation is removed from the dataset.

After matching the IPO data with the VC transaction data, the number of VC-backed IPO companies in the final dataset is 231. In other words, the sample size (N) of the dataset used in this thesis is 231.

6.1.3 Exchange rate data

To normalize the dataset's financial VC-backed IPO and VC transaction values, daily exchange rate data for the period 2000–2023 are gathered from the LSEG Workspace for eight currency pairs (*CHF/USD*, *DKK/USD*, *EUR/USD*, *GBP/USD*, *HUF/USD*, *NOK/USD*, *PLN/USD*, *SEK/USD*). All financial data are converted into a common currency (USD) to make the data comparable among the VC-backed IPO companies. This is done using the daily closing price of the specific IPO company's listing date. Using these unique exchange rates, all financial values on a company basis are converted to USD.

6.2 Dataset analysis

Figures 4 and 5 illustrate basic characteristics specific to the thesis's dataset. Figure 4 demonstrates the number of IPOs by year in the EEA and EFTA countries during 2000–2023. This figure also shows the distribution of the VC-backed IPOs by country, indicating in which country IPOs have taken place. One can see that France, the United Kingdom, and Germany are the most active IPO countries, with France being the number one

country in IPO activity (91 IPOs). The United Kingdom follows with 70 IPOs, and Germany is third with 23 IPOs. This evidence demonstrates the greater frequency of these three countries once VC-backed companies enter public equity markets.

During 2000–2009, a total of 103 IPOs took place, followed by 106 IPOs during 2010–2019, and 22 IPOs during 2020–2023. Figure 4 also demonstrates how concentrated the IPO activity is during 2000–2009 for three specific years. Years 2005 to 2007 (the period before the global financial crisis) contribute approximately 83% of the total IPO activity during 2000–2009. With a total of 38 IPOs, 2006 is the most active year.

During 2010–2019, the number of IPOs is more evenly distributed across years, demonstrating an economically sound decade for VC-backed companies to go public (compared to the unstable decade of 2000–2009). The broader IPO count trend for 2020–2023 is generally consistent with the numbers presented in Section 1.1. In 2021, IPO activity boomed, followed by a significant decrease in 2022, and nonexistent activity in 2023.

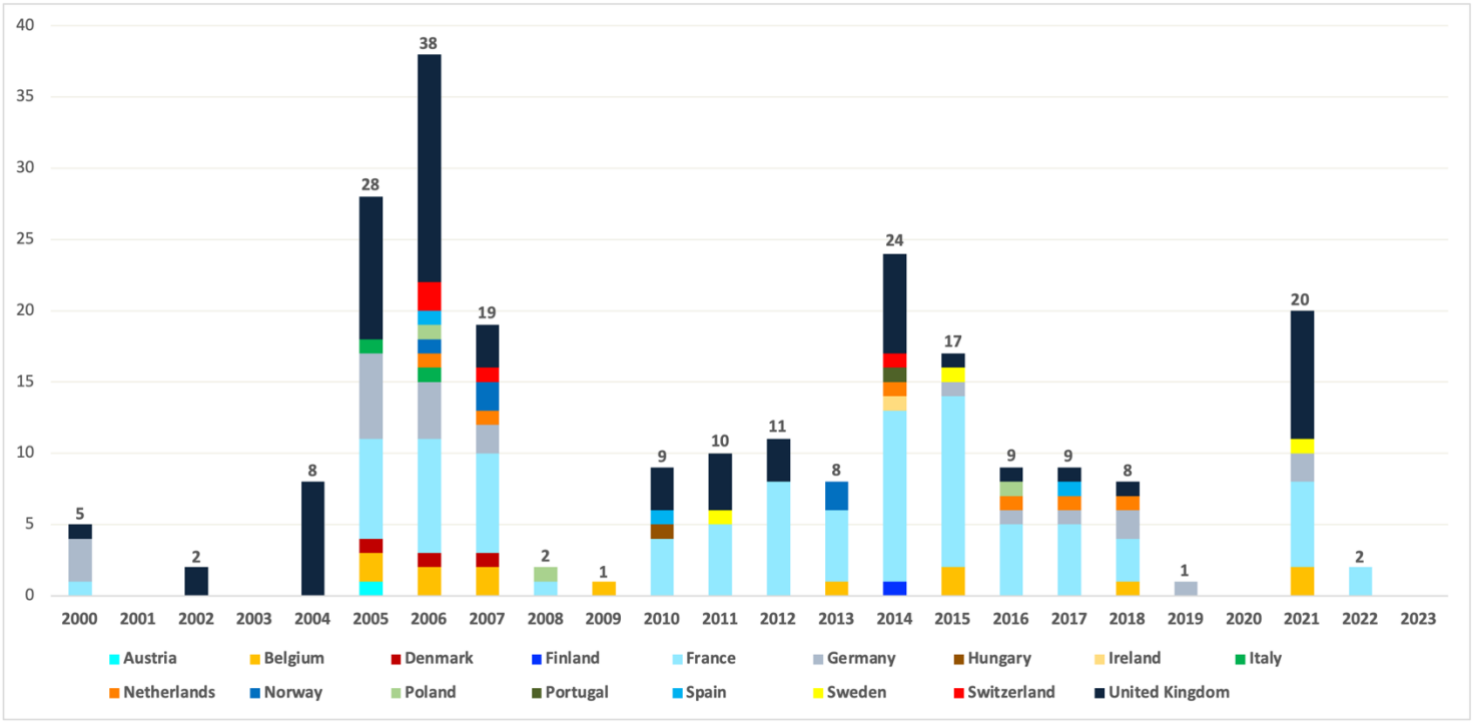


Figure 4. Number of VC-backed IPOs in the EEA and EFTA countries during 2000–2023, by IPO country of the listed company.

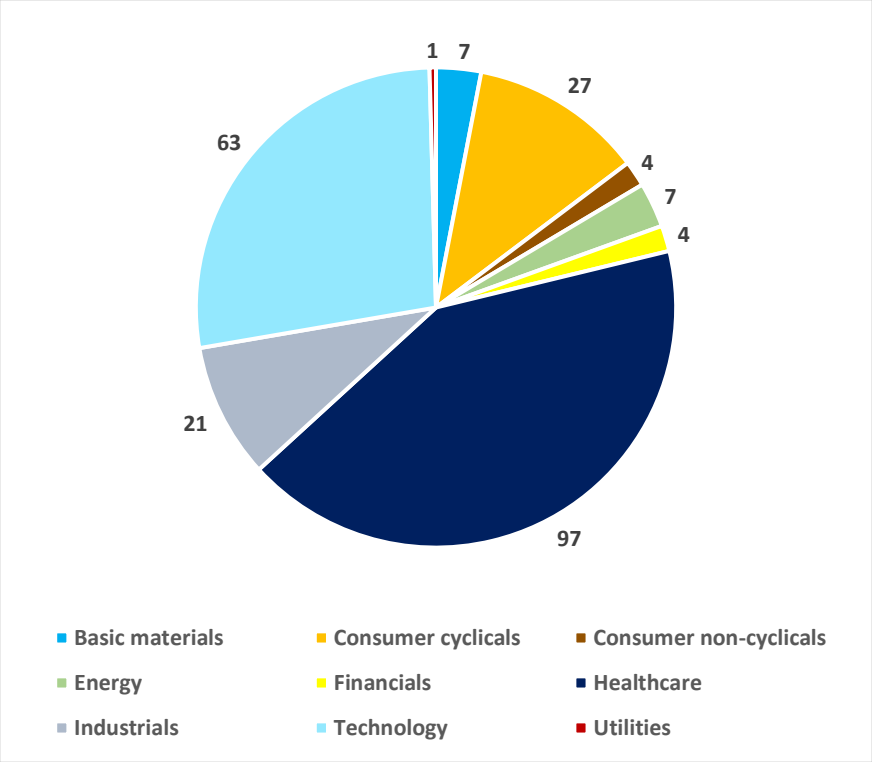


Figure 5. Distribution of TRBC economic sectors of the VC-backed IPO companies in the EEA and EFTA countries during 2000–2023.

Figure 5 illustrates the distribution of economic sectors among all VC-backed IPO companies in the dataset. The sector classification is based on the Refinitiv Business Classification (TRBC) system. The chart shows that most VC-backed IPO companies in the dataset are in the healthcare sector. Among the top three sectors, technology and consumer cyclicals industries are the second- and third-most frequent, respectively. Approximately 81% of the dataset's VC-backed companies fall into these three sectors, highlighting their importance relative to the other sectors.

6.3 Variables in the thesis

This section describes the relevant variables in this thesis. Additionally, Sections from 6.3.1 to 6.3.3 provide comprehensive descriptive statistics for them. A summary table of the variable descriptions can be found in Appendix 1.

IPO proceeds, USD captures the dollar value of the total gross proceeds company generates in an IPO, including sold overallotment shares. Proceeds is a widely used success metric in the IPO literature (Gulati and Higgins, 2003; Wu, Li, and Li, 2013; Lehnertz, Plagmann, and Lutz, 2022). In addition, Botazzi and Da Rin (2002) use a modified version of proceeds by scaling total raised IPO funds by the company's total assets. Equation 1 presents the mathematics of the *Scaled IPO proceeds, USD* variable in this thesis.

$$Scaled\ IPO\ proceeds, USD = \frac{IPO\ proceeds, USD}{Total\ assets\ before\ IPO, USD} \quad (1)$$

However, this thesis uses a binary *Scaled IPO proceeds* variable to measure whether an IPO is successful. The binary *Scaled IPO proceeds* variable indicates successful IPO (1) if the value of the *Scaled IPO proceeds, USD* is above the sample median, and unsuccessful IPO (0) otherwise. Scaling proceeds is necessary to control for the differences in company size. Without scaling, the proceeds simply reflect how big the company is.

A binary *IPO underpricing* is the second IPO success variable in this thesis. Lehnertz et al. (2022) also use underpricing as a success measure. *IPO underpricing, %* is the percentage difference between the closing price per share on the first day of trading and the IPO offer price. Equation 2 presents the mathematics of the variable.

$$IPO\ underpricing, \% = \frac{(First\ day\ closing\ price - Offer\ price)}{Offer\ price} \quad (2)$$

As discussed in Section 5.1, the *market feedback hypothesis*, the *signaling hypothesis*, the *winner's curse hypothesis*, the *bandwagon hypothesis*, the *investment banker's monopsony power hypothesis*, and the *ownership dispersion hypothesis* all suggest that companies and underwriters deliberately underprice their IPOs. If underpricing is an intentional process in the markets, as these six hypotheses suggest, it is reasonable to

consider in this thesis that only positive underpricing is a signal of a successful IPO. In addition, Bessler and Seim (2012) show that European VC-backed companies have high positive IPO underpricing, suggesting that IPOs are generally underpriced in Europe. Therefore, the binary *IPO underpricing* variable indicates a successful IPO (1) if underpricing is positive (greater than zero), and an unsuccessful IPO (0) otherwise.

Round equity, USD and *ln(Round equity, USD)* are the foundational variables for the purpose of this thesis. *Round equity, USD* captures the size, in dollars, of the largest disclosed venture capital equity financing round since the founding of a VC-backed IPO company. *ln(Round equity, USD)* is the logarithm of *Round equity, USD*. In this thesis, the empirical sections favor *ln(Round equity, USD)* variable over *Round equity, USD* to compress extreme values, reduce skewness, and increase economic interpretability. The log transformation enables analysis of proportional (%) rather than absolute (USD) changes, which makes more sense given the purpose of this thesis. For example, an increase of 10 million USD is not economically equivalent across the VC equity financing round distribution. Moving from a disclosed venture capital equity financing round of 10 million USD to 20 million USD represents a larger relative change than moving from a disclosed venture capital equity financing round of 200 million USD to 210 million USD.

Total assets before IPO, USD is defined similarly to Lehnertz, Plagmann, and Lutz (2022). According to the authors, this size factor is used frequently in the IPO performance literature. It captures the dollar value of total assets before a VC-backed IPO. *ln(Total assets before IPO, USD)* is the logarithm of *Total assets before IPO, USD*. The thesis's empirical sections favor the log transformed variable to compress extreme values, reduce skewness, and increase economic interpretability.

Age at IPO is also defined similarly to Lehnertz et al. (2022). The variable captures the number of years between the IPO year and the founding year of a VC-backed company. In other words, it is the difference between the IPO year and the founding year of a company. It can serve as a proxy for the organizational maturity of a VC-backed company.

VC rounds before IPO is the number of venture capital rounds a VC-backed company has completed before the IPO. This variable is employed in the thesis as a proxy for the historical equity financing intensity of VC-backed companies (how frequently new rounds of equity are raised).

Profitability is defined identically to Lehnertz et al. (2022). Based on net income after taxes before the IPO, *Profitability* is always 1 or 0. A company is profitable (1) if its net income after taxes before the IPO is greater than zero, and unprofitable (0) otherwise.

France, *Germany, United Kingdom,* and *Rest of the EEA/EFTA* are binary country variables indicating whether a VC-backed company's IPO country (exchange nation) is the specified country. The variable is 1 if a company's country matches with the variable, and 0 otherwise. *France*, *Germany,* and *United Kingdom* are ungrouped individual variables, as the analysis in Section 6.2 demonstrates their high relevance in the dataset. The *Rest of the EEA/EFTA* variable includes Austria, Belgium, Denmark, Finland, Hungary, Ireland, Italy, Netherlands, Norway, Poland, Portugal, Spain, Sweden, and Switzerland.

Consumer cyclicals, *Healthcare, Technology,* and *Other sectors* are binary economic sector variables indicating whether a VC-backed company's economic sector is the specified sector. The variable is 1 if a company's sector matches with the variable, and 0 otherwise. *Consumer cyclicals*, *Healthcare,* and *Technology* are ungrouped individual variables, as the analysis in Section 6.2 demonstrates their high relevance in the dataset. The *Other sectors* variable includes the following economic sectors: basic materials, consumer noncyclicals, energy, financials, industrials, and utilities.

IPO year 2000–2009, *IPO year 2010–2019,* and *IPO year 2020–2023* are IPO year (period) variables indicating whether a VC-backed company's IPO took place during the specified period. The variable is 1 if a company's IPO year matches with the variable, and 0 otherwise. As discussed in Section 1.1, IPO activity is subject to year-to-year fluctuations and is influenced by factors such as macroeconomic conditions, political developments,

and investor sentiment. These IPO year periods are used instead of introducing many one-year dummies for practical reasons, given the thesis's small dataset (N = 231).

6.3.1 Descriptive statistics (full sample)

Table 1. Descriptive statistics. All VC-backed IPOs.

Variable	N	Mean	Standard deviation	Median	Minimum	Maximum	Range	Skewness	Kurtosis
IPO proceeds, USD	231	102 592 147	242 895 077	34 274 048	8 421	2 195 432 009	2 195 423 588	5.919	43.297
Scaled IPO proceeds, USD	231	3.144	4.673	1.790	0.003	38.740	38.737	4.383	23.820
IPO underpricing, %	231	0.0350	0.1411	0.0053	-0.4154	0.8958	1.3112	1.661	8.836
Round equity, USD	231	28 721 027	65 713 312	12 189 372	11 536	623 781 733	623 770 197	6.547	49.7792
ln(Round equity, USD)	231	16.185	1.554	16.316	9.353	20.251	10.898	-0.989	3.080
Total assets before IPO, USD	231	116 289 979	519 849 574	17 054 211	457 344	7 056 932 861	7 056 475 516	10.943	140.330
ln(Total assets before IPO, USD)	231	16.821	1.612	16.652	13.033	22.677	9.644	0.599	0.932
Age at IPO	231	10.338	9.396	8	1	81	80	4.827	30.196
VC rounds before IPO	231	3.506	2.565	3	1	19	18	1.877	6.583
Scaled IPO proceeds	231	0.498	–	–	–	–	–	–	–
IPO underpricing	231	0.584	–	–	–	–	–	–	–
Profitability	231	0.290	–	–	–	–	–	–	–

Table 1 shows descriptive statistics for all 231 VC-backed IPO companies in the EEA/EFTA countries from 2000–2023. Binary country, economic sector and IPO year variables are not included in the statistics. On average, a VC-backed company generates approximately 102.59 million USD in IPO proceeds, and the proceeds raised are, on average, 3.14 times the issuing company's total assets before the IPO, as measured by *Scaled IPO proceeds, USD*. Mean IPO underpricing on the first day of trading is 3.50%.

The average size of the largest VC financing round that a VC-backed company has secured before its IPO is approximately 28.72 million USD, and the median is 12.19 million USD. The largest VC round is 623.78 million USD. The skewness of the *Round equity, USD* variable is 6.55, which justifies the log transformation. The skewness of *Total asset before IPO, USD* is 10.94, which is why the log transformation is applied also to this variable. The average size of a VC-backed company's total asset before IPO is 116.29 million USD.

Moreover, the average age of a VC-backed company is 10.34 years at the time of its IPO, and a VC-backed company raises, on average, 3.51 VC financing rounds before its IPO.

6.3.2 Descriptive statistics (Scaled IPO proceeds)

Table 2 reports statistics for successful VC-backed IPOs (where *Scaled IPO proceeds* equals 1), and Table 3 reports statistics for unsuccessful VC-backed IPOs (where *Scaled IPO proceeds* equals 0). The number of successful IPOs is 115, and the number of unsuccessful IPOs is 116.

Table 2. Descriptive statistics. Successful VC-backed IPOs (*Scaled IPO proceeds* = 1).

Variable	N	Mean	Standard deviation	Median	Minimum	Maximum	Range	Skewness	Kurtosis
IPO proceeds, USD	115	121 469 378	297 428 488	41 622 632	6 939 625	2 195 432 009	2 188 492 384	5.701	36.094
Scaled IPO proceeds, USD	115	5.396	5.801	3.532	1.801	38.740	36.939	3.487	13.745
IPO underpricing, %	115	0.0446	0.1356	0.0130	-0.2947	0.7368	1.0316	1.626	6.012
Round equity, USD	115	31 021 172	78 168 992	15 435 468	104 094	623 781 733	623 677 639	6.549	45.654
ln(Round equity, USD)	115	16.350	1.327	16.552	11.553	20.251	8.698	-0.402	1.930
Total assets before IPO, USD	115	36 622 611	100 905 739	11 898 427	457 344	839 817 881	839 360 536	6.353	44.689
ln(Total assets before IPO, USD)	115	16.304	1.397	16.292	13.033	20.549	7.516	0.189	0.637
Age at IPO	115	8.835	7.353	8	1	72	71	5.803	47.809
VC rounds before IPO	115	3.661	2.714	3	1	19	18	1.984	7.742
Scaled IPO proceeds	115	1.000	–	–	–	–	–	–	–
IPO underpricing	115	0.609	–	–	–	–	–	–	–
Profitability	115	0.157	–	–	–	–	–	–	–

Table 2 shows that an average VC-backed company completing a successful IPO generates approximately 121.47 million USD in IPO proceeds, and the proceeds raised are, on average, 5.40 times the issuing company's total assets before the IPO. The average size of the largest VC financing round that a successful VC-backed company has secured before its IPO is approximately 31.02 million USD. In addition, the average age of a successful VC-backed IPO company is 8.84 years at the time of the IPO, and on average, a successful company raises 3.66 VC financing rounds before its IPO.

On the other hand, Table 3 shows that an average VC-backed company completing an unsuccessful IPO generates approximately 83.88 million USD in IPO proceeds, and the proceeds raised are, on average, 0.91 times the issuing company's total assets before the IPO. The average size of the largest VC financing round before IPO is approximately 26.44 million USD. In addition, the average age of an unsuccessful VC-backed IPO company is 11.83 years at the time of the IPO, and on average, an unsuccessful company raises 3.35 VC financing rounds before its IPO.

Table 3. Descriptive statistics. Unsuccessful VC-backed IPOs (*Scaled IPO proceeds* = 0).

Variable	N	Mean	Standard deviation	Median	Minimum	Maximum	Range	Skewness	Kurtosis
IPO proceeds, USD	116	83 877 650	172 023 118	22 322 412	8 421	1 114 771 462	1 114 763 041	3.746	16.241
Scaled IPO proceeds, USD	116	0.912	0.477	0.907	0.003	1.790	1.787	-0.030	-1.098
IPO underpricing, %	116	0.0255	0.1463	0.0045	-0.4154	0.8958	1.3112	1.750	11.438
Round equity, USD	116	26 440 711	50 682 202	10 333 604	11 536	372 909 936	372 898 400	4.786	26.550
ln(Round equity, USD)	116	16.023	1.740	16.151	9.353	19.737	10.384	-1.123	2.769
Total assets before IPO, USD	116	195 270 560	719 551 711	26 947 147	705 168	7 056 932 861	7 056 227 691	7.979	73.387
ln(Total assets before IPO, USD)	116	17.333	1.652	17.109	13.466	22.677	9.211	0.740	0.554
Age at IPO	116	11.828	10.885	10	1	81	80	4.264	22.727
VC rounds before IPO	116	3.353	2.411	3	1	15	14	1.704	4.666
Scaled IPO proceeds	116	0.000	–	–	–	–	–	–	–
IPO underpricing	116	0.560	–	–	–	–	–	–	–
Profitability	116	0.422	–	–	–	–	–	–	–

To summarize the descriptive statistics sorted by Scaled *IPO proceeds*, it appears that, on average, a successful VC-backed IPO company has secured a larger equity financing round before its IPO compared to a company that completes an unsuccessful IPO. Moreover, a successful IPO company is, on average, younger at the time of its IPO and has raised more financing rounds than a company with an unsuccessful IPO. The descriptive statistics results are in line with the alternative hypothesis (H_1) of *Scaled IPO proceeds*.

6.3.3 Descriptive statistics (IPO underpricing)

Table 4 reports statistics for successful VC-backed IPOs (where *IPO underpricing* equals 1), and Table 5 reports statistics for unsuccessful VC-backed IPOs (where *IPO underpricing* equals 0). The number of successful IPOs is 135, and the number of unsuccessful IPOs is 96.

Table 4. Descriptive statistics. Successful VC-backed IPOs (*IPO underpricing* = 1).

Variable	N	Mean	Standard deviation	Median	Minimum	Maximum	Range	Skewness	Kurtosis
IPO proceeds, USD	135	101 804 538	239 819 741	34 116 335	873 204	2 195 432 009	2 194 558 805	5.967	45.496
Scaled IPO proceeds, USD	135	3.450	5.261	1.839	0.194	38.740	38.546	3.987	19.116
IPO underpricing, %	135	0.1034	0.1342	0.0559	0.0001	0.8958	0.8958	2.863	11.620
Round equity, USD	135	26 755 574	58 771 630	11 941 750	104 094	551 010 385	550 906 291	6.481	50.910
ln(Round equity, USD)	135	16.210	1.380	16.296	11.553	20.127	8.574	-0.458	1.374
Total assets before IPO, USD	135	95 573 240	286 088 807	15 729 095	457 344	2 085 706 944	2 085 249 600	4.905	26.360
ln(Total assets before IPO, USD)	135	16.762	1.604	16.571	13.033	21.458	8.425	0.552	0.978
Age at IPO	135	10.259	8.299	9	1	72	71	4.153	25.916
VC rounds before IPO	135	3.526	2.769	3	1	19	18	2.169	7.981
Scaled IPO proceeds	135	0.519	–	–	–	–	–	–	–
IPO underpricing	135	1.000	–	–	–	–	–	–	–
Profitability	135	0.281	–	–	–	–	–	–	–

Table 4 shows that a VC-backed company completing a successful IPO has, on average, IPO underpricing of 10.34%. The average size of the largest VC financing round that a successful VC-backed company has secured before its IPO is approximately 26.76 million USD. In addition, the average age of a successful VC-backed IPO company is 10.26 years at the time of the IPO, and on average, a successful company raises 3.53 VC financing rounds before its IPO.

Table 5 shows that a VC-backed company completing an unsuccessful IPO has, on average, IPO underpricing of –6.11%. The average size of the largest VC financing round secured before IPO is approximately 31.48 million USD. In addition, the average age of

an unsuccessful VC-backed IPO company is 10.45 years at the time of the IPO, and on average, an unsuccessful company raises 3.48 VC financing rounds before its IPO.

Table 5. Descriptive statistics. Unsuccessful VC-backed IPOs (*IPO underpricing* = 0).

Variable	N	Mean	Standard deviation	Median	Minimum	Maximum	Range	Skewness	Kurtosis
IPO proceeds, USD	96	103 699 721	248 416 725	35 191 910	8 421	2 069 536 421	2 069 527 999	5.942	42.699
Scaled IPO proceeds, USD	96	2.714	3.676	1.637	0.003	30.795	30.792	5.111	35.969
IPO underpricing, %	96	-0.0611	0.0826	-0.0333	-0.4154	0.0000	0.4154	-2.199	5.547
Round equity, USD	96	31 484 945	74 628 215	13 919 900	11 536	623 781 733	623 770 197	6.422	46.362
ln(Round equity, USD)	96	16.151	1.777	16.445	9.353	20.251	10.898	-1.293	3.483
Total assets before IPO, USD	96	145 422 894	733 047 955	20 600 298	705 168	7 056 932 8601	7 056 227 692	9.049	85.565
ln(Total assets before IPO, USD)	96	16.903	1.629	16.841	13.466	22.677	9.211	0.671	0.960
Age at IPO	96	10.448	10.798	8	1	81	80	5.108	30.120
VC rounds before IPO	96	3.479	2.262	3	1	12	11	1.045	1.091
Scaled IPO proceeds	96	0.469	–	–	–	–	–	–	–
IPO underpricing	96	0.000	–	–	–	–	–	–	–
Profitability	96	0.302	–	–	–	–	–	–	–

To summarize the descriptive statistics sorted by *IPO underpricing*, it appears that, on average, a successful VC-backed IPO company has secured a smaller largest VC round before its IPO compared to a company that completes an unsuccessful IPO. Moreover, a successful IPO company is, on average, younger at the time of its IPO and has raised more financing rounds than a company with an unsuccessful IPO. The descriptive statistics results are in line with the alternative hypothesis (H_1) of *IPO underpricing*.

7 Methodology

To statistically estimate whether the size of the largest disclosed VC equity financing round is associated with IPO success, this thesis specifies two binary classification models with *ln(Round equity, USD)* as the variable of interest. The primary analyses use traditional logistic regression (logit), while robustness checks are performed with random forest (RF) and extreme gradient boosting (XGBoost) machine learning algorithms, strongly inspired by the methodology of Sapkota (2025). Section 7.1 introduces model variables, Section 7.2 describes the logistic regression models, and Section 7.3 focuses on the machine learning algorithms.

7.1 Model variables

Table 6. Variable composition of the binary classification models.

Binary classification model 1	Binary classification model 2
Dependent variable:	**Dependent variable:**
Scaled IPO proceeds	IPO underpricing
Independent (predictor) variables:	**Independent (predictor) variables:**
ln(Round equity, USD)	ln(Round equity, USD)
Age at IPO	ln(Total assets before IPO, USD)
VC rounds before IPO	Age at IPO
Profitability	VC rounds before IPO
	Profitability
Country fixed effects (omitted category: *Rest of the EEA/EFTA*):	**Country fixed effects (omitted category: *Rest of the EEA/EFTA*):**
France	France
Germany	Germany
United Kingdom	United Kingdom
Sector fixed effects (omitted category: *Other sectors*):	**Sector fixed effects (omitted category: *Other sectors*):**
Consumer cyclicals	Consumer cyclicals
Healthcare	Healthcare
Technology	Technology
Year fixed effects (omitted category: *IPO year 2000–2009*):	**Year fixed effects (omitted category: *IPO year 2000–2009*):**
IPO Year 2010–2019	IPO Year 2010–2019
IPO Year 2020–2023	IPO Year 2020–2023

This thesis uses two binary dependent IPO success variables: *Scaled IPO proceeds* and *IPO underpricing*. The main independent (predictor) variable of interest is *ln(Round equity, USD)*. Models of the thesis also include predictor variables *ln(Total assets before IPO, USD)*, *Age at IPO*, *VC rounds before IPO*, and *Profitability*, as well as country, sector, and year fixed effects. Country fixed effects are implemented using dummy variables for *France*, *Germany*, and *United Kingdom*, with *Rest of the EEA/EFTA* as the omitted (baseline) category. Sector fixed effects are implemented using dummies for *Consumer cyclicals*, *Healthcare*, and *Technology*, with *Other sectors* as the omitted (baseline) category. Year fixed effects are implemented using dummies for *IPO year 2010–2019* and *IPO year 2020–2023*, with *IPO year 2000–2009* as the omitted (baseline) category. Omissions are necessary to identify an intercept in statistical models. Table 6 underlines the variable composition of the two binary classification models with fixed effects used in this thesis.

Table 7. Pearson and point-biserial correlation matrix.

	Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	**Scaled IPO proceeds**	1.000						
2	**IPO underpricing**	0.049	1.000					
3	**ln(Round equity, USD)**	0.106	0.019	1.000				
4	**ln(Total assets before IPO, USD)**	-0.320	-0.043	0.460	1.000			
5	**Age at IPO**	-0.160	-0.010	-0.022	0.176	1.000		
6	**VC rounds before IPO**	0.060	0.009	0.339	0.008	-0.008	1.000	
7	**Profitability**	-0.293	-0.022	-0.182	0.297	0.201	-0.220	1.000

Note: The Pearson correlation measure is used for pairs of continuous and/or discrete variables. The point-biserial correlation measure is used for pairs consisting of a continuous or discrete variable and a binary variable. For pairs of binary variables, the phi coefficient is used. The matrix is computed using the `numpy` and `pandas` Python libraries.

Table 7 presents the Pearson and point-biserial correlation matrix for the thesis's dependent and independent variables. The Pearson correlation measure is used for pairs of continuous and/or discrete variables. The point-biserial correlation measure is used for pairs consisting of a continuous or discrete variable and a binary variable. For pairs

of binary variables, the phi coefficient is used. Appendix 2 presents the complete Pearson and point-biserial correlation matrix, including the fixed effect dummy variables.

Table 8. Variance inflator factors (VIFs).

	Scaled IPO proceeds model			IPO underpricing model	
	Variable	**VIF**		**Variable**	**VIF**
1	**ln(Round Equity, USD)**	1.199	1	**ln(Round Equity, USD)**	1.687
2	**Age at IPO**	1.085	2	**ln(Total Assets Before IPO, USD)**	1.909
3	**VC rounds before IPO**	1.176	3	**Age at IPO**	1.118
4	**Profitability**	1.349	4	**VC rounds before IPO**	1.192
			5	**Profitability**	1.517

Note:
VIF = 1 → No multicollinearity.
1 < VIF ≤ 5 → Moderate multicollinearity.
5 < VIF ≤ 10 → Higher multicollinearity. Potential problem.
VIF > 10 → Serious multicollinearity. Remove or adjust the variable.

VIFs are estimated using the `numpy`, `pandas`, and `statsmodels` Python library, with the function `statsmodels.stats.outliers_influence.variance_inflation_factor`.

Table 8 shows variance inflator factor (VIF) measures for the thesis's independent variables. The VIFs are computed including the fixed effects. Based on the VIF tables for the *Scaled IPO proceeds* and *IPO underpricing* models, it appears that there is only moderate multicollinearity between the predictor variables. Thus, variable adjustments or removals are not required.

7.2 Logistic regression (logit) models

The traditional logistic regression (logit) model is a frequently used statistical model for binary classification (Ranta, 2023, Chapter 8.2.). The *Scaled IPO proceeds* and *IPO underpricing* dependent variables are binary indicators of IPO success, and the logit models in the thesis estimate the log-odds that these outcome variables equal 1 as a function of predictor variables. In this thesis, the log-odds are defined as the logarithm of the odds of a successful IPO, measured by either *Scaled IPO proceeds* or *IPO underpricing*.

Based on Sapkota (2025, p. 12), a linear combination of independent variables and their coefficients with the intercept (the right-hand side of a logit model) equals the log-odds of the outcome, which is then transformed into a probability between 0 and 1 using the logistic function. The logit coefficient estimates in this thesis represent changes in the log-odds of a successful IPO for a unit change in the specified predictor.

7.2.1 Logit (Scaled IPO proceeds)

The *Scaled IPO proceeds* logistic regression model is presented in Equation 3.

$$\begin{aligned}\log\left(\frac{P(Scaled\ IPO\ proceeds_i\ =\ 1)}{1\ -\ P(Scaled\ IPO\ proceeds_i\ =\ 1)}\right) &= Z_{scaled\ proceeds_i} \\ &= \beta_1^{scaled\ proceeds} \ln(Round\ equity, USD)_i \\ &+\ \beta_2^{scaled\ proceeds}\ Age\ at\ IPO_i \\ &+\ \beta_3^{scaled\ proceeds}\ VC\ rounds\ before\ IPO_i \\ &+\ \beta_4^{scaled\ proceeds}\ Profitability_i + \gamma_{country_i} + \delta_{sector_i} + \theta_{year_i},\end{aligned} \tag{3}$$

where the $Z_{scaled\ proceeds_i}$ is the linear combination of independent variables with their coefficients and fixed effect intercepts $\gamma_{country_i}$, δ_{sector_i}, and θ_{year_i}.

7.2.2 Logit (IPO underpricing)

The *IPO underpricing* logistic regression model is presented in Equation 4.

$$\begin{aligned}\log\left(\frac{P(IPO\ underpricing_i\ =\ 1)}{1\ -\ P(IPO\ underpricing_i\ =\ 1)}\right) &= Z_{underpricing_i} \\ &= \beta_1^{underpricing} \ln(Round\ equity, USD)_i \\ &+ \beta_2^{underpricing} \ln(Total\ assets\ before\ IPO, USD)_i \\ &+\ \beta_3^{underpricing}\ Age\ at\ IPO_i \\ &+\ \beta_4^{underpricing}\ VC\ rounds\ before\ IPO_i \\ &+\ \beta_5^{underpricing}\ Profitability_i + \gamma_{country_i} + \delta_{sector_i} + \theta_{year_i},\end{aligned} \tag{4}$$

where the $Z_{underpricing_i}$ is the linear combination of independent variables with their coefficients and fixed effect intercepts $\gamma_{country_i}$, δ_{sector_i}, and θ_{year_i}.

7.3 Machine learning algorithms

This thesis uses random forest (RF) and extreme gradient boosting (XGBoost) machine learning algorithms to check robustness of the logit estimations. These algorithms are ensemble methods that combine weak estimators into a single strong model for predicting a binary class. Weak estimators in ensemble methods are commonly decision trees, which split observations during training based on feature values, and each new leaf node created by a split represents a prediction (Ranta, 2023, Chapter 8.6.2.).

The random forest and XGBoost methods use the same data as the logistic regression but generate and train multiple different weak estimators through different mechanisms. These mechanisms are discussed in Sections 7.3.1 and 7.3.2. By combining a diverse set of weak estimators, the ensemble methods can create a strong predictive model with high accuracy (Ranta, 2023, Chapter 8.6.).

In this thesis, the random forest and XGBoost algorithms are trained using train–test splits of 60/40, 70/30, 80/20, and 90/10 to provide comprehensive evidence of how the models perform across different training proportions. The hyperparameters are optimized for model efficiency using 5-fold cross-validation on the training set only, to control the learning process for predicting IPO success (Ranta, 2023, Chapter 8.9.1.).

7.3.1 Random Forest (RF)

According to Sapkota (2025, p. 19), the random forest machine learning algorithm can handle complex interactions and nonlinear relationships between variables. On its own, moderate multicollinearity among the predictor variables (as demonstrated in Table 8) does not provide sufficient justification for using ensemble methods. However, the random forest is used as a robustness check because it can capture potential

nonlinearities that logit models (which assume linearity in the predictors) may miss. In addition, the random forest algorithm reduces variance and improves accuracy by combining many decision trees, thereby effectively resisting overfitting in small- to moderate-sized datasets (Sapkota, 2025, p. 23). The dataset of this thesis is rather small (N = 231), which suggests that the random forest method could be suitable.

The random forest algorithm employs both random feature selection and bagging. Random feature selection involves selecting a random subset of features for a weak estimator at each leaf node (Ranta, 2023, Chapter 8.6.2.2.). Bagging involves creating bootstrap samples (subsamples of training data) on which the weak estimators are trained in parallel (Ranta, 2023, Chapter 8.6.2.1.).

7.3.2 eXtreme Gradient Boosting (XGBoost)

The XGBoost machine learning algorithm has a similar concept to its precursor, gradient boosting, which has empirically demonstrated strong effectiveness in classification problems. Since its introduction in 2014, XGBoost has rapidly gained popularity and has won many competitions in the machine learning community. (Nielsen, 2016, pp. 1–2)

The justification for using XGBoost as a robustness check lies in its ensemble learning characteristics. As an ensemble model, XGBoost can capture potential nonlinear effects (like the random forest method) that logit models may miss. Moreover, XGBoost provides strong out-of-sample predictive performance, well suited for robustness analysis.

The XGBoost algorithm employs boosting ensemble method principles, where weak estimators are trained sequentially rather than in parallel (Ranta, 2023, Chapter 8.6.2.3.). Each weak estimator is trained to minimize a loss function using gradient-based optimization, thereby prioritizing observations with higher prediction errors (IBM, n.d.). In addition, based on (IBM, n.d.), XGBoost employs regularization as a part of the learning process, and the final prediction is computed as a weighted sum of all weak estimators.

8 Results and discussions

Section 8 presents and analyzes the results of the thesis's empirical tests. Sections 8.1 and 8.2 discuss coefficient estimates of the logit models, and Sections 8.3 to 8.9 focus on the robustness checks with the random forest and XGBoost machine learning algorithms. The robustness check sections include model performance assessment, variable importance analysis, and 5-fold cross-validation testing.

8.1 Logit model coefficients (Scaled IPO proceeds)

Table 9 reports the Scaled IPO proceeds logit model coefficient estimates with their statistical significance. The coefficient estimates represent changes in the log-odds of *Scaled IPO proceeds* for a one-unit change in the respective predictor. Based on the findings, only *Profitability* coefficient is statistically significant (at the 1% level). Therefore, the model provides limited insights into the associations between the predictor variables and *Scaled IPO proceeds*.

The logit coefficient estimates are conditional on the other predictor variables and fixed effects. The *ln(Round equity, USD)* coefficient suggests a positive relationship between the size of the largest VC equity round and *Scaled IPO proceeds*, but it is not statistically different from zero, so no true conclusions of the association can be drawn. However, if analyzing solely the positive sign of the coefficient, VC-backed companies that raise larger all-time VC equity rounds prior to the IPO could have higher odds of generating above-median scaled IPO proceeds. The coefficient of 0.1010, implies that a one-unit increase in *ln(Round equity, USD)* is associated with approximately 10.6% ($e^{0.1010}$) higher odds of above-median scaled IPO proceeds. More intuitively, a doubling of *Round equity, USD* is associated with approximately 7.3% ($e^{0.1010 \times \ln(2)}$) increase in the odds of generating above-median scaled IPO proceeds. Based on the average marginal effect (AME), doubling *Round equity, USD* increases the probability of above-median scaled IPO proceeds by approximately 1.5 percentage points ($AME \times \ln(2)$). The positive sign of the statistically insignificant coefficient is in line with the expected direction of the

alternative Scaled IPO proceeds hypothesis (H_1), which predicts a positive conditional association between the VC round size and *Scaled IPO proceeds*.

Table 9. Scaled IPO proceeds logit coefficient estimates.

Variable	Estimate	Standard error	z-value	Pr(>|z|)
(Intercept)	-1.0864	1.6791	-0.6470	0.5176
ln(Round equity, USD)	0.1010	0.1009	1.0008	0.3169
Age at IPO	-0.0268	0.0206	-1.3005	0.1934
VC rounds before IPO	-0.0080	0.0592	-0.1353	0.8924
Profitability	-1.1082	0.3707	-2.9890	0.0028 **
Country fixed effects	Included	Included	Included	Included
Sector fixed effects	Included	Included	Included	Included
Year fixed effects	Included	Included	Included	Included

Note: This table presents predictor variable coefficient estimates and statistical significance levels based on two-sided z-test for the Scaled IPO proceeds logit model estimating whether IPO is successful (success if *Scaled IPO proceeds* = 1, and no IPO success otherwise). Variable of interest is the *ln(Round equity, USD)*, which is logarithm of the largest disclosed venture capital equity financing round before IPO. Predictor variable *Age at IPO* is the company's age at the time of IPO, VC *round before IPO* is the number of completed VC rounds before IPO, and *Profitability* is dummy variable (1/0) indicating whether company is profitable at the time of IPO. The logit model includes country fixed effects ($\gamma_{country_i}$), sector fixed effects (δ_{sector_i}), and year fixed effects (θ_{year_i}). Data period is 2000–2023.

The logit model is the following:
$\log\left(\frac{P(Scaled\ IPO\ proceeds_i = 1)}{1 - P(Scaled\ IPO\ proceeds_i = 1)}\right) = Z_{scaled\ proceeds_i} = \beta_1^{scaled\ proceeds} \ln(Round\ equity, USD)_i + \beta_2^{scaled\ proceeds}\ Age\ at\ IPO_i + \beta_3^{scaled\ proceeds}\ VC\ rounds\ before\ IPO_i + \beta_4^{scaled\ proceeds}\ Profitability_i + \gamma_{country_i} + \delta_{sector_i} + \theta_{year_i}$.

The logistic function $\sigma(Z)$ and the probability P that *Scaled IPO proceeds* occur is: $P_{scaled\ proceeds_i} = \sigma\left(Z_{scaled\ proceeds_i}\right) = \frac{1}{1+e^{-Z_{scaled\ proceeds_i}}}$.

The intercept is the log-odds of *Scaled IPO proceeds* for the omitted country, sector, and year fixed effects categories when all other predictor variables equal zero. Pr(>|z|) is the p-value of a two-sided z-test of the Scaled IPO proceeds null hypothesis (H_0) that the coefficient is zero. Logit estimations are performed using the `pandas`, and `statsmodels` Python libraries.

Country fixed effects omitted category: *Rest of the EEA/EFTA*
Sector fixed effects omitted category: *Other sectors*
Year fixed effects omitted category: *IPO year 2000–2009*
Number of observations: 231
AME of *ln(Round equity, USD)*: 0.0218
$R^2_{McFadden}$: 0.1035

***** significant at the 5% level.
****** significant at the 1% level.
******* significant at the 0.1% level.

$R^2_{McFadden}$ of 0.1035 for the Scaled IPO proceeds logit model indicates modest model fit. This suggests that the logit model, with its predictor variables and fixed effects, improves the model fit by 10.35% relative to an intercept-only model. Additionally, the confusion matrix in Appendix 4 suggest that the Scaled IPO proceeds logit model can

predict IPO success with moderate accuracy. The model correctly predicts IPO success, as measured by *Scaled IPO proceeds*, in 91 out of 115 cases (true positive rate of 79.13%), and correctly predicts no IPO success in 70 out of 116 cases (false positive rate of 60.34%). However, the logit model fails to predict IPO success correctly 24 times out of 115, which means that approximately 20.87% of the predictions are wrong. This is also known as the Type II error (Sapkota, 2025, p. 14).

8.2 Logit model coefficients (IPO underpricing)

Table 10 reports the IPO underpricing logit model coefficient estimates. The coefficient estimates represent changes in the log-odds of *IPO underpricing* for a one-unit change in the respective predictor. Based on the findings, none of the coefficients are statistically different from zero, implying that the logit model does not provide true insights into the associations between the predictor variables and *IPO underpricing*.

The findings cannot demonstrate with statistical significance that the size of the largest VC equity round (conditional on the other predictor variables and fixed effects) is associated with IPO success, measured by *IPO underpricing*. However, if analyzing solely the positive sign of the *ln(Round equity, USD*) coefficient of 0.1056, the findings suggest a positive relationship. The positive sign of the statistically insignificant coefficient is opposite with the expected direction of the alternative IPO underpricing hypothesis (H_1), which predicts a negative conditional association between the VC round size and *IPO underpricing*.

If the positive *ln(Round equity, USD*) coefficient was statistically significant with H_0 and H_1, the contradictory to the direction of the alternative IPO underpricing hypothesis could be explained by stronger demand-based mechanisms driving the IPO underpricing up more than the hypothesized signaling, VC certification, VC monitoring, and staged financing effects reduce it. In other words, while these effects may still reduce IPO underpricing through a larger disclosed VC equity round, the association is not strong enough to demonstrate negative relationship between *IPO underpricing* and larger VC

round size. Under the thesis's definition of *IPO underpricing* (underpricing is greater than zero), the hypothesized negative association essentially implies that larger VC rounds would reduce the probability of positive underpricing.

As discussed in Section 5.2, Lehnertz, Plagmann, and Lutz (2022) show that securing a mega-deal generates higher IPO underpricing. The thesis's statistically insignificant coefficient is in line with this evidence if analyzed solely based on its positive sign. The authors explain that mega-deals create additional demand, demonstrated by simultaneously higher IPO offer price revision. The price revision is related to the *market feedback hypothesis* (see Section 5.1), which implies that underwriters tend to adjust the final IPO offer price only partially upward (producing higher IPO underpricing) when overall pre-IPO demand is strong. This could partly explain a hypothetically significant positive *ln(Round equity, USD*) coefficient. However, other explanations are also likely. Behavioral mechanics not covered in this thesis may explain special investor attraction to larger VC rounds, increasing IPO demand and, ultimately, the first day share price.

$R^2_{McFadden}$ of 0.0782 for the IPO underpricing logit model indicates limited model fit. The low $R^2_{McFadden}$ value, compared to the higher $R^2_{McFadden}$ of Scaled IPO proceeds model, implies that pricing outcomes such as *IPO underpricing* are difficult to explain by firm characteristic variables, which is what the thesis's logit model essentially tries to do. In other words, the logit model's predictor variables and fixed effects do not explain much of the variation in whether IPO is successful measured by *IPO underpricing*. This observation is consistent with the existing literature, which demonstrates that underpricing is largely driven by, for example, market conditions, hot issue markets, and underwriter reputation, which the IPO underpricing logit model does not include as predictors (Ritter and Welch, 2002; Ibbotson and Jaffe, 1975; Carter and Manaster, 1990).

Consistent with the evidence about the weak predictive power, the confusion matrix in Appendix 5 reinforces the fact that IPO underpricing logit model has a limited accuracy predicting IPO success. The model fails to predict IPO success correctly 40 times out of

135, which corresponds to a Type II error rate of approximately 29.63%. On the other hand, the model correctly predicts IPO success in 95 out of 135 cases (true positive rate of 70.37%), and correctly predicts no IPO success in 57 out of 96 cases (false positive rate of 59.38%).

Table 10. IPO underpricing logit coefficient estimates.

Variable	Estimate	Standard error	z-value	Pr(>|z|)
(Intercept)	1.0598	1.9821	0.5347	0.5929
ln(Round equity, USD)	0.1056	0.1187	0.8900	0.3735
ln(Total assets before IPO, USD)	-0.1339	0.1225	-1.0931	0.2744
Age at IPO	-0.0023	0.0162	-0.1428	0.8864
VC rounds before IPO	-0.0173	0.0603	-0.2862	0.7747
Profitability	0.0324	0.3865	0.0839	0.9331
Country fixed effects	Included	Included	Included	Included
Sector fixed effects	Included	Included	Included	Included
Year fixed effects	Included	Included	Included	Included

Note: This table presents predictor variable coefficient estimates and statistical significance levels based on two-sided z-test for the IPO underpricing logit model estimating whether IPO is successful (success if *IPO underpricing* = 1, and no IPO success otherwise). Variable of interest is the *ln(Round equity, USD)*, which is logarithm of the largest disclosed venture capital equity financing round before IPO. Predictor variable *ln(Total assets before IPO, USD)* is logarithm of the value of total assets before a IPO, *Age at IPO* is the company's age at the time of IPO, *VC round before IPO* is the number of completed VC rounds before IPO, and *Profitability* is dummy variable (1/0) indicating whether company is profitable at the time of IPO. The logit model includes country fixed effects ($\gamma_{country_i}$), sector fixed effects (δ_{sector_i}), and year fixed effects (θ_{year_i}). Data period is 2000–2023.

The logit model is the following:
$\log\left(\frac{P(IPO\ underpricing_i = 1)}{1 - P(IPO\ underpricing_i = 1)}\right) = Z_{underpricing_i} = \beta_1^{underpricing} \ln(Round\ equity, USD)_i +$
$\beta_2^{underpricing} \ln(Total\ assets\ before\ IPO, USD)_i + \beta_3^{underpricing} Age\ at\ IPO_i + \beta_4^{underpricing} VC\ rounds\ before\ IPO_i +$
$\beta_5^{underpricing} Profitability_i + \gamma_{country_i} + \delta_{sector_i} + \theta_{year_i}$.

The logistic function $\sigma(Z)$ and the probability P that *IPO underpricing* occur is: $P_{underpricing_i} = \sigma\left(Z_{underpricing_i}\right) = \frac{1}{1+e^{-Z_{underpricing_i}}}$.

The intercept is the log-odds of *IPO underpricing* for the omitted country, sector, and year fixed effects categories when all other predictor variables equal zero. Pr(>|z|) is the p-value of a two-sided z-test of the IPO underpricing null hypothesis (H_0) that the coefficient is zero. Logit estimations are performed using the `pandas`, and `statsmodels` Python libraries.

Country fixed effects omitted category: *Rest of the EEA/EFTA*
Sector fixed effects omitted category: *Other sectors*
Year fixed effects omitted category: *IPO year 2000–2009*
Number of observations: 231
AME of *ln(Round equity, USD)*: 0.0230
$R^2_{McFadden}$: 0.0782

***** significant at the 5% level.
****** significant at the 1% level.
******* significant at the 0.1% level.

8.3 Random Forest model performance

Table 11. Performance metrics for the Scaled IPO proceeds and IPO underpricing logit and random forest models.

Model (hyperparameters)	**Train–test**	**Performance metric:**							
		Accuracy	F1 score	AUC	95% CI, AUC	Specificity	Recall	Precision	NPV
Scaled IPO proceeds (logit)	Full sample	0.6970	0.7222	0.6953	[0.6774, 0.7991]	0.6034	0.7913	0.6642	0.7447
Scaled IPO proceeds (RF) (n_estimators = 500, max_features = 0.5)	60% / 40%	0.5957	0.6020	0.6382	[0.6187, 0.6577]	0.5695	0.6225	0.5871	0.6079
Scaled IPO proceeds (RF) (n_estimators = 500, max_features = sqrt)	70% / 30%	0.6333	0.6618	0.6672	[0.6446, 0.6897]	0.5448	0.7219	0.6154	0.6665
Scaled IPO proceeds (RF) (n_estimators = 500, max_features = 0.5)	80% / 20%	0.6319	0.6413	0.6809	[0.6540, 0.7077]	0.5875	0.6783	0.6149	0.6596
Scaled IPO proceeds (RF) (n_estimators = 500, max_features = sqrt)	90% / 10%	0.6347	0.6565	0.6833	[0.6492, 0.7175]	0.5611	0.7083	0.6224	0.6650
IPO underpricing (logit)	Full sample	0.6580	0.7063	0.6841	[0.6613, 0.7939]	0.5938	0.7037	0.7090	0.5876
IPO underpricing (RF) (n_estimators = 200, max_features = sqrt)	60% / 40%	0.5785	0.6796	0.6165	[0.6026, 0.6304]	0.3094	0.7728	0.6080	0.4957
IPO underpricing (RF) (n_estimators = 500, max_features = sqrt)	70% / 30%	0.5762	0.6802	0.6071	[0.5871, 0.6272]	0.2977	0.7732	0.6094	0.4834
IPO underpricing (RF) (n_estimators = 500, max_features = sqrt)	80% / 20%	0.5567	0.6593	0.5954	[0.5677, 0.6230]	0.2850	0.7580	0.5873	0.4721
IPO underpricing (RF) (n_estimators = 200, max_features = sqrt)	90% / 10%	0.5875	0.6789	0.5971	[0.5592, 0.6351]	0.3533	0.7548	0.6243	0.4985

Note: This table reports the full sample model performance metrics for Scaled IPO proceeds and IPO underpricing logit models, and mean metrics for Scaled IPO proceeds random forest models (with train–test splits of 60/40, 70/30, 80/20, and 90/10) and IPO underpricing random forest models (with train–test splits of 60/40, 70/30, 80/20, and 90/10). The Scaled IPO proceeds and IPO underpricing random forest model performance metrics are mean values across 30 repeated stratified splits. The hyperparameters are optimized using 5-fold cross-validation on the training set only. The table in Appendix 3 provides a detailed description of the performance metrics. Appendix 6 includes the confusion matrices. The performance metrics are calculated using the `numpy`, `pandas`, and `sklearn.metrics` Python libraries.

The random forest machine learning algorithm is used as a robustness check for the logit model estimations. Table 11 reports various metrics to assess the performance of the Scaled IPO proceeds and IPO underpricing random forest models across four different train–test proportions (60/40, 70/30, 80/20, and 90/10). Additionally, the table reports the performance metrics of the full sample Scaled IPO proceeds and IPO underpricing logit models. Appendix 3 provides a detailed description of the performance metrics.

Table 11 shows that the mean accuracy of the Scaled IPO proceeds random forest models almost consistently increase as the training proportion increases. With a 90/10 train–test split, the random forest adaptation of the Scaled IPO proceeds model achieves its

best accuracy (63.47%). The IPO underpricing random forest mean accuracies evolve in a nonmonotonic way across train–test proportions. However, as with the Scaled IPO proceeds random forest models, the IPO underpricing random forest accuracy is highest (58.75%) with the 90/10 split, where the training proportion is the largest. Comparing Scaled IPO proceeds and IPO underpricing random forest accuracies across all train–test proportions, the Scaled IPO proceeds is more accurate in every case.

The finding that accuracies are highest with the 90/10 train–test split implies that the Scaled IPO proceeds and IPO underpricing random forest models benefit from the largest training proportion (90%) to better learn existing weak patterns. With the largest training proportion in the rather small dataset of this thesis (N = 231), Scaled IPO proceeds and IPO underpricing models are less sensitive to which data observations end up in the test set, which could potentially distort the mean accuracies.

Based on Table 11, the IPO underpricing random forest model's recall appears to be high, while specificity is very low across all train–test proportions. This implies that the IPO underpricing random forest models tend to correctly predict IPO success, but struggles to predict no IPO success, which is reflected also in low NPV values. This is consistent with the discussions of the IPO underpricing model's weaknesses (see Section 8.2). The differences between recall and specificity are not as severe with the Scaled IPO proceeds random forest models as they are with the IPO underpricing models, and Scaled IPO proceeds models have higher NPVs. However, IPO underpricing random forest models have higher recall, indicating that these models predict IPO success better than the Scaled IPO proceeds random forest models.

F1 score incorporates both recall and precision, being their harmonic mean to assess a model's performance (Ranta, 2023, Chapter 8.10.1.). Based on Table 11, with a 70/30 train–test split, the Scaled IPO proceeds random forest model achieves its highest F1 score (66.18%), while the IPO underpricing random forest model's F1 score is also highest (68.02%) with the 70/30 split. Compared to the random forest mean accuracies across

all train–test proportions, the F1 scores are higher because they combine both recall and precision while ignoring lower specificity. Because the F1 score ignores specificity, the IPO underpricing random forest model's F1 values are higher than those of the Scaled IPO proceeds random forest models.

8.4 XGBoost model performance

Table 12. Performance metrics for the Scaled IPO proceeds and IPO underpricing logit and XGBoost models.

Model (hyperparameters)	**Train–test**	**Performance metric:**							
		Accuracy	F1 score	AUC	95% CI, AUC	Specificity	Recall	Precision	NPV
Scaled IPO proceeds (logit)	Full sample	0.6970	0.7222	0.6953	[0.6774, 0.7991]	0.6034	0.7913	0.6642	0.7447
Scaled IPO proceeds (XGBoost) (n_estimators = 200, max_depth = 2, learning_rate = 0.2)	60% / 40%	0.5806	0.5806	0.6207	[0.6020, 0.6395]	0.5702	0.5913	0.5744	0.5896
Scaled IPO proceeds (XGBoost) (n_estimators = 500, max_depth = 4, learning_rate = 0.1)	70% / 30%	0.5895	0.5913	0.6323	[0.6102, 0.6544]	0.5800	0.5990	0.5888	0.5937
Scaled IPO proceeds (XGBoost) (n_estimators = 1200, max_depth = 3, learning_rate = 0.01)	80% / 20%	0.6206	0.6227	0.6761	[0.6535, 0.6988]	0.5931	0.6493	0.6076	0.6464
Scaled IPO proceeds (XGBoost) (n_estimators = 200, max_depth = 6, learning_rate = 0.2)	90% / 10%	0.6181	0.6197	0.6632	[0.6184, 0.7080]	0.5972	0.6389	0.6227	0.6351
IPO underpricing (logit)	Full sample	0.6580	0.7063	0.6841	[0.6613, 0.7939]	0.5938	0.7037	0.7090	0.5876
IPO underpricing (XGBoost) (n_estimators = 200, max_depth = 2, learning_rate = 0.2)	60% / 40%	0.5918	0.6663	0.6170	[0.6011, 0.6328]	0.4350	0.7049	0.6338	0.5174
IPO underpricing (XGBoost) (n_estimators = 800, max_depth = 6, learning_rate = 0.01)	70% / 30%	0.5948	0.6809	0.6237	[0.6024, 0.6450]	0.3862	0.7423	0.6310	0.5198
IPO underpricing (XGBoost) (n_estimators = 200, max_depth = 6, learning_rate = 0.2)	80% / 20%	0.5780	0.6612	0.6322	[0.6048, 0.6596]	0.3717	0.7309	0.6083	0.5199
IPO underpricing (XGBoost) (n_estimators = 200, max_depth = 6, learning_rate = 0.2)	90% / 10%	0.5958	0.6684	0.6365	[0.6006, 0.6724]	0.4400	0.7071	0.6413	0.5219

Note: This table reports the full sample model performance metrics for Scaled IPO proceeds and IPO underpricing logit models, and mean metrics for Scaled IPO proceeds XGBoost models (with train–test splits of 60/40, 70/30, 80/20, and 90/10) and IPO underpricing XGBoost models (with train–test splits of 60/40, 70/30, 80/20, and 90/10). The Scaled IPO proceeds and IPO underpricing XGBoost model performance metrics are mean values across 30 repeated stratified splits. The hyperparameters are optimized using 5-fold cross-validation on the training set only. The table in Appendix 3 provides a detailed description of the performance metrics. Appendix 7 includes the confusion matrices. The performance metrics are calculated using the `numpy`, `pandas`, and `sklearn.metrics` Python libraries.

The XGBoost machine learning algorithm is used as the second out-of-sample robustness check for the logit model estimations. Table 12 reports various performance metrics for the Scaled IPO proceeds and IPO underpricing XGBoost models across four train–test proportions, and the full sample Scaled IPO proceeds and IPO underpricing logit models.

As with the random forest performance metrics, Table 12 shows that the mean accuracies of the Scaled IPO proceeds and IPO underpricing XGBoost models almost persistently increase as the training proportion increases. With an 80/20 train–test split, the Scaled IPO proceeds XGBoost model achieves its best accuracy (62.06%). The mean accuracy with the 90/10 train–test split (61.81%) is only slightly lower. On the other hand, the IPO underpricing XGBoost accuracy is highest (59.58%) with the 90/10 split. Comparing Scaled IPO proceeds and IPO underpricing XGBoost mean accuracies across all train–test proportions, the two models are generally equally accurate, while Scaled IPO proceeds model has a slight edge.

The finding that accuracies are generally higher with a larger training set implies that the Scaled IPO proceeds and IPO underpricing random forest models benefit from a larger training proportion to better learn existing weak patterns. With a larger training proportion in the rather small dataset of this thesis, Scaled IPO proceeds and IPO underpricing models are less sensitive to which data observations end up in the test set, which could potentially distort the mean accuracies.

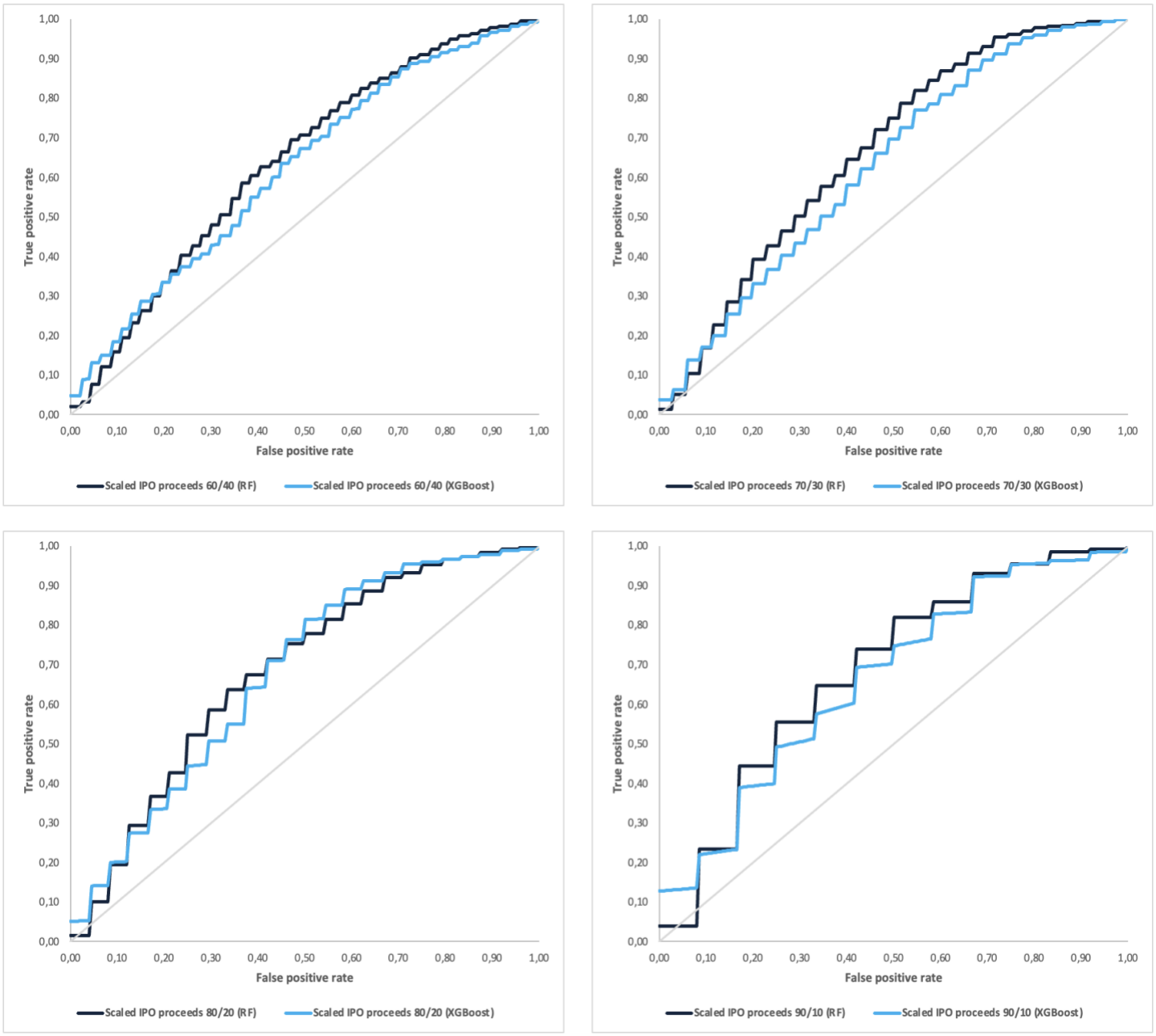
As with the random forest models, the IPO underpricing XGBoost model's recall appears to be high, while specificity is very low across all train–test proportions. This implies that the IPO underpricing XGBoost models tend to correctly predict IPO success, but struggles to predict no IPO success, which is reflected also in low NPV values. The differences between recall and specificity area generally smaller with the Scaled IPO proceeds XGBoost models compared to the IPO underpricing models, and Scaled IPO proceeds models have higher NPVs. The IPO underpricing XGBoost models have higher recall, indicating that these models predict IPO success better than the Scaled IPO proceeds XGBoost models.

Table 12 shows that, with an 80/20 train–test split, the Scaled IPO proceeds XGBoost model achieves its highest F1 score (62.27%), while the IPO underpricing XGBoost model's F1 score is highest (68.09%) with the 70/30 split. Because the F1 score ignores

lower specificity, the scores are higher for all models than the mean accuracies, and IPO underpricing XGBoost model's F1 values are higher than those of the Scaled IPO proceeds XGBoost models.

8.5 Random Forest and XGBoost ROC curves (Scaled IPO proceeds)

Figure 6. Receiver operating characteristic (ROC) curves for the Scaled IPO proceeds random forest and XGBoost models.



Note: These mean ROC curves are plotted from 30 repeated stratified splits. The ROC curves are computed using the `numpy`, `pandas`, and `sklearn.metrics` Python libraries.

According to Sapkota (2025, p. 14), the receiver operating characteristic (ROC) curves are graphical illustrations of models' performance at various classification thresholds and are plotted with true positive rate (recall) against the false positive rate (specificity). Figure 6 shows the ROC curves for Scaled IPO Proceeds random forest and XGBoost models across four train–test splits (60/40, 70/30, 80/20, and 90/10). Analyzing the visual profiles of the curves, the random forest models appear to perform better, because their dark blues curves are generally above the light blue XGBoost curves. Overall, all models perform only moderately across train–test splits, as the ROC curves are plotted far from the top-left corner.

Ranta (2023, Chapter 8.10.2.) explains that the area under the curve (AUC) metric captures ROC curve's visual information in a single value ranging from 0 to 1. Mean AUC values reported in Table 11 show that the 90/10 train–test split of the random forest IPO Scaled proceeds models outperform the other splits, while the best split with XGBoost is 80/20. The mean AUC of the Scaled IPO proceeds random forest 90/10 model is 68.33%, while the mean AUC of Scaled IPO proceeds XGBoost 80/20 model is 67.61%. As implied also in Sections 8.3 and 8.4, the Scaled IPO proceeds models have a moderate capability to predict both IPO success and no IPO success.

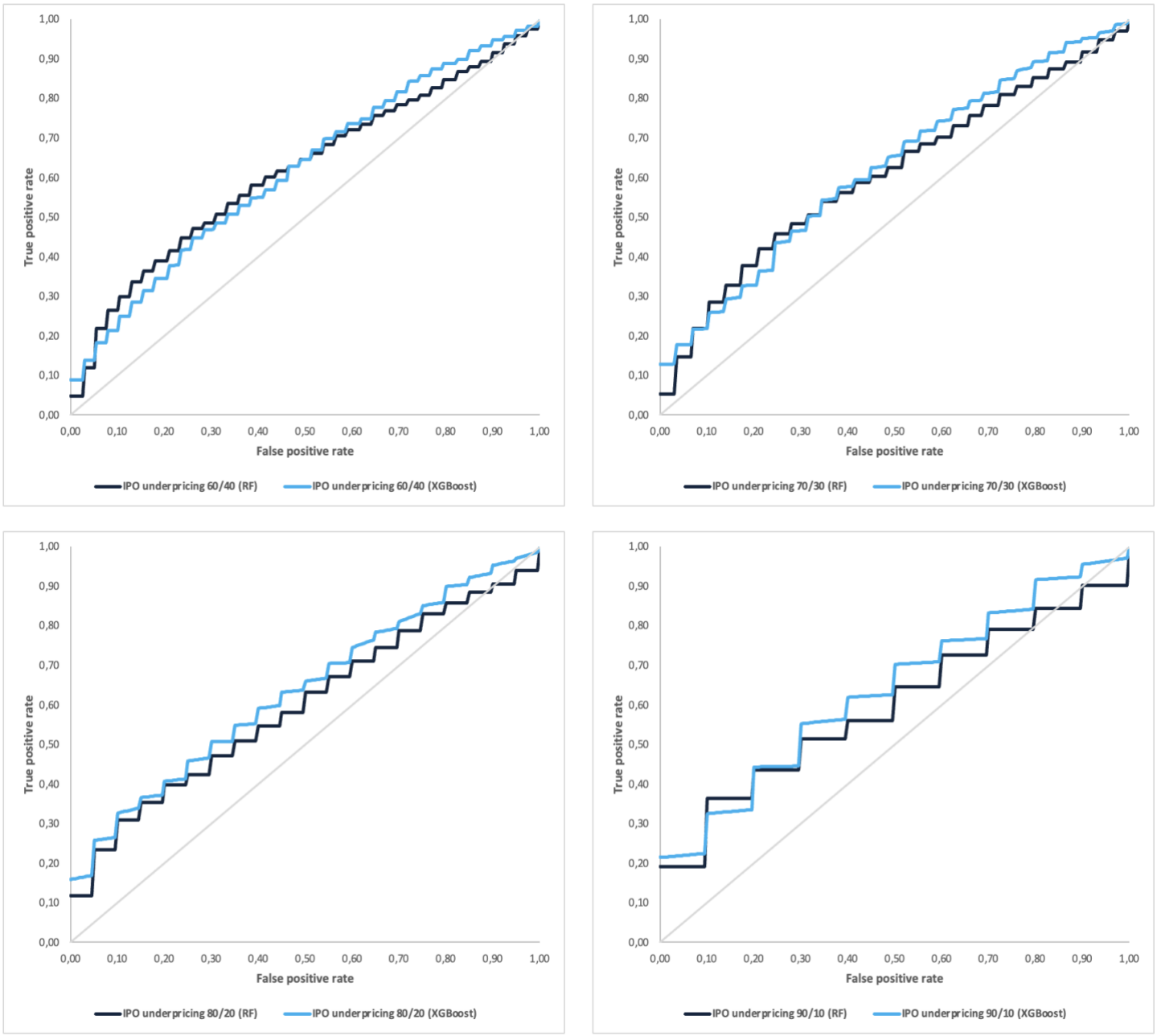
8.6 Random Forest and XGBoost ROC curves (IPO underpricing)

Figure 7 shows the ROC curves for IPO underpricing random forest and XGBoost models across four train–test splits. Based on the visual profiles of the curves, the XGBoost models appear to perform better, because their light blue curves are generally above the dark blue random forest curves. However, both machine learning algorithms seem to perform only moderately across all train–test splits, as the ROC curves are plotted far from the top-left corner.

Mean AUC values reported in Table 12 indicate that the IPO underpricing random forest and XGBoost models outperform the other splits with train–test splits of 60/40 and 90/10, respectively. The mean AUC of IPO underpricing 60/40 random forest model is

61.65%, while the AUC of IPO underpricing XGBoost 90/10 model is 63.65%. Thus, IPO underpricing models have a weaker overall capability to predict IPO success and no IPO success compared to the random forest models. As discussed in Sections 8.3 and 8.4, IPO underpricing models struggle to predict no IPO success, which explains the relatively weaker AUC values. The predictor variables' (with fixed effects) in the IPO underpricing models have weak predictive signal for *IPO underpricing*.

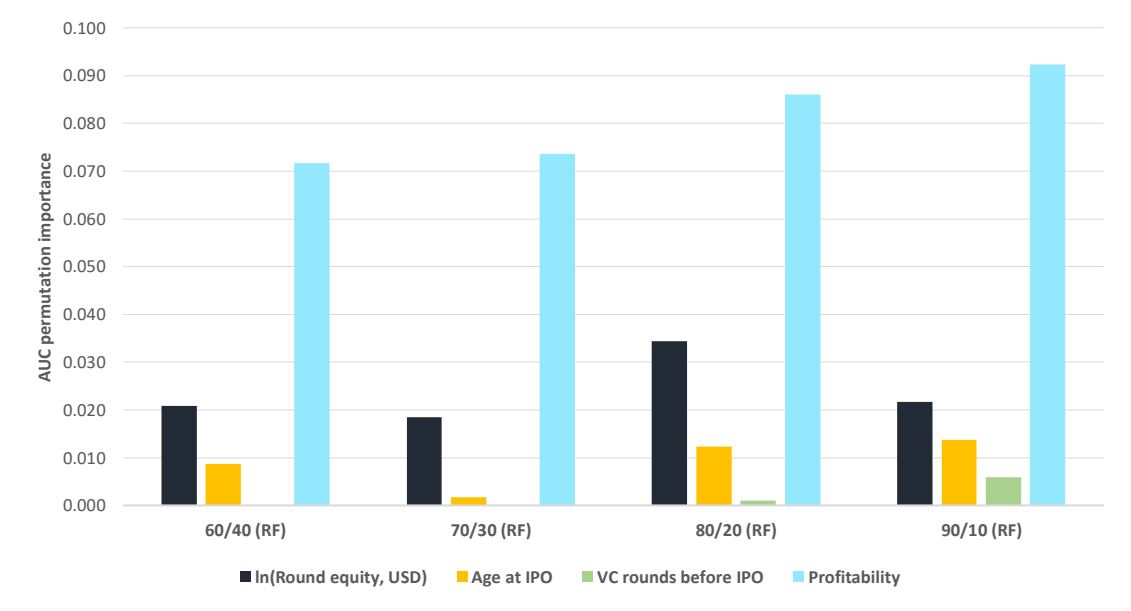
Figure 7. Receiver operating characteristic (ROC) curves for the IPO underpricing random forest and XGBoost models.



Note: These mean ROC curves are plotted from 30 repeated stratified splits. The ROC curves are computed using the `numpy`, `pandas`, and `sklearn.metrics` Python libraries.

8.7 Random Forest models' variable importance and PDPs

Figure 8. Variable importance plot for the Scaled IPO proceeds random forest models.



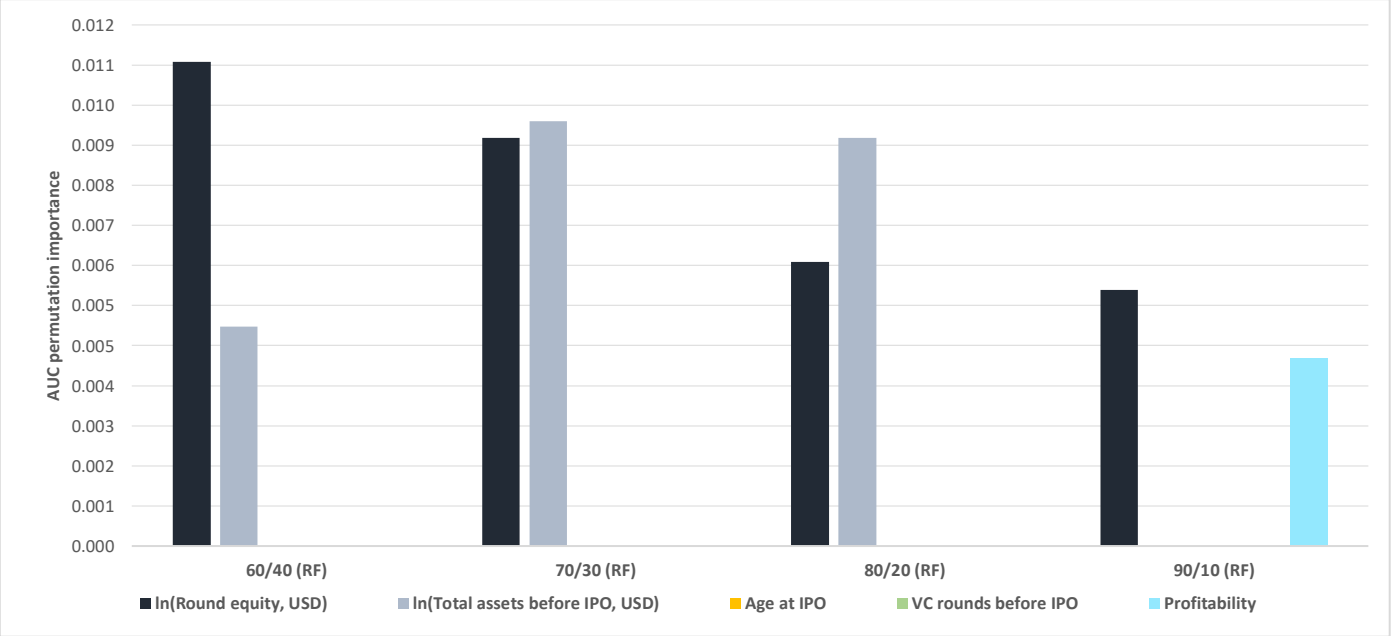
Note: This figure shows the mean AUC permutation importance for the Scaled IPO proceeds random forest model variables over 30 repeated stratified splits and across train–test proportions of 60/40, 70/30, 80/20, and 90/10 evaluated on test sets only. Permutation importance values are computed with 50 random shuffles to determine how much model's mean AUC score degrades when each variable is permuted. Negative permutation importances are set to zero. See Appendix 8 for detailed reporting of the permutation importance values. The AUC permutation importances are computed using the `numpy`, `pandas`, `statsmodels`, and `sklearn` (`sklearn.inspection.permutation_importance` function) Python libraries.

Figure 8 shows a variable importance plot for the Scaled IPO proceeds random forest models across all train–test proportions. Based on permutation importance, the size of the bar indicates how important that specific variable is in the model for that training proportion. According to Ranta (2023, Chapter 12.2.1.), permutation importance indicates how much the model's prediction error increases when predictor variable's relationships to the target are broken by repeated shuffles. The resulting degradation of the model's performance indicates how dependent the model is on the specific predictor variable while other variables remain unchanged. In this thesis, AUC is the scoring metric used for permutation importance.

Figure 8 shows that *profitability* is the dominant predictor variable for AUC in Scaled IPO proceeds random forest models. The variable of interest, *ln(Round equity, USD)*, is the second most important variable overall, and the evidence indicates that it is an

important predictor across all train–test proportions. Figure 9 suggests that, in the IPO underpricing random forest models, *ln(Round equity, USD)* is the most important predictors for AUC. However, the importance of *ln(Round equity, USD)* appears to be sensitive to the train–test ratio, and it systematically decreases as the training proportion increases. Appendix 8 reports the exact mean AUC permutation importance values.

Figure 9. Variable importance plot for the IPO underpricing random forest models.



Note: This figure shows the mean AUC permutation importance for the IPO underpricing random forest model variables over 30 repeated stratified splits and across train–test proportions of 60/40, 70/30, 80/20, and 90/10 evaluated on test sets only. Permutation importance values are computed with 50 random shuffles to determine how much model's mean AUC score degrades when each variable is permuted. Negative permutation importances are set to zero. See Appendix 8 for detailed reporting of the permutation importance values. The AUC permutation importances are computed using the `numpy`, `pandas`, `statsmodels`, and `sklearn` (`sklearn.inspection.permutation_importance` function) Python libraries.

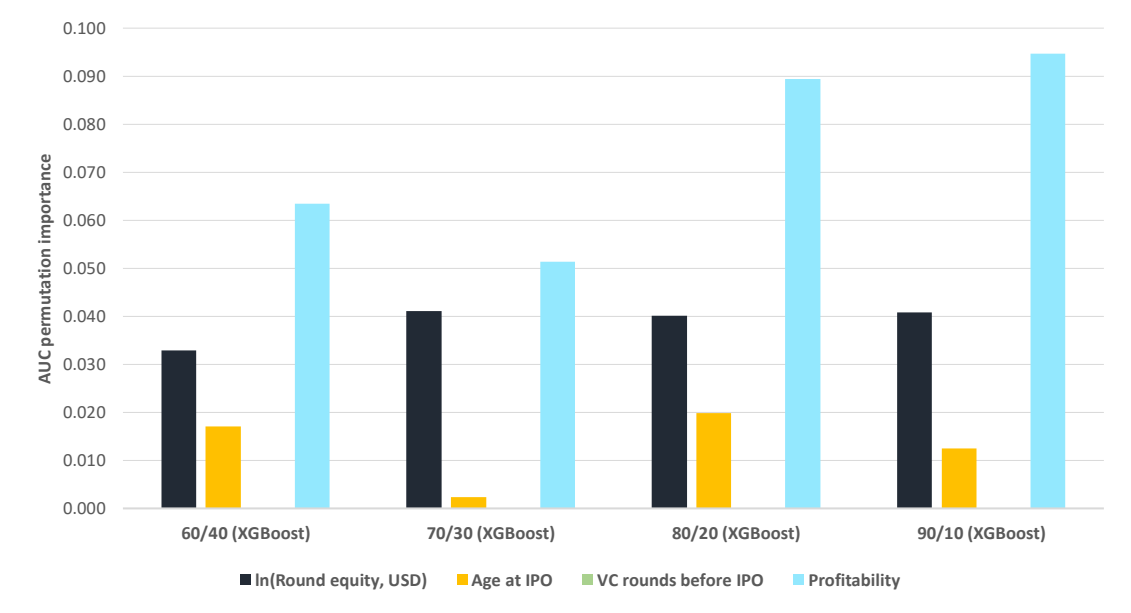
8.7.1 Random Forest PDPs

Descriptive partial dependence plots (PDPs) in Appendices 10 and 11 show the marginal effect of *ln(Round equity, USD)* on the average probability of Scaled IPO proceeds and IPO underpricing in random forest models. Ranta (2023, Chapter 12.1.1.) explains that a PDP averages out the influence of the model's other predictor variables (and fixed effects). Appendix 8 quantifies the differences (Δ) in average probabilities between the 95th percentile and 5th percentile of the variable of interest, *ln(Round equity, USD)*.

Appendix 8 show that, moving from the 5th percentile of *ln(Round equity, USD)* to the 95th percentile, the partial dependence change of Scaled IPO proceeds is positive across all four train–test proportions. Specifically, the average probabilities of above-median scaled IPO proceeds using the random forest 60/40, 70/30, 80/20, and 90/10 models increase by 10.52, 6.88, 10.73, and 7.13 percentage points, respectively. Appendix 10 illustrates the partial dependence change within the *ln(Round equity, USD)* grid. On the other hand, Δ PDP (p95) – PDP (p05) changes of the IPO underpricing models are generally slightly negative across the train–test ratios. Additionally, Appendix 11 shows that the average probability of IPO underpricing is plotted nonmonotonically within the grid, making the magnitude of the PDP changes arguable. A linear trendline through the grid is almost flat for all train–test proportions.

8.8 XGBoost models' variable importance and PDPs

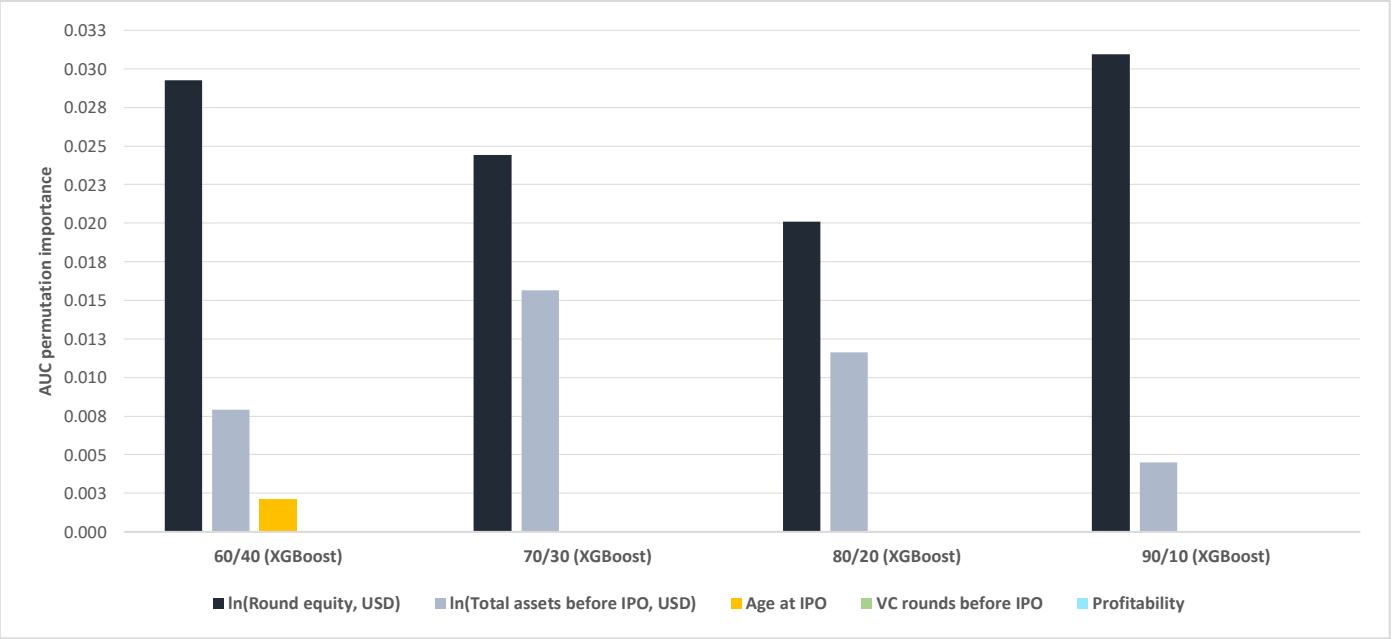
Figure 10. Variable importance plot for the Scaled IPO proceeds XGBoost models.



Note: This figure shows the mean AUC permutation importance for the Scaled IPO proceeds XGBoost model variables over 30 repeated stratified splits and across train–test proportions of 60/40, 70/30, 80/20, and 90/10 evaluated on test sets only. Permutation importance values are computed with 50 random shuffles to determine how much model's mean AUC score degrades when each variable is permuted. Negative permutation importances are set to zero. See Appendix 9 for detailed reporting of the permutation importance values. The AUC permutation importances are computed using the `numpy`, `pandas`, `statsmodels`, and `sklearn` (`sklearn.inspection.permutation_importance` function) Python libraries.

Figure 10 shows a variable importance plot for the Scaled IPO proceeds XGBoost models. *Profitability* and *ln(Round equity, USD)* are once again the two most important variables for AUC. However, with XGBoost, the importance of *ln(Round equity, USD)* appears to be stronger overall than with the random forest algorithm. This observation is consistent across all train–test proportions. Figure 11 shows that *ln(Round equity, USD)* is the dominant predictor variable for AUC in IPO underpricing XGBoost models. The importance of *ln(Round equity, USD)* appears to be train–test ratio sensitive. The highest and second-highest mean AUC permutation importance values are at 90/10 and 60/40, respectively. Appendix 9 reports the exact mean AUC permutation importance values.

Figure 11. Variable importance plot for the IPO underpricing XGBoost models.



Note: This figure shows the mean AUC permutation importance for the IPO underpricing XGBoost model variables over 30 repeated stratified splits and across train–test proportions of 60/40, 70/30, 80/20, and 90/10 evaluated on test sets only. Permutation importance values are computed with 50 random shuffles to determine how much model's mean AUC score degrades when each variable is permuted. Negative permutation importances are set to zero. See Appendix 9 for detailed reporting of the permutation importance values. The AUC permutation importances are computed using the `numpy`, `pandas`, `statsmodels`, and `sklearn` (`sklearn.inspection.permutation_importance` function) Python libraries.

8.8.1 XGBoost PDPs

Appendix 9 shows that, moving from the 5th percentile of *ln(Round equity, USD)* to the 95th percentile, the partial dependence change of Scaled IPO proceeds is generally negative across train–test proportions. This evidence is contrary to the Δ PDP (p95) – PDP (p05) changes using the random forest algorithm. At 70/30, the PDP change is

positive, but this appears to be an outlier among the other train–test ratios. Based on Appendix 12, the average probability of Scaled IPO proceeds is plotted nonmonotonically within the grid with all train–test proportions, making the magnitude of the PDP changes arguable. Linear trendlines through the *ln(Round equity, USD)* grid are positive for all train–test proportions. On the other hand, Δ PDP (p95) – PDP (p05) changes of the IPO underpricing models are negative across all train–test ratios, and as with the Scaled IPO proceeds XGBoost models, the average probabilities are plotted nonmonotonically within the grid in Appendix 13.

8.9 Research question reformulation and 5-fold cross-validation tests

To properly align with the machine learning methodology and the type of results the algorithms provide, the main research question is reformulated into three sub-research questions, while maintaining consistency with the thesis's hypotheses:

RQ1: *Is the size of the largest disclosed VC equity round conditionally associated with Scaled IPO proceeds/IPO underpricing by adding predictive importance?*

RQ2: *Does the size of the largest disclosed VC equity round have a positive conditional association with Scaled IPO proceeds by increasing its average predicted probability?*

RQ3: *Does the size of the largest disclosed VC equity round have a negative conditional association with IPO underpricing by decreasing its average predicted probability?*

While evidence to answer these research questions has already been presented, 5-fold cross-validation tests are run to support the robustness of the findings and to confirm or deny them. According to Sapkota (2025, p. 25) cross-validation tests provide more reliable results on a machine learning model's capability to generalize to out-of-sample data compared to single train–test splits.

Table 13 shows that the 5-fold cross-validation AUCs for the Scaled IPO proceeds random forest (66.58%) and XGBoost (66.97%) models are marginally lower than the AUCs of the corresponding single train–test split 90/10 random forest (68.33%) and 80/20 XGBoost (67.61%) models. By contrast, the AUCs for the IPO underpricing random forest (61.27%) and XGBoost (65.18%) models are lower and higher, respectively, compared to the single train–test split AUCs of 61.65% (60/40 RF) and 63.55% (90/10 XGBoost). Scaled IPO proceeds models are better at predicting IPO success and no IPO success and have a moderate overall AUC performance.

Table 13. 5-fold cross-validation test results.

Scaled IPO proceeds	**N**	**Metric**	**Predictor variable**	**5-fold cross-validation**	**95% CI**	**Sample sizes**
Random Forest	231	AUC		0.6658	[0.6189, 0.7128]	184, 185, 185, 185, 185
	231	AUC permutation importance	ln(Round equity, USD)	0.0408	[0.0061, 0.0755]	184, 185, 185, 185, 185
	231	Δ PDP (p95) – PDP (p05)	ln(Round equity, USD)	0.1116	[0.0240, 0.1993]	184, 185, 185, 185, 185
XGBoost	231	AUC		0.6697	[0.6329, 0.7065]	184, 185, 185, 185, 185
	231	AUC permutation importance	ln(Round equity, USD)	0.0533	[0.0143, 0.0924]	184, 185, 185, 185, 185
	231	Δ PDP (p95) – PDP (p05)	ln(Round equity, USD)	-0.0128	[-0.0619, 0.0364]	184, 185, 185, 185, 185
IPO underpricing	**N**	**Metric**	**Predictor variable**	**5-fold cross-validation**	**95% CI**	**Sample sizes**
Random Forest	231	AUC		0.6127	[0.5303, 0.6951]	184, 185, 185, 185, 185
	231	AUC permutation importance	ln(Round equity, USD)	0.0030	[-0.0408, 0.0468]	184, 185, 185, 185, 185
	231	Δ PDP (p95) – PDP (p05)	ln(Round equity, USD)	-0.0075	[-0.0500, 0.0349]	184, 185, 185, 185, 185
XGBoost	231	AUC		0.6518	[0.5591, 0.7445]	184, 185, 185, 185, 185
	231	AUC permutation importance	ln(Round equity, USD)	0.0272	[-0.0371, 0.0914]	184, 185, 185, 185, 185
	231	Δ PDP (p95) – PDP (p05)	ln(Round equity, USD)	-0.0503	[-0.0819, -0.0188]	184, 185, 185, 185, 185

Note: This table shows mean 5-fold cross-validation test results for AUC, AUC permutation importance, and Δ PDP (p95) – PDP (p05). AUC captures model's capability to correctly predict IPO success and no IPO success across different classification thresholds, AUC permutation importance captures how much model's AUC degrades when relationships of *ln(Round equity, USD)* to the target are broken by 50 repeated shuffles, and Δ PDP (p95) – PDP (p05) is the difference of the average probability of IPO success between the 95th percentile and 5th percentile of *ln(Round equity, USD)*. 80% of the data goes into training, and 20% to validation for each fold. The 5-fold cross-validation tests are computed using the `numpy`, `pandas`, `xgboost`, and `sklearn` (`sklearn.model_selection.StratifiedKFold` function) Python libraries.

Based on Table 13, positive *ln(Round equity, USD)* AUC permutation importance values across all models confirm RQ1. Thus, the results suggest that the size of the largest disclosed VC equity round is conditionally associated with *Scaled IPO proceeds* and *IPO underpricing* success measures by adding predictive importance for the models' AUC. RQ2 is not consistently supported by the 5-fold cross-validation results; Δ PDP (p95) –

PDP (p05) is positive with random forest (0.1116) but negative with XGBoost (–0.0128). On the other hand, both random forest and XGboost algorithms confirm RQ3, as Δ PDP (p95) – PDP (p05) changes are negative (–0.0075 and –0.0503, respectively). Therefore, the 5-fold cross-validation results confirm that the size of the largest disclosed VC equity round has a negative conditional association with *IPO underpricing* by decreasing its average predicted probability.

Table 14 summarizes all relevant logit and machine learning findings of this thesis considering the association between the size of the largest disclosed VC equity round and IPO success, as measured by *Scaled IPO proceeds* and *IPO underpricing*.

Table 14. Summary of the thesis's results.

	IPO success measure:	
	Scaled IPO proceeds	**IPO underpricing**
Logit (full sample):		
ln(Round equity, USD) coefficient estimate	0.1010	0.1056
Statistical significance	No statistical significance	No statistical significance
Fails to reject H_0 / Rejects H_0	Fails to reject H_0	Fails to reject H_0
Supports H_1 / Do not support H_1	Do not support H_1	Do not support H_1
Random Forest (5-fold cross-validation):		
AUC permutation importance	Positive	Positive
Δ PDP (p95) – PDP (p05)	Positive	Negative
Confirms RQ1 / Denies RQ1	Confirms RQ1	Confirms RQ1
Confirms RQ2 / Denies RQ2	Confirms RQ2	–
Confirms RQ3 / Denies RQ3	–	Confirms RQ3
XGBoost (5-fold cross-validation):		
AUC permutation importance	Positive	Positive
Δ PDP (p95) – PDP (p05)	Negative	Negative
Confirms RQ1 / Denies RQ1	Confirms RQ1	Confirms RQ1
Confirms RQ2 / Denies RQ2	Denies RQ2	–
Confirms RQ3 / Denies RQ3	–	Confirms RQ3

9 Conclusions

Studying VC-backed IPO companies in the EEA and EFTA countries during 2000–2023, this thesis finds no statistically significant conditional association between the size of the largest disclosed VC round and IPO success when using traditional logistic regression. Therefore, the null hypotheses (H_0) cannot be rejected, and alternative hypotheses (H_1) are not supported. However, the results from the random forest and XGBoost machine learning models suggest different findings.

All machine learning models in this thesis indicate that the size of the largest disclosed VC equity round is conditionally associated with *Scaled IPO proceeds* and *IPO underpricing* by adding predictive importance for the models' AUC. In other words, *ln(Round equity, USD)* is an important variable in the thesis's models predicting IPO success. Therefore, sub-research question RQ1 is confirmed by the evidence.

However, the machine learning algorithms do not consistently support RQ2 – that the size of the largest disclosed VC equity round has a positive conditional association with *Scaled IPO proceeds* by increasing its average predicted probability. The Scaled IPO proceeds random forest model suggest a positive association, but the XGBoost model implies a negative association. Therefore, the RQ2 is denied based on the mixed evidence. On the other hand, both the IPO underpricing random forest and XGBoost models agree in confirming RQ3, as the Δ PDP (p95) – PDP (p05) changes are negative. Specifically, the evidence suggests that the size of the largest disclosed VC equity round has a negative conditional association with *IPO underpricing* by decreasing its average predicted probability.

While the machine learning models answer slightly different questions about the conditional association between the size of the largest VC round and IPO success than the logit models, the evidence presented in this thesis is still valuable for issuers and investors considering participating in an IPO. The random forest and XGBoost algorithms

can capture nonlinearities and interactions that cannot be captured by using logistic regression. This could be one of the main drivers of the differing findings. The thesis successfully fills a small part of the gap in the academic European VC-backed IPO performance literature by introducing a perspective on the VC round size by means of machine learning.

When considering the limitations of this thesis, the empirical findings should be interpreted only as conditional associations between the size of the largest disclosed VC equity round and IPO success measures, and not as causal effects. The studied associations are conditional on other predictor variables and fixed effects in the models.

The selected IPO underpricing model predictor variables and fixed effects do not explain much of the variation in terms of whether an IPO is successful. This implies that changing or adding other variables is necessary to improve the predictive power of the model. Adding controls that better explain pricing outcomes, such as market conditions, hot issue markets, and underwriter reputation, could yield improved model fit. In addition, the IPO underpricing success measure can be viewed as a logic opposite to that used in this thesis, highlighting that it matters how IPO success is interpreted. Under this opposite logic, also negative IPO underpricing could be interpreted as a sign of success, as less IPO proceeds are left on the table. Moreover, the *ln(Round equity, USD)* captures only information on disclosed VC equity rounds, and if disclosure is incomplete, it may not accurately represent the size of the largest VC equity round raised prior to an IPO.

Future researchers could study the true causal effects of the size of the largest disclosed VC equity round on IPO success. A potential method for this is an instrumental variables (IV) two-stage least squares (2SLS) regression analysis. Lehnertz, Plagmann, and Lutz (2022) use this in their empirical tests to avoid conflating the causal effect of the mega-deal with underlying firm quality. In addition, using other machine learning algorithms could yield improved and more accurate results. Sapkota (2025) uses support vector machine, neural networks, and Naive Bayes algorithms.

References

Abrahamson, M., Jenkinson, T., & Jones, H. (2011). Why don't US issuers demand European fees for IPOs?. *Journal of Finance*, 66(6), 2055-2082.

Akerlof, G. A. (1970). The market for “lemons”: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3), 488-500.

Allen, F., & Faulhaber, G. R. (1989). Signalling by underpricing in the IPO market. *Journal of Financial Economics*, 23(2), 303-323.

Ang, Y. Q., Chia, A., & Saghafian, S. (2021). Using machine learning to demystify startups' funding, post-money valuation, and success. *Innovative Technology at the Interface of Finance and Operations*: Volume I (pp. 271-296). Cham: Springer International Publishing.

Assoil, A., & Laporte, J. M. (2024). Underpricing of IPOs: Evidence from the Euronext Paris market. *Economics Bulletin*, 44(1), 416-429.

Barry, C. B., Muscarella, C. J., Peavy Iii, J. W., & Vetsuypens, M. R. (1990). The role of venture capital in the creation of public companies: Evidence from the going-public process. *Journal of Financial Economics*, 27(2), 447-471.

Bastı, E., Kuzey, C., & Delen, D. (2015). Analyzing initial public offerings' short-term performance using decision trees and SVMs. *Decision Support Systems*, 73, 15-27.

Benveniste, L. M., & Spindt, P. A. (1989). How investment bankers determine the offer price and allocation of new issues. *Journal of Financial Economics*, 24(2), 343-361.

Bessler, W., & Seim, M. (2012). The performance of venture-backed IPOs in Europe. *Venture Capital*, 14(4), 215-239.

Bodie, Z., Kane, A., & Marcus, A. J. (2014). *Investments*. McGraw-Hill Education.

Bottazzi, L., & Da Rin, M. (2002). Venture capital in Europe and the financing of innovative companies. *Economic Policy*, 17(34), 229-270.

Carter, R., & Manaster, S. (1990). Initial public offerings and underwriter reputation. *Journal of Finance*, 45(4), 1045-1067.

Caselli, S., & Negri, G. (2021). *Private Equity and Venture Capital in Europe: Markets, Techniques, and Deals*. Academic Press.

Chemmanur, T. J., Krishnan, K., & Nandy, D. K. (2011). How does venture capital financing improve efficiency in private firms? A look beneath the surface. *Review of Financial Studies*, 24(12), 4037-4090.

Coakley, J., Hadass, L., & Wood, A. (2009). UK IPO underpricing and venture capitalists. *The European Journal of Finance*, 15(4), 421-435.

CompaniesMarketcap.com. (n.d.). *Largest Companies by Marketcap*. Retrieved April 3, 2026, from https://companiesmarketcap.com

Cumming, D. (Ed.). (2012). *The Oxford handbook of venture capital*. Oxford University Press.

Euronext (2025). *IPO Guide: A Guide to Listing on the Stock Exchange*. Retrieved April 3, 2026, from https://www.euronext.com/en/listing/raise-capital/how-go-public/ipo-journey

Euronext (n.d.). *Our Business*. Euronext. Retrieved April 3, 2026, from https://www.euronext.com/en/about/our-business

European Union (n.d. -a). *Glossary: European Economic Area (EEA)*. Eurostat. Retrieved April 3, 2026, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:European_Economic_Area_(EEA)

European Union (n.d. -b). *Glossary: European Free Trade Association (EFTA)*. Eurostat. Retrieved April 3, 2026, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:European_Free_Trade_Association_(EFTA)

European Union (n.d. -c). *Brexit*. European Comission. Retrieved April 3, 2026, from https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/brexit_en

Flickinger, M. (2023, March 29). *Venture Capital Fundamentals: Why VC Is A Driving Force Of Innovation.* Forbes. Retrieved April 3, 2026, from https://www.forbes.com/sites/markflickinger/2023/03/29/venture-capital-fundamentals-why-vc-is-a-driving-force-of-innovation/

Gajewski, J. F., & Gresse, C. (2006). A survey of the European IPO market. *ECMI Research Paper*, (2).

Gompers, P. A. (1995). Optimal Investment, Monitoring, and the Staging of Venture Capital. *Journal of Finance*, 50(5), 1461-1489.

Gompers, P. A. (1996). Grandstanding in the venture capital industry. *Journal of Financial Economics*, 42(1), 133-156.

Gulati, R., & Higgins, M. C. (2003). Which ties matter when? The contingent effects of interorganizational partnerships on IPO success. *Strategic Management Journal*, 24(2), 127-144.

Hanley, K. W. (1993). The underpricing of initial public offerings and the partial adjustment phenomenon. *Journal of Financial economics*, 34(2), 231-250.

Helwege, J., & Liang, N. (2004). Initial public offerings in hot and cold markets. *Journal of financial and quantitative analysis*, 39(3), 541-569.

Hochberg, Y. V., Ljungqvist, A., & Lu, Y. (2007). Whom you know matters: Venture capital networks and investment performance. *Journal of Finance*, 62(1), 251-301.

Honorine, A. N. D., & Emmanuelle, D. (2019). Stage financing and syndication in the IPO underpricing of venture-backed firms: Venture capital and IPO underpricing. *International Journal of Entrepreneurship and Innovation*, 20(4), 289-300.

Ibbotson, R. G., & Jaffe, J. F. (1975). "Hot issue" markets. *Journal of Finance*, 30(4), 1027-1042.

IBM (n.d.). *What is XGBoost?*. *Think*. IBM. Retrieved April 3, 2026, from https://www.ibm.com/think/topics/xgboost

Knickerboecker, K. (n.d.). *PitchBook's guide to VC fundraising for startup.* PitchBook blog. Retrieved April 3, 2026, from https://pitchbook.com/blog/vc-fundraising-for-startups-guide

Krishnan, C. N. V., Ivanov, V. I., Masulis, R. W., & Singh, A. K. (2011). Venture capital reputation, post-IPO performance, and corporate governance. *Journal of Financial and Quantitative Analysis*, 46(5), 1295-1333.

Lee, P. M., & Wahal, S. (2004). Grandstanding, certification and the underpricing of venture capital backed IPOs. *Journal of Financial Economics*, 73(2), 375-407.

Lerner, J. (1994). The Syndication of Venture Capital Investments. *Financial Management*, 23(3), 16-27.

Lerner, J., & Nanda, R. (2020). Venture capital's role in financing innovation: What we know and how much we still need to learn. *Journal of Economic Perspectives*, 34(3), 237-261.

Leland, H. E., & Pyle, D. H. (1977). Informational asymmetries, financial structure, and financial intermediation. Journal of Finance, 32(2), 371-387.

Loughran, T., & Ritter, J. (2004). Why has IPO underpricing changed over time?. *Financial Management*, 5-37.

Megginson, W. L., & Weiss, K. A. (1991). Venture capitalist certification in initial public offerings. *Journal of Finance*, 46(3), 879-903.

Metrick, A., & Yasuda, A. (2021). *Venture capital and the finance of innovation*. John Wiley & Sons.

Nielsen, D. (2016). Tree boosting with xgboost-why does xgboost win" every" machine learning competition. *Master's thesis*, NTNU.

Pantea, S., & Tkacik, M. (2025). Venture capital and high-tech start-ups in Europe: a systematic review of the empirical evidence. *Venture Capital*, 27(4), 435-458.

Patel, N. (2022, January 19). *PitchBook's Annual 2021 European Venture Report*. PitchBook. Retrieved April 3, 2026, from https://files.pitchbook.com/website/files/pdf/2021_Annual_European_Venture_Report.pdf

Patel, N. (2023, January 18). *PitchBook's Annual 2022 European Venture Report*. PitchBook. Retrieved April 3, 2026, from https://files.pitchbook.com/website/files/pdf/2022_Annual_European_Venture_Report.pdf

Pennacchio, L. (2014). The causal effect of venture capital backing on the underpricing of Italian initial public offerings. *Venture Capital*, 16(2), 131-155.

PitchBook. (2023a, October 12). *Public vs. private sector: What's the difference?*. PitchBook blog. Retrieved April 3, 2026, from https://pitchbook.com/blog/public-vs-private-sector-whats-the-difference

PitchBook. (2023b, October 18). *PitchBook's Q3 2023 European Venture Report*. Retrieved Retrieved April 3, 2026, from https://files.pitchbook.com/website/files/xls/Q3_2023_European_Venture_Report_Summary_XLS.xlsx

PitchBook. (2024, July 17). *PitchBook's Q2 2024 European Venture Report*. Retrieved April 3, 2026, from https://files.pitchbook.com/website/files/xls/Q2_2024_European_Venture_Report_Summary_XLS.xlsx

PitchBook. (n.d. -a). *Broadcom Overview*. Retrieved April 3, 2026, from https://pitchbook.com/profiles/company/10010-89#overview

PitchBook. (n.d. -b). *Nvidia Overview*. Pitchbook. Retrieved April 3, 2026, from https://pitchbook.com/profiles/company/41161-24#overview

Quintana, D., Sáez, Y., & Isasi, P. (2017). Random forest prediction of IPO underpricing. *Applied Sciences*, 7(6), 636.

Rajan, N. (2024, January 17). *PitchBook's Annual 2023 European Venture Report.* PitchBook. Retrieved April 3, 2026, from https://files.pitchbook.com/website/files/pdf/2023_Annual_European_Venture_Report.pdf

Rajan, N. (2025a, January 22). *PitchBook's Annual 2024 European Venture Report.* PitchBook. Retrieved April 3, 2026, from https://files.pitchbook.com/website/files/pdf/2024_Annual_European_Venture_Report.pdf

Rajan, N. (2025b, July 10). *PitchBook's Q2 2025 European Venture Report.* PitchBook. Retrieved April 3, 2026, from https://files.pitchbook.com/website/files/pdf/Q2_2025_European_Venture_Report.pdf

Ranta, M. (2023). *Introduction to Data Analytics in Accounting and Finance*. Github. Retrieved April 3, 2026, from https://mranta-ai.github.io/Data_analytics_in_accounting/book_intro.html

Ritter, J. R (1998). *Initial Public Offerings*. Retrieved April 3, 2026, from https://site.warrington.ufl.edu/ritter/files/CFD.pdf

Ritter, J. R., & Welch, I. (2002). A review of IPO activity, pricing, and allocations. *Journal of Finance*, 57(4), 1795-1828.

Shalman, W. A. (1990). The structure and governance of venture-capital organizations. *Journal of Financial Economics*, 27(2), 473-521.

Sapkota, N. (2025). The crypto collapse chronicles: Decoding cryptocurrency exchange defaults. *Journal of International Financial Markets, Institutions and Money*, 99, 102093.

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3), 355-374.

Tanda, A., & Manzi, G. (2020). Underpricing of venture backed IPOs: a meta-analysis approach. *Economics of Innovation and new technology*, 29(4), 331-348.

Tastan, M., Falconieri, S., & Filatotchev, I. (2013). Does venture capital syndicate size matter?. *European Financial Management Association 2013 Annual Meeting*.

Tian, X. (2011). The causes and consequences of venture capital stage financing. *Journal of Financial Economics*, 101(1), 132-159.

Tian, X. (2012). The role of venture capital syndication in value creation for entrepreneurial firms. *Review of Finance*, 16(1), 245-283.

Welch, I. (1989). Seasoned offerings, imitation costs, and the underpricing of initial public offerings. *Journal of Finance*, 44(2), 421-449.

Wu, J., Li, S., & Li, Z. (2013). The contingent value of CEO political connections: A study on IPO performance in China. *Asia Pacific Journal of Management*, 30(4), 1087-1114.

Zuniga, R. (2024, August 1). *The IPO process explained.* PitchBook blog. Retrieved April 3, 2026, from https://pitchbook.com/blog/ipo-process-explained

Appendices

Appendix 1. Variable descriptions

Variable	Type	Description
IPO proceeds, USD	Continuous	Captures the dollar value of the total gross proceeds company generates in an IPO, including sold overallotment shares.
Scaled IPO proceeds, USD	Continuous	Scales *IPO proceeds, USD* by *Total assets before IPO, USD*.
IPO underpricing, %	Continuous	Calculated as the percentage difference between the closing price per share on the first day of trading and the IPO offer price.
Round equity, USD	Continuous	Captures the size, in dollars, of the largest disclosed venture capital equity financing round since the founding of a VC-backed IPO company.
ln(Round equity, USD)	Continuous	Logarithm of *Round equity, USD*.
Total assets before IPO, USD	Continuous	Captures the dollar value of total assets before a VC-backed IPO.
ln(Total assets before IPO, USD)	Continuous	Logarithm of *Total assets before IPO, USD*.
Age at IPO	Discrete	Captures the number of years between the IPO year and the founding year of a VC-backed company.
VC rounds before IPO	Discrete	Captures the number of venture capital rounds a VC-backed company completed before the IPO.
Scaled IPO proceeds	Binary	A binary variable of *Scaled IPO proceeds, USD*, which captures whether a VC-backed company's IPO is successful. The IPO is successful (1) if the scaled proceeds value is above the sample median, and unsuccessful (0) otherwise.
IPO underpricing	Binary	A binary variable of *IPO underpricing, %*, which captures whether a VC-backed company's IPO is successful. The IPO is successful (1) if it is underpriced (underpricing is greater than zero), and unsuccessful (0) otherwise.
Profitability	Binary	Captures whether a VC-backed company is profitable before the IPO. A company is profitable (1) if the net income after taxes before the IPO is greater than zero, and unprofitable (0) otherwise.
France	Binary	Captures whether the exchange nation of a VC-backed IPO is France.
Germany	Binary	Captures whether the exchange nation of a VC-backed IPO is Germany.
United Kingdom	Binary	Captures whether the exchange nation of a VC-backed IPO is the United Kingdom.
Rest of the EEA/EFTA	Binary	Captures whether the exchange nation of a VC-backed IPO is Austria, Belgium, Denmark, Finland, Hungary, Ireland, Italy, Netherlands, Norway, Poland, Portugal, Spain, Sweden, or Switzerland.
Consumer cyclicals	Binary	Captures whether the TRBC economic sector of a VC-backed IPO company is consumer cyclicals.
Healthcare	Binary	Captures whether the TRBC economic sector of a VC-backed IPO company is healthcare.
Technology	Binary	Captures whether the TRBC economic sector of a VC-backed IPO company is technology.
Other sectors	Binary	Captures whether the TRBC economic sector of a VC-backed IPO company is basic materials, consumer noncyclicals, energy, financials, industrials, or utilities.
IPO year 2000–2009	Binary	Captures whether the IPO year of the VC-backed company took place during 2000–2009.
IPO year 2010–2019	Binary	Captures whether the IPO year of the VC-backed company took place during 2010–2019.
IPO year 2020–2023	Binary	Captures whether the IPO year of the VC-backed company took place during 2020–2023.

Appendix 2. Pearson and point-biserial correlation matrix

	Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
1	**Scaled IPO proceeds**	1.000														
2	**IPO underpricing**	0.049	1.000													
3	**ln(Round equity, USD)**	0.106	0.019	1.000												
4	**ln(Total assets before IPO, USD)**	-0.320	-0.043	0.460	1.000											
5	**Age at IPO**	-0.160	-0.010	-0.022	0.176	1.000										
6	**VC rounds before IPO**	0.060	0.009	0.339	0.008	-0.008	1.000									
7	**Profitability**	-0.293	-0.022	-0.182	0.297	0.201	-0.220	1.000								
8	**France**	-0.059	-0.237	-0.056	-0.179	0.011	-0.011	-0.027	1.000							
9	**Germany**	0.103	-0.072	0.042	0.084	-0.078	-0.015	0.010	-0.268	1.000						
10	**United Kingdom**	0.022	0.250	-0.009	0.001	0.088	-0.002	0.014	-0.532	-0.219	1.000					
11	**Consumer cyclicals**	-0.120	-0.049	0.104	0.266	0.102	-0.014	0.153	-0.045	0.149	0.024	1.000				
12	**Healthcare**	0.170	0.006	0.044	-0.246	-0.089	0.065	-0.370	0.050	-0.136	-0.160	-0.310	1.000			
13	**Technology**	-0.143	0.043	-0.129	-0.071	0.099	0.000	0.208	0.043	0.024	0.040	-0.223	-0.521	1.000		
14	**IPO year 2010–2019**	0.021	-0.034	0.066	0.021	-0.015	0.045	-0.167	0.307	-0.132	-0.210	0.017	0.132	-0.115	1.000	
15	**IPO year 2020–2023**	0.001	-0.056	0.111	0.073	0.064	0.028	-0.077	-0.020	-0.009	0.075	0.111	-0.037	-0.033	-0.299	1.000

Note: The Pearson correlation measure is used for pairs of continuous and/or discrete variables. The point-biserial correlation measure is used for pairs consisting of a continuous or discrete variable and a binary variable. For pairs of binary variables, the phi coefficient is used.

The matrix is computed using the `numpy` and `pandas` Python libraries.

Appendix 3. Performance metric descriptions

Metric	Formula	Description
True positive (TP)	*TP*	The number of cases correctly predicting IPO success when the IPO is actually successful in the data.
False positive (FP)	*FP*	The number of cases incorrectly predicting IPO success when the IPO is not actually successful in the data.
True negative (TN)	*TN*	The number of cases correctly predicting no IPO success when the IPO is not actually successful in the data.
False negative (FN)	*FN*	The number of cases incorrectly predicting no IPO success when the IPO is actually successful in the data.
Accuracy	*(TP + TN) / (TP + FP + TN + FN)*	The proportion of correct predictions among all predictions made by the model.
F1 score	*2 / (1 / Precision + 1 / Recall)*	A single value ranging from 0 to 1, which combines precision and recall metrics.
Area under the curve (AUC)	–	A single value ranging from 0 to 1 representing the models' capability to correctly predict IPO success and no IPO success across different classification thresholds.
95% confidence interval (95% CI, AUC)	–	95% confidence interval of AUC, providing a range of its possible values.
False positive rate (specificity)	*TN / (FP + TN)*	The proportion of times the model correctly predicts no IPO success among all cases, when the IPO is not actually successful in the data.
True positive rate (recall)	*TP / (TP + FN)*	The proportion of times the model correctly predicts IPO success among all cases, when the IPO is actually successful in the data.
Positive predictive value (precision)	*TP / (TP +FP)*	The proportion of times the model correctly predicts IPO success among all cases, when it predicts IPO success.
Negative predictive value (NPV)	*TN / (FN + TN)*	The proportion of times the model correctly predicts no IPO success among all cases, when it predicts no IPO success.

Appendix 4. Logit confusion matrix (Scaled IPO proceeds)

Actual outcome	**Predicted outcome by the logit function**	
	IPO success (by *Scaled IPO proceeds*)	No IPO success (by *Scaled IPO proceeds*)
IPO success (by *Scaled IPO proceeds*)	91 (*TP*)	24 (*FN*)
No IPO success (by *Scaled IPO proceeds*)	46 (*FP*)	70 (*TN*)

Note: This confusion matrix reports the numbers (in cells) of how many times the Scaled IPO proceeds logit model predicts IPO success outcome (columns) correctly or incorrectly relative to the actual outcome (rows) of an observation. In cells, where both the predicted and actual outcomes correspond to IPO success, or where both correspond to no IPO success, the matrix reports the number of correct predictions. Otherwise, the cell reports the number of incorrect predictions. The overall point accuracy of the Scaled IPO proceeds logit model is 0.6970, with 95% confidence interval of [0.6147, 0.7576].

The matrix is computed using the `numpy`, `pandas`, and `sklearn` Python library, with the function `sklearn.metrics.confusion_matrix`.

Appendix 5. Logit confusion matrix (IPO underpricing)

Actual outcome	**Predicted outcome by the logit function**	
	IPO success (by *IPO underpricing*)	No IPO success (by *IPO underpricing*)
IPO success (by *IPO underpricing*)	95 (*TP*)	40 (*FN*)
No IPO success (by *IPO underpricing*)	39 (*FP*)	57 (*TN*)

Note: This confusion matrix reports the numbers (in cells) of how many times the IPO underpricing logit model predicts IPO success outcome (columns) correctly or incorrectly relative to the actual outcome (rows) of an observation. In cells, where both the predicted and actual outcomes correspond to IPO success, or where both correspond to no IPO success, the matrix reports the number of correct predictions. Otherwise, the cell reports the number of incorrect predictions. The overall accuracy of the IPO underpricing logit model is 0.6580, with 95% confidence interval of [0.6104, 0.7532].

The matrix is computed using the `numpy`, `pandas`, and `sklearn` Python library, with the function `sklearn.metrics.confusion_matrix`.

Appendix 6. Random Forest confusion matrices

Actual outcome	**Predicted outcome by Random Forest algorithm (60/40 train–test split)**	
	IPO success (by *Scaled IPO proceeds*)	No IPO success (by *Scaled IPO proceeds*)
IPO success (by *Scaled IPO proceeds*)	29 (*TP*)	17 (*FN*)
No IPO success (by *Scaled IPO proceeds*)	20 (*FP*)	27 (*TN*)
	Predicted outcome by Random Forest algorithm (70/30 train–test split)	
	IPO success (by *Scaled IPO proceeds*)	No IPO success (by *Scaled IPO proceeds*)
IPO success (by *Scaled IPO proceeds*)	25 (*TP*)	10 (*FN*)
No IPO success (by *Scaled IPO proceeds*)	16 (*FP*)	19 (*TN*)
	Predicted outcome by Random Forest algorithm (80/20 train–test split)	
	IPO success (by *Scaled IPO proceeds*)	No IPO success (by *Scaled IPO proceeds*)
IPO success (by *Scaled IPO proceeds*)	16 (*TP*)	7 (*FN*)
No IPO success (by *Scaled IPO proceeds*)	10 (*FP*)	14 (*TN*)
	Predicted outcome by Random Forest algorithm (90/10 train–test split)	
	IPO success (by *Scaled IPO proceeds*)	No IPO success (by *Scaled IPO proceeds*)
IPO success (by *Scaled IPO proceeds*)	9 (*TP*)	4 (*FN*)
No IPO success (by *Scaled IPO proceeds*)	5 (*FP*)	7 (*TN*)
	Predicted outcome by Random Forest algorithm (60/40 train–test split)	
	IPO success (by *IPO underpricing*)	No IPO success (by *IPO underpricing*)
IPO success (by *IPO underpricing*)	42 (*TP*)	12 (*FN*)
No IPO success (by *IPO underpricing*)	27 (*FP*)	12 (*TN*)
	Predicted outcome by Random Forest algorithm (70/30 train–test split)	
	IPO success (by *IPO underpricing*)	No IPO success (by *IPO underpricing*)
IPO success (by *IPO underpricing*)	32 (*TP*)	9 (*FN*)
No IPO success (by *IPO underpricing*)	20 (*FP*)	9 (*TN*)
	Predicted outcome by Random Forest algorithm (80/20 train–test split)	
	IPO success (by *IPO underpricing*)	No IPO success (by *IPO underpricing*)
IPO success (by *IPO underpricing*)	20 (*TP*)	7 (*FN*)
No IPO success (by *IPO underpricing*)	14 (*FP*)	6 (*TN*)
	Predicted outcome by Random Forest algorithm (90/10 train–test split)	
	IPO success (by *IPO underpricing*)	No IPO success (by *IPO underpricing*)
IPO success (by *IPO underpricing*)	11 (*TP*)	3 (*FN*)
No IPO success (by *IPO underpricing*)	6 (*FP*)	4 (*TN*)

Note: This confusion matrix reports, at specified train–test splits, the numbers (in cells) of how many times the Scaled IPO proceeds and IPO underpricing random forest model predicts IPO success outcome (columns) correctly or incorrectly relative to the actual outcome (rows) of an observation. In cells, where both the predicted and actual outcomes correspond to IPO success, or where both correspond to no IPO success, the matrix reports the number of correct predictions. Otherwise, the cell reports the number of incorrect predictions.

The matrix is derived from 30 stratified splits. Therefore, TP, FP, TN, and FN are means across splits and rounded to whole numbers in this matrix.

The matrix is computed using the `numpy`, `pandas`, and `sklearn` Python library, with the function `sklearn.metrics.confusion_matrix`.

Appendix 7. XGBoost confusion matrices

Actual outcome	**Predicted outcome by XGBoost algorithm (60/40 train–test split)**	
	IPO success (by *Scaled IPO proceeds*)	No IPO success (by *Scaled IPO proceeds*)
IPO success (by *Scaled IPO proceeds*)	38 (*TP*)	16 (*FN*)
No IPO success (by *Scaled IPO proceeds*)	22 (*FP*)	17 (*TN*)
	Predicted outcome by XGBoost algorithm (70/30 train–test split)	
	IPO success (by *Scaled IPO proceeds*)	No IPO success (by *Scaled IPO proceeds*)
IPO success (by *Scaled IPO proceeds*)	30 (*TP*)	11 (*FN*)
No IPO success (by *Scaled IPO proceeds*)	18 (*FP*)	11 (*TN*)
	Predicted outcome by XGBoost algorithm (80/20 train–test split)	
	IPO success (by *Scaled IPO proceeds*)	No IPO success (by *Scaled IPO proceeds*)
IPO success (by *Scaled IPO proceeds*)	20 (*TP*)	7 (*FN*)
No IPO success (by *Scaled IPO proceeds*)	13 (*FP*)	7 (*TN*)
	Predicted outcome by XGBoost algorithm (90/10 train–test split)	
	IPO success (by *Scaled IPO proceeds*)	No IPO success (by *Scaled IPO proceeds*)
IPO success (by *Scaled IPO proceeds*)	10 (*TP*)	4 (*FN*)
No IPO success (by *Scaled IPO proceeds*)	6 (*FP*)	4 (*TN*)
	Predicted outcome by XGBoost algorithm (60/40 train–test split)	
	IPO success (by *IPO underpricing*)	No IPO success (by *IPO underpricing*)
IPO success (by *IPO underpricing*)	27 (*TP*)	19 (*FN*)
No IPO success (by *IPO underpricing*)	20 (*FP*)	27 (*TN*)
	Predicted outcome by XGBoost algorithm (70/30 train–test split)	
	IPO success (by *IPO underpricing*)	No IPO success (by *IPO underpricing*)
IPO success (by *IPO underpricing*)	21 (*TP*)	14 (*FN*)
No IPO success (by *IPO underpricing*)	15 (*FP*)	20 (*TN*)
	Predicted outcome by XGBoost algorithm (80/20 train–test split)	
	IPO success (by *IPO underpricing*)	No IPO success (by *IPO underpricing*)
IPO success (by *IPO underpricing*)	15 (*TP*)	8 (*FN*)
No IPO success (by *IPO underpricing*)	10 (*FP*)	14 (*TN*)
	Predicted outcome by XGBoost algorithm (90/10 train–test split)	
	IPO success (by *IPO underpricing*)	No IPO success (by *IPO underpricing*)
IPO success (by *IPO underpricing*)	8 (*TP*)	4 (*FN*)
No IPO success (by *IPO underpricing*)	5 (*FP*)	7 (*TN*)

Note: This confusion matrix reports, at specified train–test splits, the numbers (in cells) of how many times the Scaled IPO proceeds and IPO underpricing XGBoost model predicts IPO success outcome (columns) correctly or incorrectly relative to the actual outcome (rows) of an observation. In cells, where both the predicted and actual outcomes correspond to IPO success, or where both correspond to no IPO success, the matrix reports the number of correct predictions. Otherwise, the cell reports the number of incorrect predictions.

The matrix is derived from 30 stratified splits. Therefore, TP, FP, TN, and FN are means across splits and rounded to whole numbers in this matrix.

The matrix is computed using the `numpy`, `pandas`, and `sklearn` Python library, with the function `sklearn.metrics.confusion_matrix`.

Appendix 8. Random Forest variable importance

	Scaled IPO proceeds			
	60/40 (RF)	**70/30 (RF)**	**80/20 (RF)**	**90/10 (RF)**
AUC permutation importance:				
ln(Round equity, USD)	0.0209	0.0184	0.0344	0.0217
Age at IPO	0.0087	0.0018	0.0124	0.0137
VC rounds before IPO	-0.0011	-0.0021	0.0009	0.0059
Profitability	0.0717	0.0736	0.0860	0.0923
Δ PDP (p95) – PDP (p05):				
ln(Round equity, USD)	0.1052	0.0688	0.1073	0.0713
	IPO underpricing			
	60/40 (RF)	**70/30 (RF)**	**80/20 (RF)**	**90/10 (RF)**
AUC permutation importance:				
ln(Round equity, USD)	0.0109	0.0092	0.0064	0.0058
ln(Total assets before IPO, USD)	0.0049	0.0095	0.0092	-0.0028
Age at IPO	-0.0068	-0.0061	-0.0119	-0.0158
VC rounds before IPO	-0.0163	-0.0195	-0.0188	-0.0196
Profitability	-0.0009	-0.0029	-0.0031	0.0042
Δ PDP (p95) – PDP (p05):				
ln(Round equity, USD)	0.0018	-0.0085	-0.0037	-0.0043

Note: This table shows the mean permutation importance for the Scaled IPO proceeds and IPO underpricing random forest model variables over 30 repeated stratified splits and across train–test proportions of 60/40, 70/30, 80/20, and 90/10 evaluated on test sets only. The permutation importance scoring metric is AUC. Permutation importance values are computed with 50 random shuffles to determine how much model's mean AUC score degrades when each variable is permuted. Δ PDP (p95) – PDP (p05*)* is the difference of the average probability of Scaled IPO proceeds between the 95th percentile and 5th percentile of the variable of interest, *ln(Round equity, USD)*. *95% CIs* (confidence intervals) show the range of possible mean AUC permutation importance and Δ PDP (p95) – PDP (p05) values across the 30 repeated stratified splits. The table in Appendix 1 provides a detailed description of the model variables.

The AUC permutation importances are computed using the `numpy`, `pandas`, `statsmodels`, and `sklearn` (`sklearn.inspection.permutation_importance` function) Python libraries.

Δ PDPs is computed using the `numpy`, `pandas`, and `sklearn`, Python libraries.

Appendix 9. XGBoost variable importance

	Scaled IPO proceeds			
	60/40 (XGBoost)	**70/30 (XGBoost)**	**80/20 (XGBoost)**	**90/10 (XGBoost)**
AUC permutation importance:				
ln(Round equity, USD)	0.0329	0.0411	0.0401	0.0409
Age at IPO	0.0171	0.0023	0.0199	0.0125
VC rounds before IPO	-0.0029	-0.0081	-0.0067	-0.0017
Profitability	0.0634	0.0513	0.0893	0.0947
Δ PDP (p95) – PDP (p05):				
ln(Round equity, USD)	-0.0184	0.1435	-0.0038	-0.0051

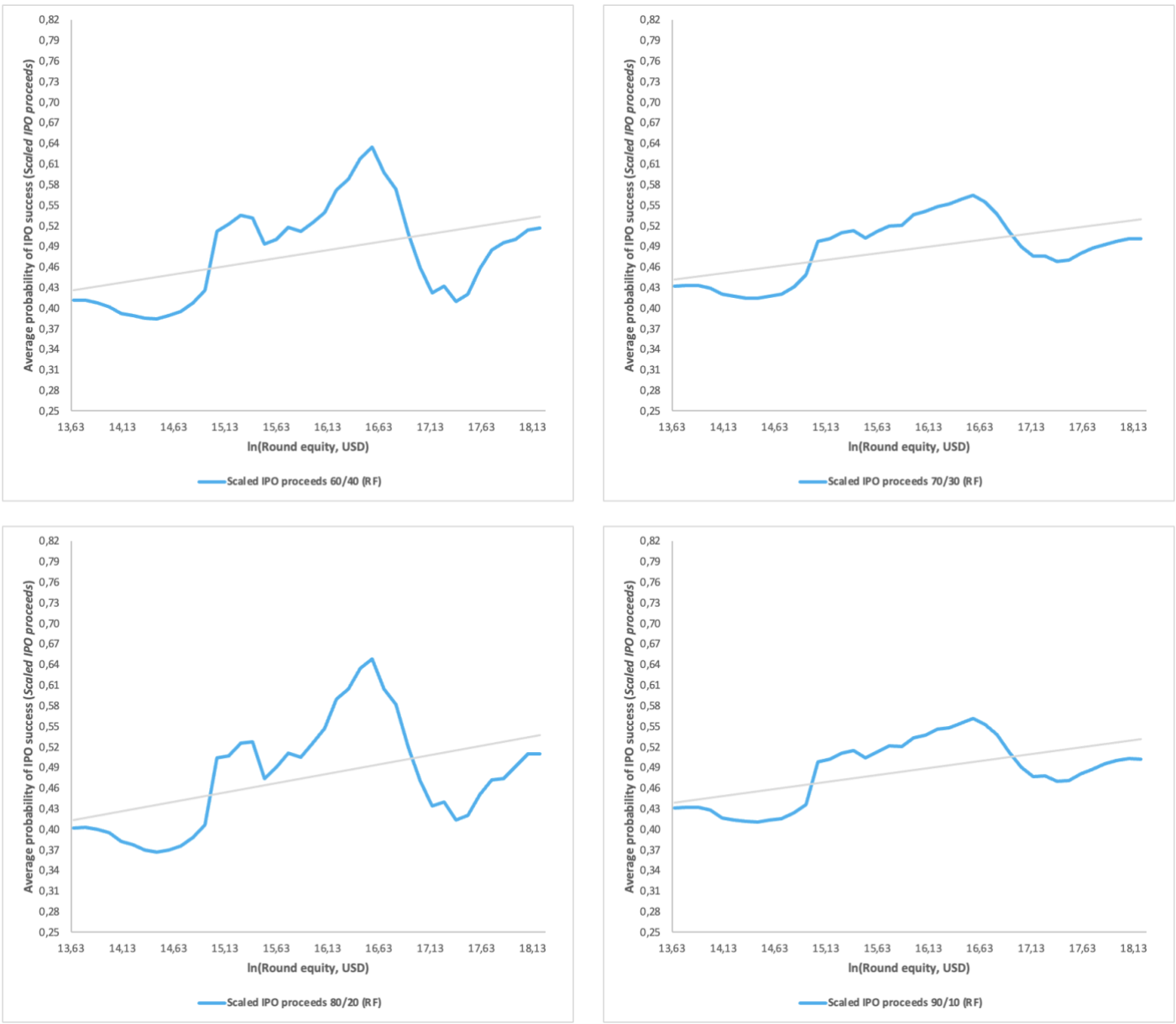
	IPO underpricing			
	60/40 (XGBoost)	**70/30 (XGBoost)**	**80/20 (XGBoost)**	**90/10 (XGBoost)**
AUC permutation importance:				
ln(Round equity, USD)	0.0292	0.0244	0.0201	0.0309
ln(Total assets before IPO, USD)	0.0079	0.0156	0.0116	0.0045
Age at IPO	0.0021	-0.0072	-0.0062	-0.0069
VC rounds before IPO	-0.0113	-0.0039	-0.0026	-0.0021
Profitability	-0.0032	-0.0024	-0.0019	-0.0019
Δ PDP (p95) – PDP (p05):				
ln(Round equity, USD)	-0.0165	-0.0833	-0.0505	-0.0330

Note: This table shows the mean permutation importance for the Scaled IPO proceeds and IPO underpricing XGBoost model variables over 30 repeated stratified splits and across train–test proportions of 60/40, 70/30, 80/20, and 90/10 evaluated on test sets only. The permutation importance scoring metric is AUC. Permutation importance values are computed with 50 random shuffles to determine how much model's mean AUC score degrades when each variable is permuted. Δ PDP (p95) – PDP (p05*)* is the difference of the average probability of Scaled IPO proceeds between the 95th percentile and 5th percentile of the variable of interest, *ln(Round equity, USD)*. *95% CIs* (confidence intervals) show the range of possible mean AUC permutation importance and Δ PDP (p95) – PDP (p05) values across the 30 repeated stratified splits. The table in Appendix 1 provides a detailed description of the model variables.

The AUC permutation importances are computed using the `numpy`, `pandas`, `statsmodels`, `xgboost`, and `sklearn` (`sklearn.inspection.permutation_importance` function) Python libraries.

Δ PDPs is computed using the `numpy`, `pandas`, `sklearn`, and `xgboost` Python libraries.

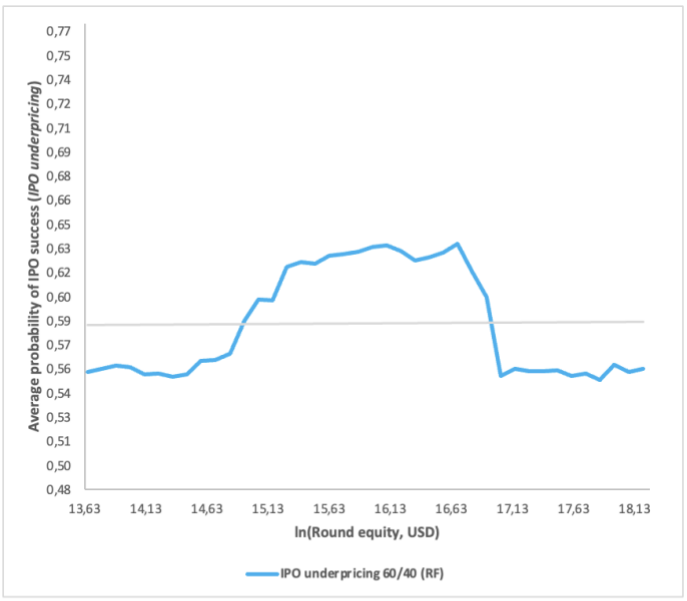
Appendix 10. PDPs (Random Forest, Scaled IPO proceeds)

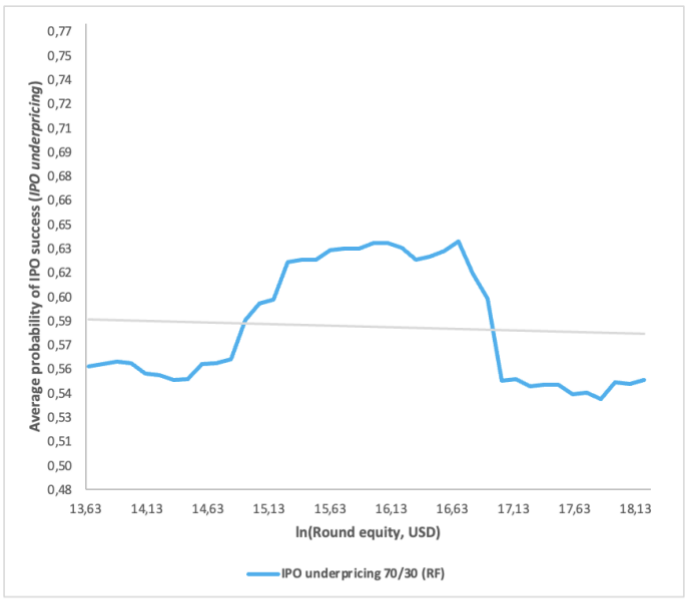


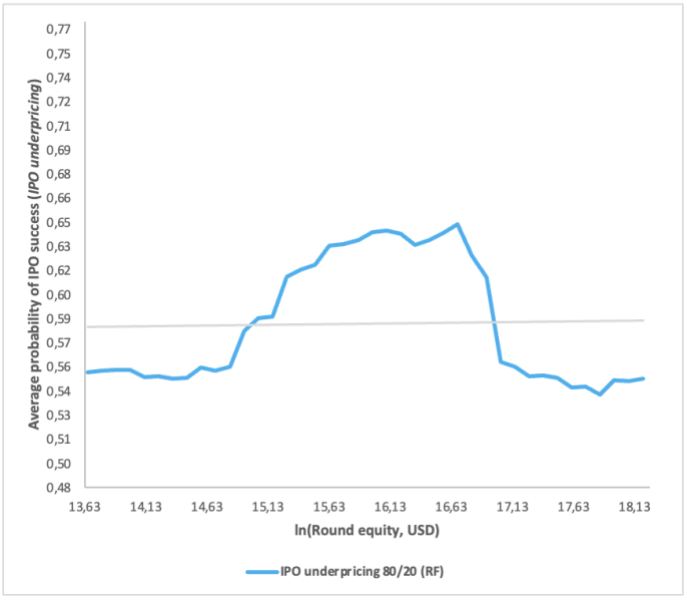
Note: These partial dependence plots (PDPs) show the relationship between *ln(Round equity, USD)* and IPO success, measured by *Scaled IPO proceeds*. The random forest PDPs (across train–test proportions of 60/40, 70/30, 80/20, and 90/10) illustrate the average probability of IPO success against the variable of interest, *ln(Round equity, USD),* over 30 repeated stratified splits. The partial dependence y-axis (the average probability of IPO success) represents the mean probability: for each fixed value of *ln(Round equity, USD)*, the predicted probabilities generated by a model are averaged across the sample, while keeping other predictor variables at their observed values. The *ln(Round equity, USD)* x-axis spans the 5th–95th percentile of *ln(Round equity, USD)* in the full dataset of the thesis. The grey trendline represents the slope of the PDP curve through the grid.

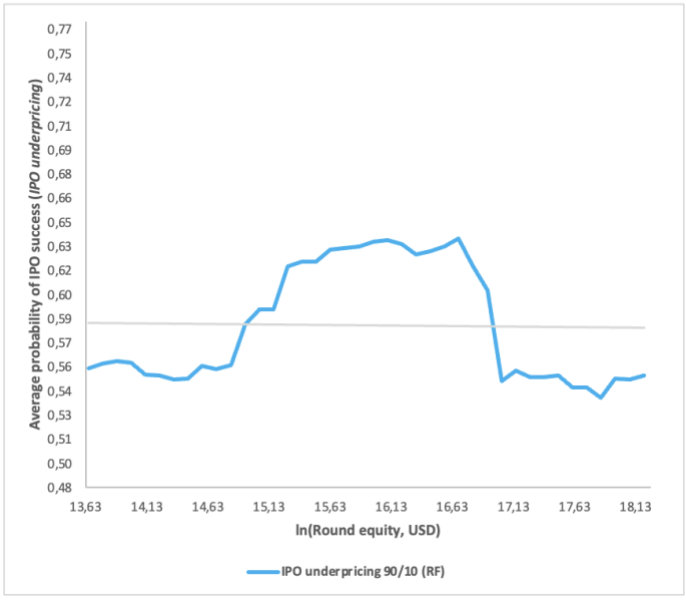
The PDPs are computed using the `numpy`, `pandas`, `statsmodels`, and `sklearn` Python libraries.

Appendix 11. PDPs (Random Forest, IPO underpricing)





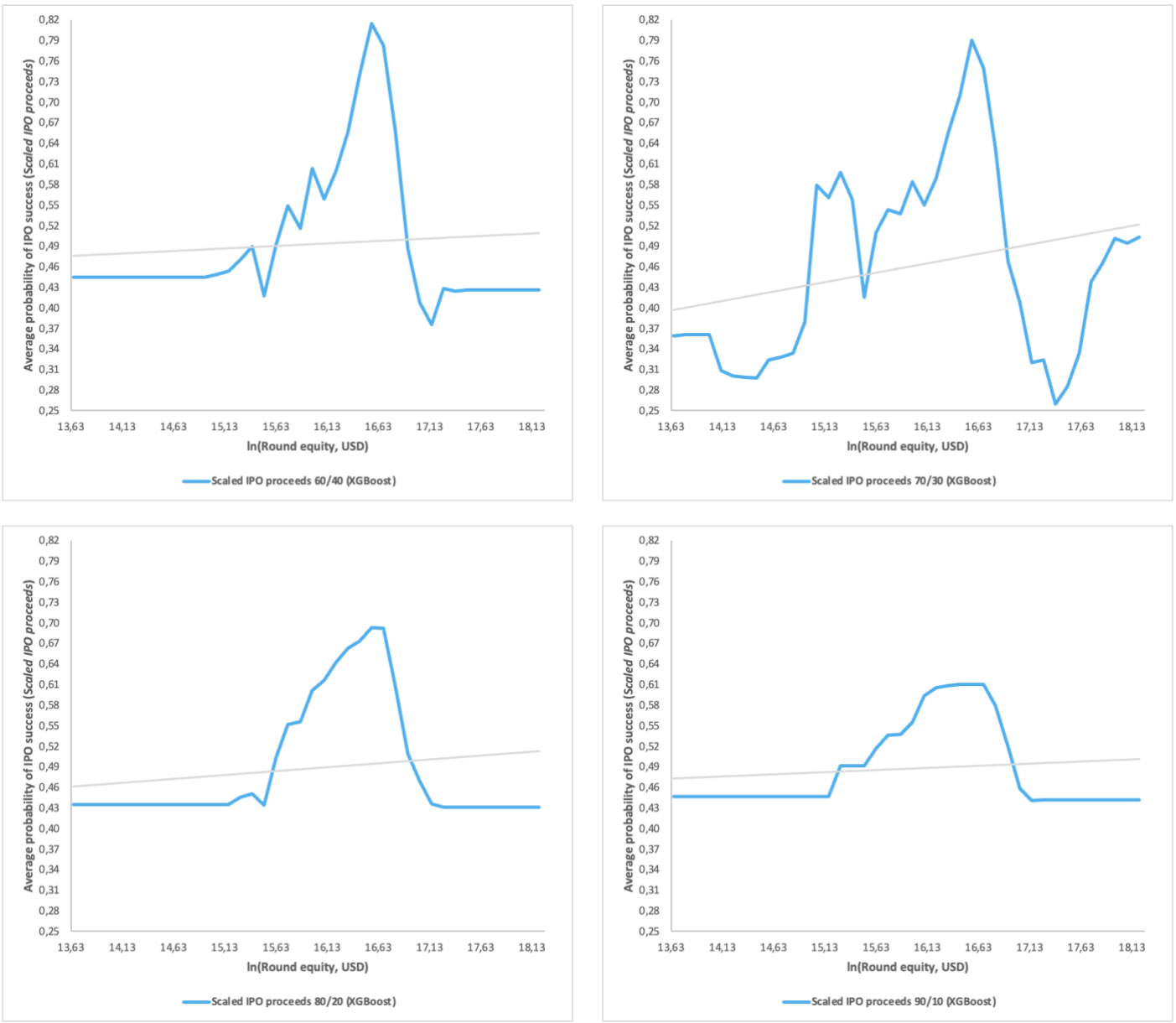




Note: These partial dependence plots (PDPs) show the relationship between *ln(Round equity, USD)* and IPO success, measured by *IPO underpricing*. The random forest PDPs (across train–test proportions of 60/40, 70/30, 80/20, and 90/10) illustrate the average probability of IPO success against the variable of interest, *ln(Round equity, USD),* over 30 repeated stratified splits. The partial dependence y-axis (the average probability of IPO success) represents the mean probability: for each fixed value *of ln(Round equity, USD),* the predicted probabilities generated by a model are averaged across the sample, while keeping other predictor variables at their observed values. The *ln(Round equity, USD)* x-axis spans the 5th–95th percentile of *ln(Round equity, USD)* in the full dataset of the thesis. The grey trendline represents the slope of the PDP curve through the grid.

The PDPs are computed using the `numpy`, `pandas`, `statsmodels`, and `sklearn` Python libraries.

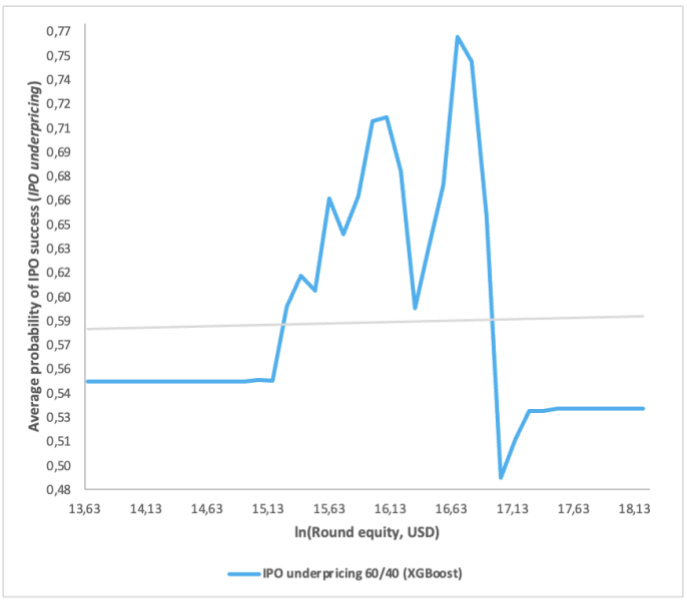
Appendix 12. PDPs (XGBoost, Scaled IPO proceeds)

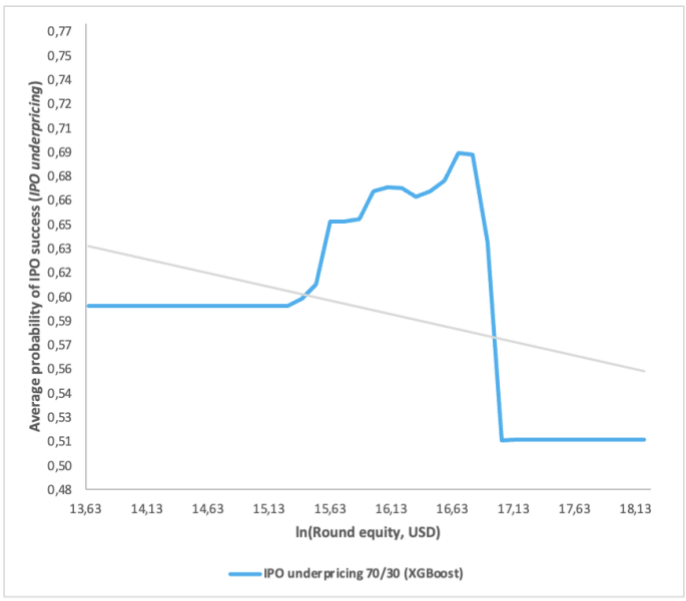


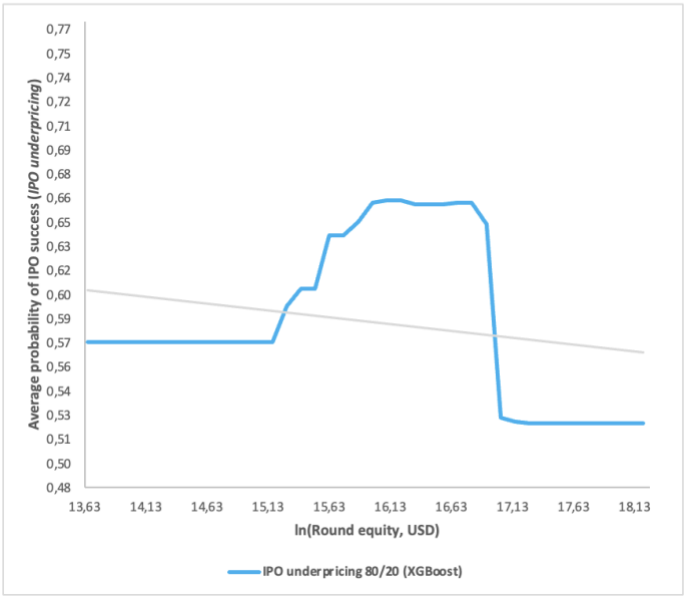
Note: These partial dependence plots (PDPs) show the relationship between *ln(Round equity, USD)* and IPO success, measured by *Scaled IPO proceeds*. The XGBoost PDPs (across train–test proportions of 60/40, 70/30, 80/20, and 90/10) illustrate the average probability of IPO success against the variable of interest, *ln(Round equity, USD),* over 30 repeated stratified splits. The partial dependence y-axis (the average probability of IPO success) represents the mean probability: for each fixed value of *ln(Round equity, USD),* the predicted probabilities generated by a model are averaged across the sample, while keeping other predictor variables at their observed values. The *ln(Round equity, USD)* x-axis spans the 5th–95th percentile of *ln(Round equity, USD)* in the full dataset of the thesis. The grey trendline represents the slope of the PDP curve through the grid.

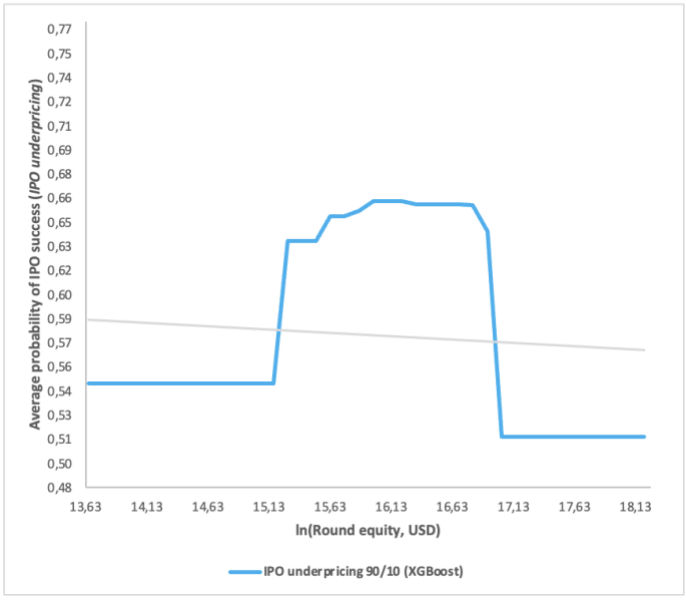
The PDPs are computed using the `numpy`, `pandas`, `statsmodels`, and `sklearn` Python libraries.

Appendix 13. PDPs (XGBoost, IPO underpricing)









Note: These partial dependence plots (PDPs) show the relationship between *ln(Round equity, USD)* and IPO success, measured by *IPO underpricing*. The XGBoost PDPs (across train–test proportions of 60/40, 70/30, 80/20, and 90/10) illustrate the average probability of IPO success against the variable of interest, *ln(Round equity, USD),* over 30 repeated stratified splits. The partial dependence y-axis (the average probability of IPO success) represents the mean probability: for each fixed value of *ln(Round equity, USD),* the predicted probabilities generated by a model are averaged across the sample, while keeping other predictor variables at their observed values. *The ln(Round equity, USD)* x-axis spans the 5th–95th percentile of *ln(Round equity, USD)* in the full dataset of the thesis. The grey trendline represents the slope of the PDP curve through the grid.

The PDPs are computed using the `numpy`, `pandas`, `statsmodels`, and `sklearn` Python libraries.