CATÓLICA LISBON
BUSINESS & ECONOMICS

BI Norwegian Business School

Using Big Data in Startup Selection: Exploring machine learning as a tool to predict successful startups in the age of social media

Barnabás Kiss

Dissertation written under the supervision of Alessandra Luzzi

Dissertation submitted in partial fulfilment of requirements for the MSc in International Management, at Universidade Católica Portuguesa and for the MSc in Business at BI Norwegian Business School, 2022. 09. 12.

Contents

Abstract

This research aims to further explore the possibilities in the usage of Machine Learning within the Venture Capital industry. Building on previous research the goal of this paper is to determine whether social media analyses can improve the accuracy of Machine Learning models to predict startup outcomes and valuations for startup companies. The research is built on the following models: Multilayer Perceptron, XGBoost, RandomForest, Naive Bayes, and Voting Regressor. The data used in this research comes from Crunchbase, USPTO, and Twitter.

The models in this research achieved an adjusted R2 of 0.5281 for value prediction, which shows that exit value is explainable to a large extent by using publicly available qualitative and quantitative data. Outcome prediction had precision for IPO between 0.1447 to 0.4193 and F1-scores between 0.2360 to 0.4449 for models built from Series A to Series C funding rounds.

The results of this research show that Venture Capital firms investing from Series A to Series C would be able to outperform the market in terms of returns by implementing Machine Learning in their investment decision-making process. To further improve these results extracting further social media data is a beneficial future resource. Compared to previous models this research built models for 3 specific early funding rounds and can outperform the markets with data available for VCs at these points in time.

Keywords: Venture Capital; Machine Learning; MLP; XGBoost; Random Forest; Naive Bayes; Voting Regressor; Value Prediction; Outcome Prediction; Investment Strategy;

Title: Using Big Data in Startup Selection: Exploring machine learning as a tool to predict successful startups in the age of social media

Author: Barnabás Kiss

Abstrato

Esta investigação visa explorar mais profundamente as possibilidades de utilização da aprendizagem mecânica na indústria do Venture Capital. Com base em pesquisas anteriores, o objectivo deste trabalho é determinar se as análises dos meios de comunicação social podem melhorar a precisão dos modelos de Machine Learning para prever os resultados e as avaliações das empresas em fase de arranque. A investigação baseia-se nos seguintes modelos: Multilayer Perceptron, XGBoost, RandomForest, Naive Bayes, e Voting Regressor. Os dados utilizados nesta pesquisa provêm de Crunchbase, USPTO, e Twitter.

Os modelos nesta pesquisa alcançaram um R2 ajustado de 0,5281 para previsão de valor, o que mostra que o valor de saída é explicável em grande medida através da utilização de dados qualitativos e quantitativos disponíveis publicamente. A previsão de resultados teve precisão para IPO entre 0,1447 a 0,4193 e pontuações F1 entre 0,2360 a 0,4449 para modelos construídos das séries A a séries C de financiamento.

Os resultados desta investigação mostram que as empresas de Venture Capital que investem da Série A à Série C seriam capazes de superar o mercado em termos de retorno, implementando a Machine Learning no seu processo de tomada de decisões de investimento. Para melhorar ainda mais estes resultados, extrair mais dados dos meios de comunicação social é um recurso futuro benéfico. Em comparação com modelos anteriores, esta investigação construiu modelos para 3 rondas de financiamento antecipado específicas e pode superar os mercados com dados disponíveis para VC nestes pontos no tempo.

Palavras-chave: Venture Capital; Machine Learning; MLP; XGBoost; Random Forest; Naive Bayes; Voting Regressor; Value Prediction; Outcome Prediction; Investment Strategy;

Título: Using Big Data in Startup Selection: Exploring machine learning as a tool to predict successful startups in the age of social media

Autor: Barnabás Kiss

Introduction

Evaluating startups has a long history but in the past, it has faced a lot of constraints. Firm evaluation traditionally has been based on financial performance and for public companies, the available information makes it possible to make informed decisions. For startups, the evaluation faces the constraint that past performance is limited or does not exist. Despite this, the venture capital industry has been growing rapidly since its inception. Most firms rely on qualitative analysis next to (if it exists) financial analysis. Qualitative analysis may rely on score cards (Bai & Zhao, 2021). Previous research also suggested that an average VC firm enjoys limited success during its life cycle (Hong, Serfes, & Thiele, 2020).

The parties to a Venture Capital investment include the VC firm as the investor and the startup which receives the investment. The two main operations for the investor can be characterized as venture selection and monitoring. Improvement in venture selection and monitoring thus benefits both parties and further democratization could be beneficial for the sector and the economy. One method to achieve this is Machine Learning (ML), which has been researched previously by academics (Xiang et. al, 2012.) (Ross, Das, Sciro, & Raza, 2021.) (Krishna, Agrawal, & Choudhary, 2016). The results showed positive results based on quantitative self-reported data from Crunchbase with additional use of publicly available data from USPTO.

Previous research has also suggested the possibility of using this method for the monitoring process by VCs could further benefit the sector. The monitoring aspect has received less attention in academic research and the existing research mostly focused on theoretical aspects than practical applications. This can be explained by the fact that an outcome predictor can be used for the monitoring aspect as well, when it signals a change in the prediction VCs can use that to increase due diligence. For this reason, separate research is not needed to solve this inefficiency.

This leads us to our research questions:

Can Venture Capital firms use machine learning for better venture selection at pre-Seed, Series A, B, and C funding rounds? Can Machine Learning be used for predicting exit value? Can social media analysis be useful for Venture Capital firms in evaluating investment opportunities?

Previous research showed us there is a great potential for utilizing machine learning for VCs and startups to predict exit outcomes. Even though such research exists its results are limited, and real-life adaptations were also limited. With a successful tool that can point out the expected outcome, investment decision performance could increase in an industry that has high risks. Potential use for monitoring activities could potentially be beneficial in the future. There is also a geographical limitation observed for VCs, they prefer to operate in familiar areas. A better tool for selection could also decrease this limitation and level the playing field by improving access to capital globally.

Compared to the previous models in the literature, the models were built for this research using a narrow time-specific sub-section of available data from Series A to Series C funding rounds. At this specific time section, these models outperform the market by a considerable margin using the performance metrics available. Previous research for VC firms used all available data for observations in the Crunchbase dataset which led to unclear use cases for the industry and target leaks. This research aimed to bridge that gap by having specific time points to feed data into the models and removing any data that could cause a target leak.

Literature review

Startups and Venture Capital (VC) firms have been subject to increased research in the last 30 years, due to their importance for innovation and the economy. The following chapter is going to provide a VC industry overview to understand the current trends and market dynamics and the current state of research about the entrepreneurial exit. Then an overview of environmental and startup factors for success aims to overview all aspects of success from literature. The final part of the literature review will overview the current state of research about machine learning in the VC industry.

Current trends in the VC industry

Analyzing the venture capital industry trends requires understanding the current trends for Initial Public Offerings (IPOs) and acquisitions within the wider ecosystem that involves the startups to measure it. Venture Capital (VC) firms usually invest in more than one sector and current trends influences where the capital is flowing. VCs have been defined as conductors of capital management that flow from traditional industries to more risky innovative business ventures. The legal framework for this has the VC side, general partners as asset managers, and the investors as Limited Partners, starting from 2009. VC firms also increased their offerings besides investments with additional services including marketing and hiring advisors to startups. The market is also generally invested in high-tech firms, but their area of operation can change significantly from fund to fund and with time. (Nicholas, 2019)
Kim et. al (2021) provided an overview of these trends for the past 10 years and analyzed them in their paper. Their findings showed that according to Crunchbase the number of mergers and acquisitions (M&As) has increased from 9256 between 2010. 1. 1-2012. 12. 31 to 35943 between 2016. 1. 1-2019. 12. 31. Their findings also showed that industry convergence is dependent on the industry they operate in.
Pisoni & Onetti (2018.) analyzed the geographical effects. Acquirers like shop locally, thus M&As are more likely to happen if the local or regional incumbents are large and concentrated, with geographical and spatial distance particularly significant when observing two countries. This can be an important factor for models when trying to predict the future as analyzing incumbents can lead to an important differentiation between an M&A or an IPO based on market features. (Pisoni & Onetti, 2018.)

The performance of VC firms has limited information available to the public but relying on previous research we can have an overview of the industry performance over time. Nicholas (2019) found data that showed a 1.29% return over the benchmark (S&P Composite) for vintage VC funds between 1981-2006. Top Quartile VC funds had a 4.12% return over the benchmark and Sequoia Capital had a 5.8% score. (Nicholas, 2019)

On the other hand, we have Kaplan and Lerner's (2010) publication, which analyzed the financial performance of US VCs. They observed that IRR was more or less equal to that of NASDAQ during the observed period. (Kaplan & Lerner, 2010)

We can also analyze performance based on the definition by Hong et. al. (2020), which has more available data to gauge industry metrics. They define success as IPO or M&A and measure VC performance by analyzing successful exits by their portfolio. They divide VCs into two categories, above and below the 90th percentile based on experience. They measure VC success (based on their definition) between 1990 and 2010 and calculate the 90th percentile based on that. Their findings show a 17% successful exit rate for below 90th percentile VCs and a 24% successful exit rate for the top 10 percentile VCs. (Hong et. al., 2020)

Ross et. al. (2021) on the other hand claim that successful VC firms have roughly 20% of their portfolio reach an IPO. They also claim that 95% of VC firms fall into the unsuccessful category.

These contradicting claims mean that our research will rely on Hong et. al.'s metrics the most and also use the 20% claim by Ross et. al. (2021) to evaluate the models built.

Entrepreneurial exit

Research about entrepreneurial exit has shown that entrepreneurs' goals do not differ drastically from VC goals. Wennberg et al. (2010, 2014) in two papers have framed the idea of an entrepreneurial exit from the consensus that the founder(s)' view it negatively to the idea that founders can also look at this outcome positively. In the 2010 paper, Wennberg et al. argue that previous research neglected serial entrepreneurs and portfolio entrepreneurs in their definition of entrepreneurial exit. They also note that family firm research has more extensive knowledge about the different strategies a firm should pursue handover to the next generation or sales outside the family. (Wennberg & DiTienne, 2014.) (Wennberg et. al., 2010.) This points to the fact that a more strategic approach to the cooperation between startups and VCs would be beneficial, with a better approach the benefits could be greater for both actors. VCs pursue potential unicorns for their portfolio because one unicorn can deliver enough return to shift the VC into the positive. An issue with this is that previously we noted that VCs could improve in successful exit prediction when pursuing startups and better avoidance of failures could lead to better financial returns to investors.

The most common positive (successful) outcome for startups is acquisition. This can be divided into two types: early or late sales. Arora et. al (2021) found that the capability of the founding team decided whether they committed to an early or late exit, whereas better capabilities shifted it towards a late exit. They also saw that intermediate capabilities remain flexible, and those seek offers early, but usually sell late. (Arora et. al., 2021.)

Startup M&As have been analyzed from the point of view of incumbents as well, the potential pitfalls of integration can inform us of relevant factors that these incumbents can and or will in the future consider. Research on this topic is following the case study approach. The national cultural fit and low performance of the integration team have been identified as more specific pitfalls, but more case-specific problems relating to the startups or acquirers have also been identified. (Kurshunova et. al., 2021.)

Environmental factors

There has been extensive research about the effects of the economic environment on startup success. It has been well documented that a positive environment managed by policymakers has a positive net effect on the economy, which in turn leads to an increased interest in this area for research. Policymakers have competed to provide the best environment to attract new ventures and nurture innovation. More specifically there have been numerous papers published regarding startup exits. Ogane (2015) analyzed how competition among financial institutions relates to startup company exit. He concluded that these two are positively correlated when a start-up company is defined as five years or younger, or ten years or younger. There is a negative correlation between the probability of bankruptcy and competition among financial institutions (Ogane, 2015.).

Other authors have analyzed the impact of policy uncertainty on the M&A of startup firms. Cotei et. al (2021.) found that an increase in policy uncertainty causes a decrease in the M&A of startups. The cause they highlight is the increase in the risk premium of targeted firms and the cost of developing absorbing capacity for the buyer. They also note that startups that build competitive advantage and also possess intellectual property are more likely to exit via M&A even in highly uncertain environments. (Cotei, Farhat, & Khurana, 2021.)

Hong et. al (2020) examines the effect of competitiveness in the VC market on successful exits defined as M&A or IPO. They found that a decrease in the Herfindahl-Hirschman Index by 50% from its mean results in a 2.8% increased chance for success. (Hong, Serfes, & Thiele, 2020).

Startup attributes

Another recent focus of research has been social media and its connection to startup success. Media and social media analysis has been more common in the stock market (Pineiro-Chousa, Vizcaíno-González, & Pérez-Pico, 2017.), Jiao et al. (2020) is one example of how stock exchange investing can benefit from analyzing social media and media coverage for investments (Jiao, Veiga, & Walther, 2020.). For the VC industry, interesting research has been conducted by Gloor et. al (2020), which tried to determine the impact of the social media presence of the board members on technology startups. The active Twitter presence of board members has resulted in additional funding but did not affect sales. This paper used regression models and these findings were on the 95% confidence interval with 0.426 adjusted R^2 (Gloor, Colladon, Grippa, & Hadley, 2020.).

The background of the founders has also received attention. Roche et. al (2020) found that startups by academics have comparatively worse performance than startups founded by non-academics. This research was limited to startups in biomedicine, which leads to a more limited scope of research, limiting how many conclusions we can draw from it. (Roche, Conti, & Rothaermel, 2020)

Gender roles in the leadership of organizations have been researched in previous years. One exploratory study has found higher sales growth, higher ROI, and higher ROA for firms managed by female CEOs. These findings were distorted by the fact that only around ~1.22% of the observed firms were managed by female CEOs. (Jalbert, Jalbert, & Furumo, 2013.)

Another study focused on France, as the country passed extensive legislation to increase the number of female board members for French firms. The study using quantile difference-indifferences and dose-response models has found that high-performing firms increase their financial performance after the increase of female board members, but low-performing firms will see a decrease in performance. The observed firms were publicly listed, so investor sentiment also played a part and was not controlled in this research. (Slama, Ajina, & Lakhal, 2019.)

Machine Learning in Venture Capital

Data-driven approaches have been suggested before for Venture Capital firms to improve their performance in either selection or evaluation of startups.

One model to create startup company portfolios based on data-driven approaches has been created not utilizing ML models. One such study utilized Crunchbase data with a Bayesian modeling framework for evaluating startups and then used a Brownian model for the rest of the calculations to construct the best portfolio. This approach was successful in portfolio optimization without introducing ML to the model. (Hunter, Saini, & Zaman, 2017.)

Zhong et. al (2016) also suggested that a more methodological approach would be beneficial for the Venture Capital industry. In their approach, they were also utilizing Crunchbase data with a personalized portfolio strategy based on ones that are used on the stock market. They aimed to introduce a tool that can be used according to the investors' risk preferences while maximizing return. They successfully built a model that could implement investors' risk preferences and optimize portfolios based on geographical and industry preferences. (Zhong, Liu, Zhong, & Xiong, 2018.)

Xiang et. al. (2012) utilized TechCrunch news articles and the early version of the Crunchbase dataset to predict company acquisitions by using qualitative features. Their models performed well by True Positive, False Positive, and Receiver Operating Characteristics metrics. This research utilized Bayesian networks for its models. (Xiang, et. al, 2012.)

Krishna et. al. (2016) built a model that targeted for startups to predict their outcome in the earlier stages. They saw that it could be adapted by startups to monitor their progress. Crunchbase data was utilized for the research and included numerous ML models. They limited independent variables by funding round stage to solve the non-linear issue of the Crunchbase dataset. Their models achieved high precision, recall, and accuracy metrics. (Krishna, Agrawal, & Choudhary, 2016)

Bai & Zhao (2021) explored the possibility to use ML to evaluate VC Scorecards for investment decision support. They found that ML was able to replicate human decision-making with high accuracy and could be implemented for the VC deal-sourcing process. (Bai & Zhao, 2021)

Ross et. al. (2021) built models that were targeted at VC firms to explore the potential upsides of utilizing ML in investment strategy. They built on the Crunchbase dataset and utilized USPTO for patent data. The models achieved high precision and recall but used some variables that enabled target leak or their added values were unclear. Examples would include the number of acquisitions made by the organization, length of Crunchbase description, high employee count variables (10000+, 5000-10000..), all funding data up to series J, total funding, and the number of top degrees. The results showed that ML could potentially be used in Venture Capital but the model did not take into account target leak as these variables perform well for companies that reached one of the success metrics. A company with 10000+ employees is more likely an IPO for example. The promising model metrics show that there is actual potential in the area for practical use. (Ross, Das, Sciro, & Raza, 2021.)

Research Methodology

Data to be Used

The main dataset to be used is the Crunchbase dataset. This includes data for more than a million companies and has been used in previous research. The academic access provides us with the raw CSV files and with the opportunity to use the API.

Previous research has gathered data for patents by USPTO, despite Crunchbase data having two variables about this, patents granted and most popular patent class. This is explained by that the USPTO is a more accurate dataset that is updated by state actors. (Ross, Das, Sciro, & Raza, 2021.)

Previous research has concluded that social media presence is an important indicator of startup success. However, there has been limited research about the effect of social media performance on startup success. Twitter has been identified as an important indicator of success measured by the presence and possible influence (size of followers) of board members. Twitter has an easily usable API, that will make it possible to use it in an ML environment. (Gloor, Colladon, Grippa, & Hadley, 2020.)

Models to be used

This research is going to use a machine learning method for research. The environment to be used is going to be Python, in line with previous research conducted on the topic. This research will expand on the models used by Ross et. al. (2021.) by introducing new variables and improving the quality of the dataset used, specifically targeting investors by funding round.

MLP – Multilayer Perceptron

Multilayer Perceptron models are deep-forward neural networks with the possibility of multiple setups. (Ross, Das, Sciro, & Raza, 2021.) The models in this research all used the relu activation with the adam solver. The adam solver was needed to get the best results and have the MLP converge with the smallest used dataset size (55835 observations). The layers used in the models are dependent on the models. after testing the MLP2 and MLP3 Binary classification models reached better results using 7 layers of 32 neurons, while the other models used 10 layers of 32 neurons. The MLP Success pre-seed model used 20 layers of 32 neurons.

XGBoost

XGBoost is an open-source software library first developed by Tianqi Chen. It is a scalable end-to-end tree boosting system and it is used widely used by data scientists to solve ML challenges. It is described as the de-facto choice of the ensemble method. (Chen & Guestrin, 2016.)

XGBoost provides an efficient model that enables this research to test different data inputs and arrive at the best solutions without encountering hardware and time limitations. The final XGBoost models are using the following parameters:

```
params = {
        'objective':'binary:logistic',
        'max_depth': 4,
        'alpha': 10,
        'learning_rate': 1.0,
        'n_estimators':1000
    }
```

Random Forest

Random Forest is another ensemble method that was developed by Tin Kam Ho in 1995. In this model classifications from multiple decision trees are combined to produce a more robust classifier. (Ho, 1995.) The best results were achieved with 100 classification trees and max depth was not set to any value.

Naive Bayes

Naive Bayes has been described as one of the most effective and efficient inductive learning algorithms to be used for machine learning. The relative performance compared to other algorithms is surprising because it is based on conditional independence assumption that rarely holds in the real world. Zhang (2004) proposed that the strong performance is because naive Bayes will cancel out dependencies if they are evenly distributed. In their paper, they provided sufficient evidence to support their proposal. (Zhang, 2004)

Regression model – Voting Regressor

The Crunchbase dataset also contains extensive data on valuations of acquisitions and IPOs for companies. Although the data is not complete for the whole database it has enough data points to explore. Previous research has not explored the possibility of predicting the future value for companies based on earlier qualitative and quantitative data, so this research will attempt to bridge that gap.

The regression chosen for this study is the Voting Regressor from the Scikit Python application. Within the Voting Regressor, we have tested the following 7 regressions: Gradient Boosting Regressor, Huber Regressor, Elastic Net, Extra Trees Regressor, K-Nearest Neighbors Regressor, Light Gradient Boosting Machine Regressor, and the Passive Aggressive Regressor. (Pedregose et. al, 2011) (scikit-learn, 2022)

Throughout the model engineering process, we narrowed down the following 3 models to be useful for explaining the valuations:

- Gradient Boosting Regressor [0.85]
- Extra Trees Regressor [1]
- Light Gradient Boosting Regressor [0.7]

All 3 models used 1000 for the n_estimators value and the weights for Voting Regressor can be seen in brackets above.

Definition of outcomes and stage

The following definitions for outcomes and stages are going to be used in this research. It is necessary to clearly define the dependent variables (outcomes) to build models that have practical implications for Venture Capital. A clear definition of the stage is necessary to prevent target leaks in a non-linear dataset.

Outcomes:

- *IPO (success):*

  Startups that had an IPO (Initial Public Offering). This is the most strict but clearest way to define startup success.
- *Acquisition (success):*

  Startups that have been acquired by other companies. This is also a clear way to define success, but it has one flaw related to the database. The acquisition is a less clear signal for actual success, in some cases, failures are also categorized as acquisition. These would be the cases where a failing startup is acquired by some company due to interest in the team or some technical knowledge that represents value for that company, but that is not such a value that it represents success for the owners of the startup. It is a good way to recover some of the investment but cannot be defined as a clear success. For this reason, the separation of IPO and Acquisition is justified if the initial set has enough entries to be able to train and test a model for both outcomes separately. The earlier the stage is the more justified it is to combine both since these outcomes will represent less and less of the total observations.
- *Operating:*

  Startups that are currently operating but that have not exited. Due to Venture Capital firms' preference for acquisitions and IPOs, this functions as a control group and this is not a metric of success. Due to the large size depending on the observed outcome and scope of the model, this dataset can be modified in size to better support the models.
- *Failure:*

  Startups that are currently not operating anymore. This is a very good indicator of actual failure and useful input to the model.

Definition of stage:

The aim of this is to set up a clearly defined life stage for each company to prevent target leaks in this non-linear dataset.

This was achieved by using the industry definitions for life-stage, these are the following defined by the latest funding round or exit:

```
'acquisition', 'ipo', 'seed', 'pre_seed', 'series_b', 'angel',
       'series_a', 'series_c', 'series_d', 'series_e', 'series_f',
       'series_g', 'series_h', 'series_i', 'series_j'
```

The first aim of this was to use to identify the companies relevant to the models. When using a model that includes data up until Series B for Series C investors, the observed companies would be limited to the following stages:

```
'acquisition', 'ipo', 'series_b', 'angel', 'series_c', 'series_d',
'series_e', 'series_f', 'series_g', 'series_h', 'series_i', 'series_j'
```

The implications are best explained by the following example that uses holds for Series C investment models:

By limiting observations with this metric we only observe startups that could be in front of a Series C investor. Our model then fulfills the Series C investors' role according to this metric. This prevents the model to overperform due to it identifying exited startups at the early stage through later-stage funding data. It also lets our model make use of the pre-seed, seed, Series A, Series B, and other early funding data (grants, crowdfunding, etc.). The description would hold for earlier-stage investment models with the difference that a Series B investor model has no access to Series A funding.

Data Engineering

We have engineered 5 separate datasets that were used in this research:

- Regression dataset with valuations:
  This dataset was limited to observations that were either acquired or had an IPO, and the Crunchbase dataset had a valuation for the exits. This dataset was limited to data that is available up until and including Series B funding data. This model would be useful for Series C investors.
- Classification dataset for the following life stages: Early indicators for pre-seed investors, Series A, B, and C investors. Each set includes funding data until the previous funding round.
- The Pre-seed, Success is a modified pre-seed dataset where operating observations have been removed.

The following table details the datasets by observations, number of features, and class distribution:

			Class distribution			
Dataset	Observations	Features	IPO	Acquisition	Operating	Failed
Series C	55835	692	13.4%	60.6%	16.9%	9.6%
Series B	70933	688	10.6%	47.3%	32.9%	9.2%
Series A	119603	684	6.3%	28%	55.3%	10.3%
Pre-seed	360840	390	2.1%	9.3%	80.6%	8%
Valuation, Series C	30234	686	-	-	-	-
Pre-seed, Success	70050	390	58.6%		-	41.4%

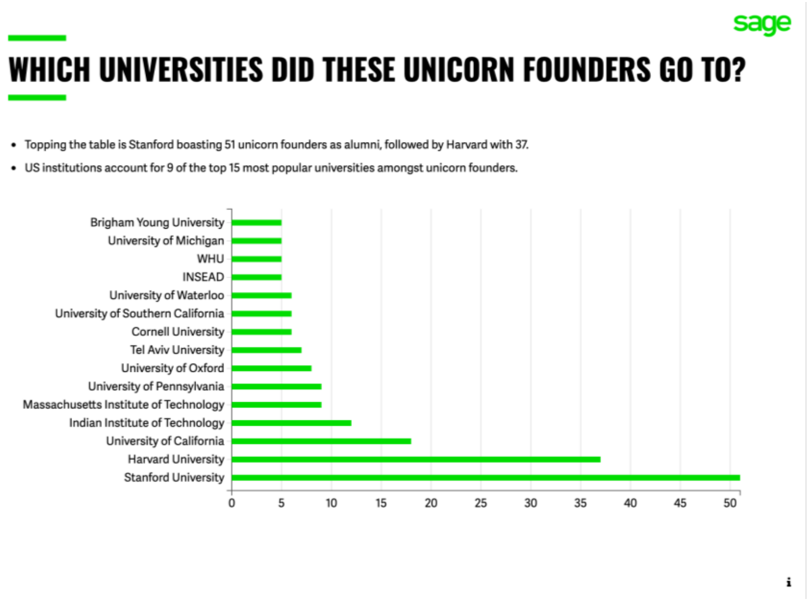
Table 1 – Detailed description of the used datasets.

*For Pre-seed, 58.6% represents IPOS+Acquisitions that were combined for those models

Educational background

Educational background has a skill and network effect that may be relevant for startup founders. For this research, we narrow down the list of successful founders to Unicorn founders. Using that list and the Crunchbase dataset we can narrow down the list of educational institutions to a more manageable size of 376 schools. With the list of educational institutions, we can create 376 dummy variables that note a founder-level connection between them and an observed startup. This can be a finished and a not-finished education for the founders. This scope might limit the findings, but it is necessary to narrow it down from the 10s of thousands of institutions in the Crunchbase dataset.

Previous research has shown that educational background is an important predictor for successful startup founders. Both the Crunchbase dataset used by this research and research by Sage (Verve Search, 2019) provided data that shows a huge discrepancy between educational institutions and the number of unicorn founders they produced. This is also the justification for the limitation decision described above. Lower ranked schools are unlikely to have a significant value-added for a startup founder that would be detected by the models. The following figures show the previously described observations that were used for the data engineering decisions.



WHICH UNIVERSITIES DID THESE UNICORN FOUNDERS GO TO?

- Topping the table is Stanford boasting 51 unicorn founders as alumni, followed by Harvard with 37.
- US institutions account for 9 of the top 15 most popular universities amongst unicorn founders.

Figure 1 – Universities by number of unicorn founders – Sage (Source: Sage, Verve Search, 2019)

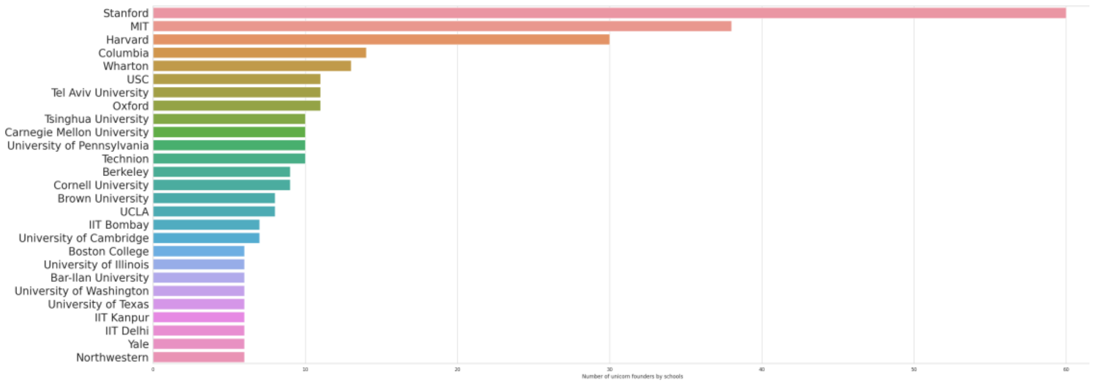


Figure 2 – Universities by number of unicorn founders – Crunchbase (Source: Crunchbase, 2022)

This variable is constructed based on a list of educational institutions that have produced unicorn founders. This also includes non-graduates due to the possible network effect. The connections are limited to founders to reduce success bias by organizations with a lot of employees and to introduce a limit on the scope. All startup organizations, therefore, have a dummy variable for each educational institution where 0 represents no established connection, 1 represents an established connection with at least 1 founder and NaN represents missing data points. Due to the models used and the characteristics of the dataset, NaNs were reconstructed as 0s for the models, because most of the models cannot handle NaNs. This was justified by satisfying coverage within the dataset across the observed startups.

Geographical location

The geographical location of a startup limits its access to capital and talent. Controlling for such factors is therefore a desired feature for any model that is trying to predict startup value and outcome. Engineering decisions were based on the Crunchbase dataset. Due to the clear signs that using country code could have a limitation for the United States, in that case, the 50 states were used instead.

After analyzing the dataset, we obtained the following data that showed a possible geographical effect.

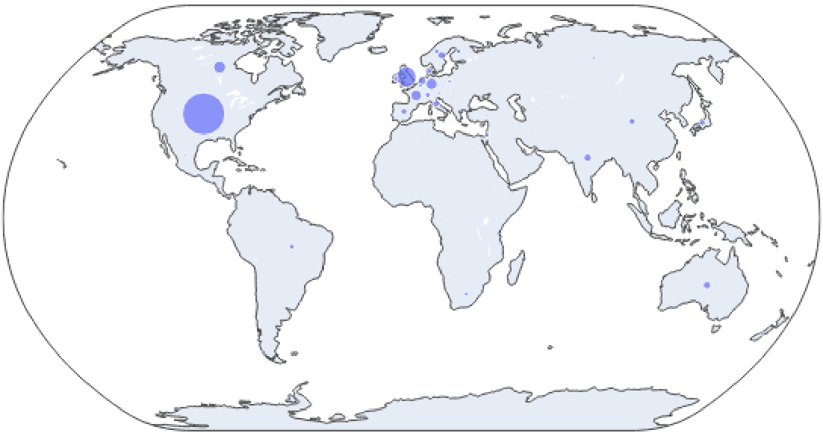


Figure 3 - Number of acquisitions by acquiree country code (Source: Crunchbase, 2022)

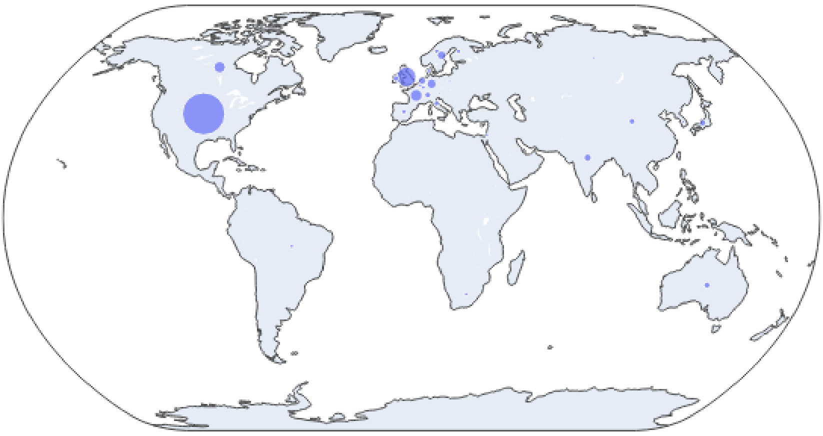


Figure 4 - Number of acquisitions by acquirer country code (Source: Crunchbase, 2022)

The two figures show that the number of acquisitions is influenced by geographical locations. Two clusters for startups can be identified, one being in Europe and one being in the United States. The European one is gravitating towards London, one of the main startup hubs in Europe.

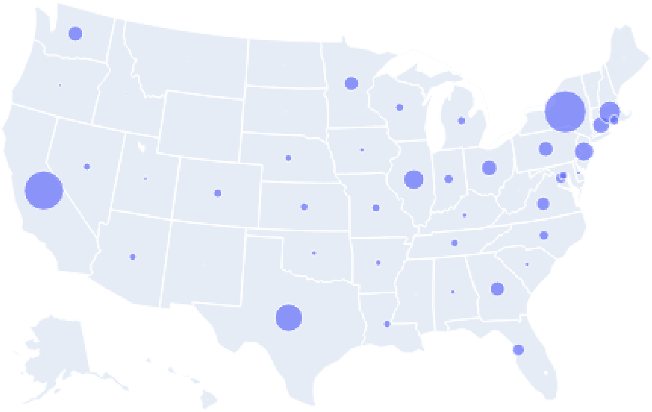


Figure 5 – Number of acquisitions in the US by states (Source: Crunchbase, 2022)

When looking at the number of acquisitions in the US by states we can see that the geographical hub effect exists within the United States. This implies that choosing a location is an important feature to control for startups.

We can also observe similar effects for IPOs:

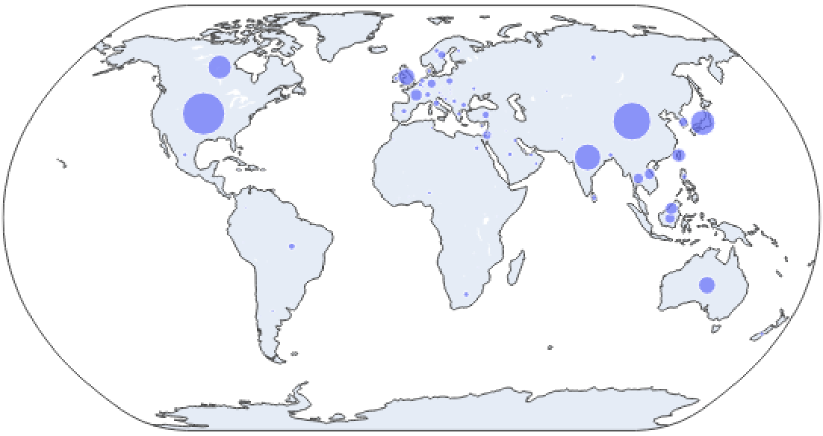


Figure 6 – Number of IPOs by country in the dataset (Source: Crunchbase, 2022)

These findings all support the decision to include geographical location as a variable and to use US states instead of the whole country for the largest market within the startup ecosystem and VC industry.

The geographical features of the dataset contain 233 variables, including all 50 US states.

Founder track record and background

The track record of founders has been identified and rated highly in background conversations with senior industry leaders. The main reasoning behind this is that if you have done it once you can do it again. Diversity on founder teams has been also rated highly, this was captured by having a female founder.

These can be categorized as the following based on the Crunchbase dataset:

- Startup founder experience (not_first_venture):
  Founders who have previously launched startups are more likely to succeed. Previous research shows that entrepreneurs rarely succeed on the first try, but it is a very relevant experience that translates into a higher chance of success. This variable is constructed as a dummy and is observed on all the founders of a venture. The value 1 represents that at least 1 founder has started a startup company before. It is constructed using the dataset.
- Successful sale before (successful_sale_before):
  One outcome that VCs aim at is acquisition. Therefore, they like to invest in founders who have a proven track record in acquisitions. This also makes it easier for founders to raise money and the track record also shows that they have good skills in such activity.
- Successful IPO before (successful_ipo_before):
  The other outcome VCs aim at is an IPO. Founders who have successfully done this before showcase all the necessary skills for success, attracting talent, ability to fundraise, and skillset to lead a company.
- Has female founder (has_female_founder):
  There is at least one female founder on the team. Previous research has shown that there is a positive effect on the outcome for such teams. VCs also aim at more diverse teams due to research-backed evidence of diversity's positive influence on capital raised. (Wise, Yeganegi, & Laplume, 2022)

These factors are important determining factors for the real-life VC investment decision process. The construction of these features is based on the Crunchbase dataset, all of them are constructed as dummy variables. The dataset covers the observed firms well and the lack of data indicates a high likelihood of the founders' lacking such attributes. Therefore, NAs were treated as 0s (lack of such attributes).

Industry categories

Kim et. al (2021) showed that industry convergence is an existing characteristic of the industry M&A deals. Because one of our success metrics is acquisition, features were required to capture this.

Controlling the industry the startup is operating in is also a way to capture the core operating areas of the startup. This descriptive feature makes it possible for the model to capture a startup sector where they operate in.

This research chose to use the category_groups_list variable from the Crunchbase dataset to capture a wider range of descriptive features. Industry variables should always be tested for target leaks when they can interact with geographical locations. This was prevented by testing, which showed the possible target leak to be at an insignificant level.

This research uses category_groups_list as dummy features, where one observation can operate in more than one category. The following categories have been included in the models:

```
'Administrative Services', 'Advertising',
'Agriculture and Farming', 'Apps',
'Artificial Intelligence', 'Biotechnology',
'Clothing and Apparel', 'Commerce and Shopping',
'Community and Lifestyle', 'Consumer Electronics',
'Consumer Goods', 'Content and Publishing',
'Data and Analytics', 'Design',
'Education', 'Energy', 'Events',
'Financial Services', 'Food and Beverage',
'Gaming', 'Government and Military',
'Hardware', 'Health Care', 'Information Technology',
'Internet Services', 'Lending and Investments',
'Manufacturing', 'Media and Entertainment',
'Messaging and Telecommunications', 'Mobile',
'Music and Audio', 'Natural Resources',
'Navigation and Mapping', 'Other',
'Payments', 'Platforms', 'Privacy and Security',
'Professional Services', 'Real Estate',
'Sales and Marketing', 'Science and Engineering',
'Software', 'Sports', 'Sustainability',
'Transportation', 'Travel and Tourism', 'Video',
```

Funding data

The amount of investment a startup can attract improves its chances of success by accessing technology and talent via that capital. Detailed funding data is therefore one of the most important aspects of the models built in this research.

With access to the Crunchbase database funding data by all funding, rounds were extracted for each organization that had such information within the database. This was available for a wide array of funding rounds listed below:

pre-seed, seed, series A to Series J, series unknown

angel, grant, private equity, debt financing, equity crowdfunding, product crowdfunding,

corporate round, convertible note, non-equity assistance, secondary market

The available data for the separate funding rounds consisted of the following points:

- Top Investor on Board:
  This dummy indicator was engineered for pre-seed, seed, Series A, and Series B funding. The reason for this was that obtaining such an investor early on is a good indicator of success. The list of the top 50 investors was obtained through the Crunchbase dataset. The list was based on the total amount invested. One entry was a duplicate of two individual investors' mutual investments that signaled a strong performance by those investors but decreased the top 50 list to a top48 list. (Appendices – Table 25)
- Money raised in funding round:
  The total amount of money raised in the round using USD. By using the CPI package in Python this was adjusted to 2022 USD values.
- Valuation after funding round:
  The valuation of the company is based on the amount of money raised in the funding round. Also adjusted for inflation using CPI.
- The number of investors on board:
  The total number of investors on board. This would signal high investor interest that would likely indicate future success.

Patents - USPTO

The list of patents was acquired from the USPTO dataset and cross-matched with the company names within our Crunchbase database. The variable uses 1 to indicate an existing patent and 0 to indicate no patent filed by the organization. This feature was included due to previous research indicating that patents are a good indicator of success and are important competitive advantages for certain sectors.

Social Media Data – Twitter

Twitter API was used to acquire the present-day number of followers for all companies that had a correct URL in the Series C dataset. After that past number of followers were acquired to part of the companies using the Wayback Machine. Estimates for companies that had no available archive on the Wayback Machine used the mean change between present-day followers and past followers. The past date for the Wayback machine was the average Series B funding age added to the date of establishment. The archive had to be extracted no later than 450 days after the Series B funding of the startup. This was a timeframe based on the dataset to only have data that would be available for a Series C investor. This data is only included in the Series C dataset to validate whether further analysis would provide more insights for Venture Capital investors.

This limited approach was chosen since academic access to Twitter in Europe is not available for graduate students due to the lack of school to verify the student's status. Future research could use more data and better estimates to explore the impact of social media analysis on startup valuation and selection.

This research also used estimations for the observations where the past follower numbers were unavailable using a mean change method based on the existing data.

Evaluation of the models

Using the confusion matrix to evaluate the performance of the models is fueled by a practical approach. If a Venture Capital firm would use such a model for investments, it could use an investment method to invest in all startups that are labeled as an IPO.

$$\begin{matrix} True\ Negative & False\ Positive \\ False\ Negative & True\ Positive \end{matrix}$$

Example for Confusion Matrix

In such a case the evaluation of the performance can be done using the following formula:

$$\frac{TP}{TP+FP} = Precision\ score$$

This real-life accuracy can be interpreted as the percentage of real IPOs out of all the predicted IPOs.

Additionally, we can use recall to see how the model performed in finding all the possible positive outcomes:

$$\frac{TP}{TP+FN} = Recall$$

Performance on negative outcomes, closed startups, will be evaluated by using the True Negative Rate:

$$\frac{TN}{N} = True\ Negative\ Rate$$

This shows us the performance of the model in avoiding failed startups.

The overall performance of the models will be evaluated using the F1 score given by the following formula:

$$\frac{2 * Precision * Recall}{Precision + Recall} = F1\ Score$$

This measure incorporates Recall and Precision and it is an adequate choice to compare overall performance depending on the dependent variable.

Additionally, ROC curves will be used on a limited number of models.

The evaluation of the Voting Regressor is going to rely on the adjusted R^2 to determine the performance of the models. This is an adjusted version of R^2 using the following formula:

$$R_{adj}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

While R^2 is calculated by the following formula where RSS equals the sum of squared of residuals and TSS equals the total sum of squares:

$$R^2 = 1 - \frac{RSS}{TSS}$$

The baseline performance of Venture Capital firms

According to our research of literature, we are going to use the following metrics as a baseline to evaluate our outcome-prediction models.

- Roughly 20% of the portfolio of successful VC firms reach an IPO. This is only 5% of all VC firms according to Ross et. al. (2021). This is not narrowed down by the stage these investors invest in.

  This is a very clear baseline metric that requires us to evaluate the IPO precision capabilities of the models. If a model can find above 20% of IPOs that would yield an above-market return.
- Hong et. al. (2020) defined a successful exit as a combination of IPO and M&As. They found the following:
  - 17% successful exit rate for below 90th percentile VCs
  - 24% successful exit rate for the top 10 percentile VCs

  To achieve this our models need to have a clear path to put together a portfolio that is predicted to achieve these exits in more than 24% of the cases.

  This research analyzed VC performance between 1990 and 2010. This research due to the time limitations can only be done accurately after a certain time passed. Therefore both metrics may be correct due to the difference in time and the difference in percentiles.

Evaluation for our value prediction model is more difficult because we have no existing baseline. Due to the lack of previous research found on such models, this research will evaluate the data based on the validity of its future use with additional data. VC has been heavily involved in the dot-com bubble and the current tech bubble also hit the VC market and startups heavily already. These two events can be argued as indicators of a need to be able to predict exit value for better valuations during the funding rounds. The lack of these tools in research indicates the difficulty to build one, so the evaluation will be made in light of these facts.

Analysis

Value prediction with Voting Regressor

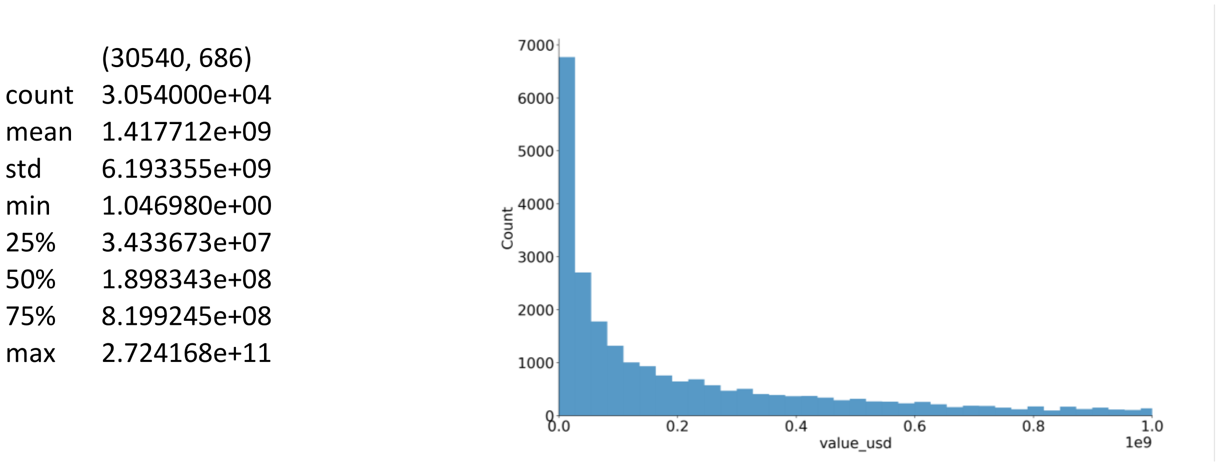
Using Machine Learning for value prediction has not yet been explored in depth in previous research. The aim was to test whether publicly available data can be used for accurate value prediction. Using Voting Regressor the model achieved an adjusted R^2 of 0.5205. This means that publicly available data used in this research was able to explain the value those companies had at their exit point (acquisition or IPO) in 52.05% of the observations. This research used the dataset Valuation, Series C, details of it can be found in the Data Engineering chapter. You can see the results in the table below:

Regression	Adjusted R2-score
Voting Regressor	0.5205
Gradient Boosting Regressor	0.4496
Extra Trees Regressor	0.4607
LGBM Regressor	0.4382

Table 2 – Regression results

Model engineering

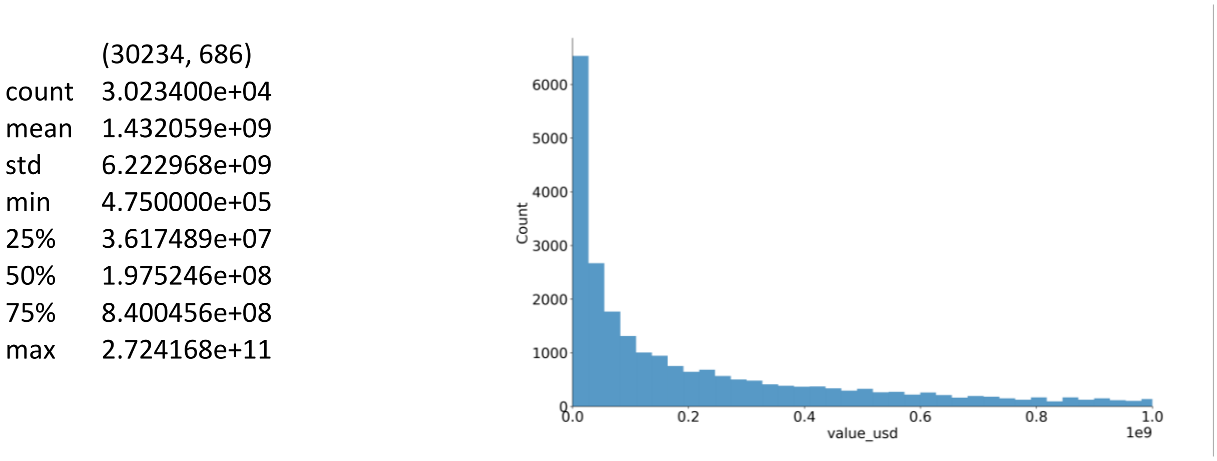
Voting Regressor was used to implement multiple regressions for improved results. At the start, 7 different regressions were included and through testing 4 of them were discarded due to low scores on the dataset. Through this process, our Voting Regression improved significantly in performance. The description of the dataset can be seen below:



```
        (30540, 686)
count   3.054000e+04
mean    1.417712e+09
std     6.193355e+09
min     1.046980e+00
25%     3.433673e+07
50%     1.898343e+08
75%     8.199245e+08
max     2.724168e+11
```

Figure 7 - Dataset description with histogram

Our dataset has a mean of 1.42 billion USD, which is skewed by the outliers at the top of the range. Most observations (75%) are below 819 million USD. This shows us that this requires treatment for outliers. The dataset has been tested with different approaches to outlier removals to improve model performance. Most of these decreased the performance of the models and the best performing approach was to remove the bottom 1% of the dataset, values below 471 949 USD. This yielded the following dataset structure:



```
        (30234, 686)
count   3.023400e+04
mean    1.432059e+09
std     6.222968e+09
min     4.750000e+05
25%     3.617489e+07
50%     1.975246e+08
75%     8.400456e+08
max     2.724168e+11
```

Figure 8 - Dataset description with histogram after removing outliers

As you can see the distribution of the dataset did not change by a large margin, but we can see significant performance improvement after removing low-end outliers, the minimum in the first description at 1.4 USD can be considered a dataset mistake rather than actual data.

After eliminating the regressions that are unable to explain the dataset and removing outliers that interfered with performance, we increased iterations and implemented the following weights based on the relative performance of the models:

Weights:

Regression	Weight
Gradient Boosting Regressor	0.85
Extra Trees Regressor	1
LGBM Regressor	0.7

Table 3 – Weights used in Voting Regression

Weights were set based on the performance of the individual models and manual testing to achieve the highest adjusted R^2. The achieved 0.52 is a strong indicator of the feasibility of such a model for practical applications.

Analysis and implications

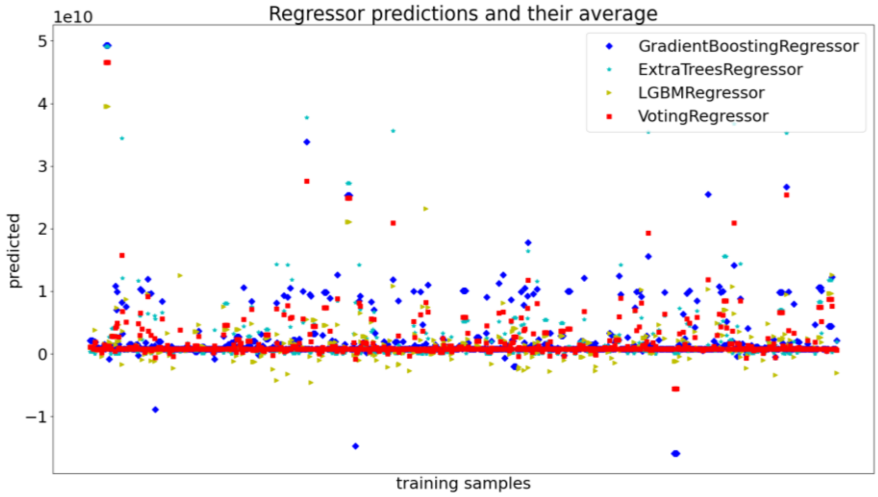


Figure 9 – Plotting the first 1000 predictions of the regressions.

The strong performance of the model indicates that combining publicly available data with other datasets can enable using machine learning models for the valuation of companies. The included features are mostly qualitative, described above in the data engineering part. The financial data included in the dataset only contained the amount raised and valuation after the funding round if applicable.

We can also observe in Figure 9 the way Voting Regressor works in practice by controlling for outlier predictions by the separate models used within. This is the reason for the increased adjusted R^2 from the included models to the Voting Regressor estimator. This indicates that in turbulent environments this methodology yields good results and is suitable for future use.

As we can see from the results the regression achieved an adjusted R2 of 0.5205, which means the model can explain 52.05% of the observations. This is a high performance that is surprising considering that the author of this paper found no previous research that predicted future company exit values based on machine learning and has no knowledge about real-life applications of such. We can therefore conclude that Venture Capital firms would benefit by exploring the further implementation of machine learning into their valuation methodology. Further data for training is necessary to achieve high enough results that warrant using it for investment decision purposes, but other data sources can be explored besides the ones used in this research.

Outcome prediction using XGBoost, MLP, RandomForest, and Naive Bayes.

The XGBoost, MLP, and Random Forest models performed well, achieving high precision scores that suggest a strong investment performance. The precision of the best models from Series A onwards reached ~30% precision and the F1-score ranged from 0.3-04. Naive Bayes underperformed on IPOs but scored high for acquisitions. Our pre-seed models had low-performance indicators besides the MLP Success model which used a different approach-.

Overview of all models

Our models performed well overall, achieving high precision in IPO prediction at 30%-40%. In the following table you can see an overview for predicting IPO across models:

	IPO			
Model Name	Precision	Recall	F1-score	Dataset
RF1_Multilabel	0.3297	0.6765	0.4363	Series C
MLP1_Multilabel	0.4193	0.4504	0.4343	Series C
XGB1_Multilabel	0.3094	0.6580	0.4209	Series C
RF2_Multilabel	0.3246	0.5966	0.4204	Series B
RF1_Binary	0.2965	0.6765	0.4123	Series C
XGB2_Multilabel	0.2924	0.6449	0.4024	Series B
MLP1_Binary	0.3392	0.4845	0.3990	Series C
XGB1_Binary	0.2788	0.6833	0.3969	Series C
MLP2_Binary	0.3311	0.4795	0.3918	Series B
MLP2_Multilabel	0.3127	0.4902	0.3819	Series B
RF2_Binary	0.2654	0.6112	0.3701	Series B
XGB2_Binary	0.2251	0.6839	0.3387	Series B
MLP3_Multilabel	0.2847	0.3881	0.3284	Series A
RF3_Multilabel	0.2322	0.4832	0.3136	Series A
XGB3_Multilabel	0.1884	0.5634	0.2824	Series A
MLP3_Binary	0.2194	0.3956	0.2823	Series A
RF3_Binary	0.1844	0.5352	0.2743	Series A
XGB3_Binary	0.1447	0.6399	0.2360	Series A
XGB4_Multilabel	0.0602	0.5294	0.1082	Pre-Seed
Naive Bayes_Multilabel	0.0176	0.2364	0.0328	Series C
Naive Bayes_Binary	0.0027	0.2857	0.0054	Series C
XGB4_Binary	0.0013	0.3333	0.0027	Pre-seed

Table 4 – Model performance for IPO

Considering the higher precision rates for post-Series A and a sharp drop for post-Seed it makes practical sense that QuantumLight Capital (who markets themselves as the world's first

AI-based VC firm) (QuantumLightCapital, 2022.) is targeting Series B and Series C investment, while promising returns above top percentile VCs (Harley-McKeown, 2022).

Our worst performing models have been the ones using data available at establishment and the Naive Bayes models. Due to this reason, we can conclude that Naive Bayes is not sufficient for capturing IPOs, which is a weakness that makes it not suitable for practical use. For acquisition, it was able to achieve high scores (Precision: 0.9414; Recall:0.6363; F1-Score:0.7593) but the model was discarded for future use due to this.
The table above shows the Precision rates, Recalls and F1-Scores that can be calculated from the Confusion Matrixes (see Appendix). This gives a potential investment performance measurement for the models given an investment strategy that invests in all companies predicted to be one or both of the defined success outcomes.
Such an investment strategy that would invest in all companies predicted to be IPOs, reached the top of the market returns at around 20% according to Ross et. al. (2021).

Our models from seed to series B data perform on precision between 13.46% up to 42%. for IPOs and 58.82% to 93.02% for acquisitions. Naive Bayes is excluded from the previous results due to non-existent performance for IPOs. A general trend for all the models is an improvement in performance for using multi-label classification. There is also a residual gain for multilabel classification models using an investment strategy that invests in both acquisitions and IPOs, mislabeled points between the two classes are captured and the benefits can be reaped after.
In the analysis, each type of model (XGBoost, Random Forest, Multi-Layer Perceptron, and Naive Bayes) and its performance is going to be analyzed separately. All results are available in the appendix grouped by ML methods.

Arguments could also be made that separating acquisitions and IPOs does not make sense on a practical level, since it is hard to make distinctions about why one would be more desired than the other from the founders' point of view. However, the IPO outcome is much more likely to contain success than the acquisition and VCs also gain value by generating publicity, more likely to achieve in the case of an IPO. The earlier the investment stage is the more sense it makes to combine acquisitions and IPOs, this research will test that approach with the best performing model used, Multi-Layer Perceptron.

Naive Bayes

	IPO			Acq			IPO_Acq			
Model	P	R	F1	P	R	F1	P1	R	F1	Data
Multilabel	0.0176	0.2364	0.0328	0.9414	0.6363	0.7593	0.7755	0.6319	0.6964	S-C
Binary	0.0027	0.2857	0.0054	0.8820	0.6564	0.7526				S-C

Table 5 – Naive Bayes for Series C dataset

The Naive Bayes was unable to identify IPOs unlike the other models used. This makes it impractical to use in real life as IPOs are the undisputed success outcome, especially at Series B. One possible and likely reason for this is that it assumes that all features are independent of each other. This is the possible reason for its underperformance that could be further tested. One indicator that this is not the case is the high performance on the Acquisition success measure. This points in the direction that the model is not able to understand the IPOs on the current dataset and further model engineering would not improve that. The good performance of other models meant that this was not explored further due to scope. Future research can explore further this method because it had the highest precision in Acquisition prediction.

XGBoost

	IPO			
Model Name	Precision	Recall	F1-score	Dataset
XGB1_Multilabel	0.3094	0.658	0.4209	Series C
XGB1_Binary	0.2788	0.6833	0.3969	Series C
XGB2_Multilabel	0.2924	0.6449	0.4024	Series B
XGB2_Binary	0.2251	0.6839	0.3387	Series B
XGB3_Multilabel	0.1884	0.5634	0.2824	Series A
XGB3_Binary	0.1447	0.6399	0.2360	Series A
XGB4_Multilabel	0.0602	0.5294	0.1082	Pre-Seed
XGB4_Binary	0.0013	0.3333	0.0027	Pre-Seed

Table 6 – XGBoost performance for predicting IPOs across all funding rounds

XGBoost performed well based on data up to seed, Series A and Series B. The investment performance can be best measured with Precision. The models would invest correctly in IPOs ~14-30% of the time. XGB2 and XGB3 are models that perform at or above the market return rate in both the binary and multilabel approaches.

It can also be seen more clearly in Table 6 that multilabel classification yields better models than primary classifications and the clear linear trend in metrics as we approach a later stage.

	Acquisition			
Model Name	Precision	Recall	F1-score	Dataset
XGB1_Multilabel	0.9098	0.7434	0.8182	Series C
XGB1_Binary	0.7583	0.8796	0.8145	Series C
XGB2_Multilabel	0.8442	0.7173	0.7756	Series B
XGB2_Binary	0.7987	0.7369	0.7666	Series B
XGB3_Multilabel	0.7433	0.6011	0.6647	Series A
XGB3_Binary	0.6246	0.649	0.6365	Series A
XGB4_Multilabel	0.0265	0.5673	0.0507	Pre-Seed
XGB4_Binary	0.0216	0.6102	0.0417	Pre-Seed

Table 7 – XGBoost performance for predicting acquisitions across all funding rounds

A clear linear trend can also be seen in terms of acquisition if we look at the performance across models. These models achieve much higher metrics on acquisitions. For this reason, a combined investment strategy can be considered to be even more beneficial.

	IPO_Acq combined			
Model Name	Precision	Recall	F1-score	Dataset

XGB1_Multilabel	0.8008	0.7346	0.7663	Series C
XGB2_Multilabel	0.7379	0.7099	0.7236	Series B
XGB3_Multilabel	0.6486	0.5938	0.6200	Series A
XGB4_Multilabel	0.0327	0.5563	0.0617	Pre-Seed

Table 8 – XGBoost performance for predicting both successful outcomes across all funding rounds

We also extracted our metrics for two minority classes, acquisitions, and IPOs in the case of the multilabel approach as we can see in Table 8. The models show a high F1-score when combining both outcomes. Investment strategies that use multilabel classification could experience a residual gain effect when investments made with the expectation of being an IPO or acquisition turn out to be the other one-. The following table details the residual gains by the invested outcome and shows the ratio of IPOs and Acquisitions in the test set:

	% of residual gains		Dataset ratio
Model	IPO to Acquisition	Acquisition to IPO	IPO to Acquisition
XGB1	60.38%	2.78%	18:82
XGB2	54.47%	2.38%	16:84
XGB3	55.32%	1.65%	18:82

Table 9 – Residual gains for XGBoost (XGB4 excluded)

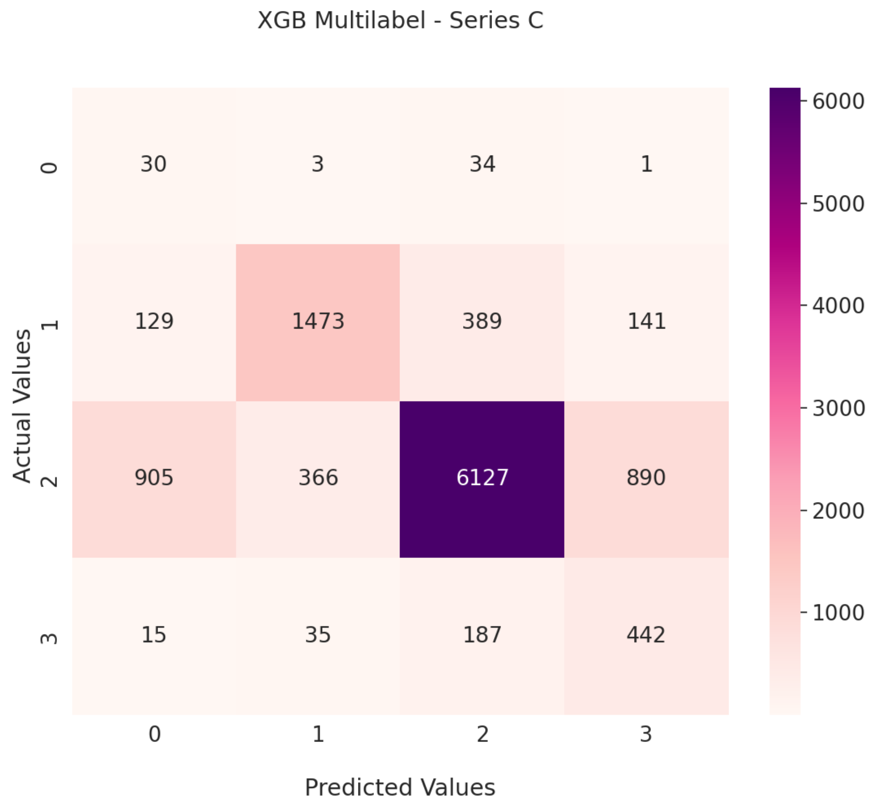


Figure 10 - XGB1 Multilabel classification confusion matrix

As we can see the residual gain from acquisitions to IPOs is minimal, the best investment strategy would be the most simple, investing in IPOs and reaping the residual acquisitions as benefits. This model achieved 30% IPOs and 60.4% acquisitions on top of that, this would lead to a 90% exit rate on its portfolio.

Based on this the most beneficial would be to invest in IPOs with the XGB1 model at Series C rounds. That would yield 60% acquisitions next to the 30% predicted IPOs. Such a portfolio would be among the best performing on the market.

Closed prediction can be evaluated by using the True Negative Rate (TNR):

	TNR	Dataset
XGB1	44.12%	Series C
XGB2	33.72%	Series B
XGB3	42.19%	Series A
XGB4	37.5%	Pre-Seed

Table 10 – TNR for XGBoost

As we can see finding more of the Closed negatives could benefit the model to some extent but looking at the confusion matrixes (see Appendix) shows that the real-life improvements would be insignificant since leaks from Operating are more significant. This means that analyzing TNR does not bear huge significance for model performance, therefore it is not going to be analyzed for the next models.

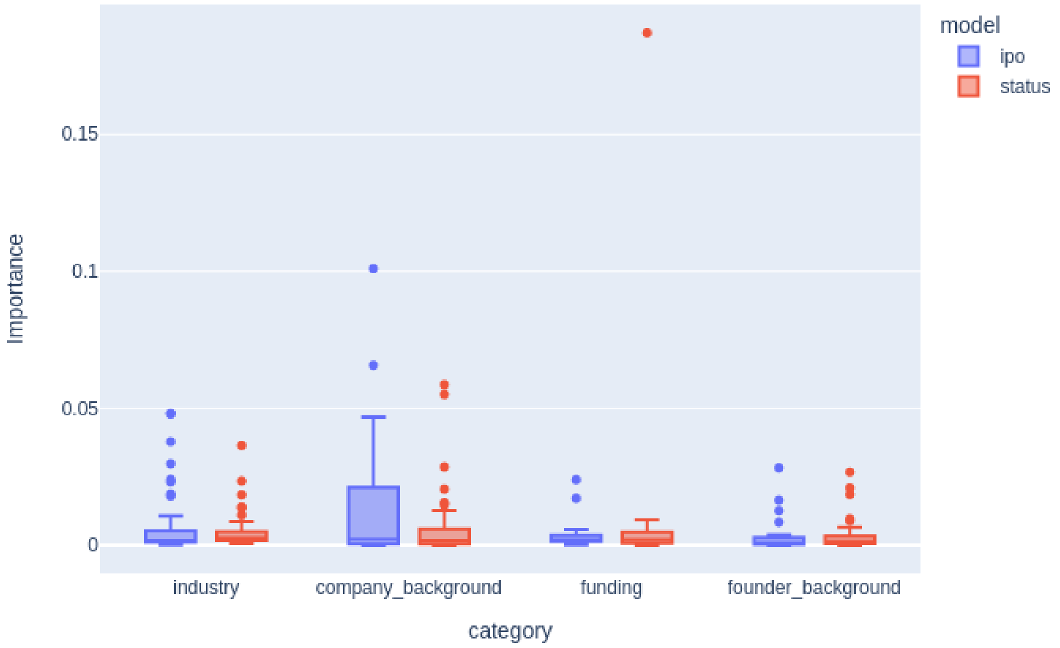
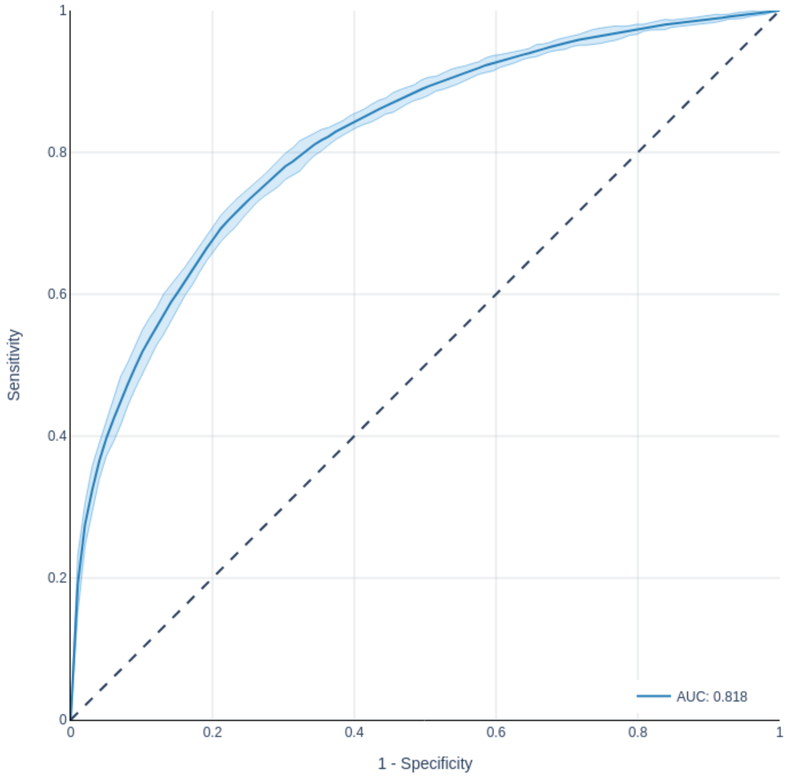


Figure 11 – XGB1 Multilabel and Binary IPO model feature importances by group

The detailed feature importance output can be found in the appendix for XGBoost. From Figure 11 we can see that most of the features lie beneath the boxes for importance. One of the most important features of our multilabel classification approach was the number of investors in Series B. This is probably important due to the recency of the data and the traction it captures for a startup. Investors help startups by introductions within their network, therefore startups do benefit from more than just the investment.

We also extracted the ROC train and test curve for our binary XGB1 model



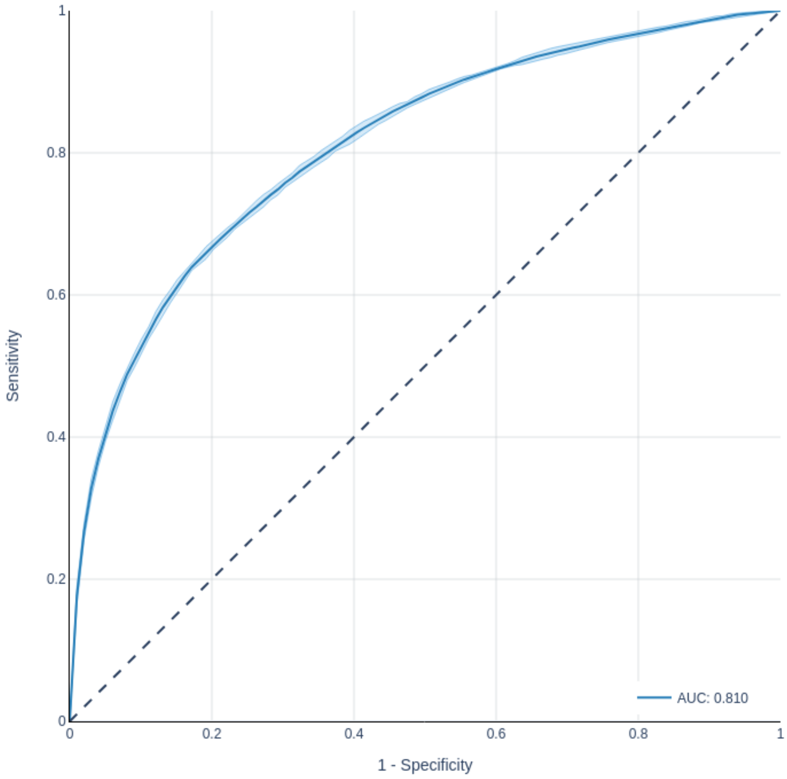


Figure 12 – ROC curves for XGB1

The AUC measure for the XGB1 model was 0.818 for the training set and 0.810 for the test set, this metric also indicates that the model is working well.

Random Forest

	IPO			
Model Name	Precision	Recall	F1-score	Dataset
RF1_Multilabel	0.3297	0.6765	0.4363	Series C
RF2_Multilabel	0.3246	0.5966	0.4204	Series B
RF1_Binary	0.2965	0.6765	0.4123	Series C
RF2_Binary	0.2654	0.6112	0.3701	Series B
RF3_Multilabel	0.2322	0.4832	0.3136	Series A
RF3_Binary	0.1844	0.5352	0.2743	Series A

Table 11 – Random Forest performance for predicting IPOs across all funding rounds

The Random Forest models on the Series B dataset showed that the multilabel approach, in this case, outperformed the Series C Binary model. For all datasets, the multilabel approach yields a 3-5% Precision gain and 2-5% F1-score. The precision scores suggest a strong investment performance on the test sets for all models. The RF1 Multilabel is the best model based on our F1-score metric across all methods tested. Rand Forest models have better recall than MLP models but lower precision. This resulted in similar F1-scores.

	Acquisition			
Model Name	Precision	Recall	F1-score	Dataset
RF1_Binary	0.8627	0.7900	0.8247	Series C
RF1_Multilabel	0.8964	0.7624	0.8240	Series C
RF2_Multilabel	0.8240	0.7491	0.7848	Series B
RF2_Binary	0.7810	0.7777	0.7793	Series B
RF3_Multilabel	0.6810	0.6318	0.6555	Series A
RF3_Binary	0.5881	0.6731	0.6277	Series A

Table 12 – Random Forest performance for predicting acquisitions across all funding rounds

Random Forest was the best performer by F1-score to predict acquisitions. The trend in IPOs disappear and the best performing model is the Series C Binary model. We can also observe a clear linear trend over life stages, which logically makes sense by having more information available to the later models.

	IPO+Acq combined			
Model Name	Precision	Recall	F1-score	Dataset
RF1_Multilabel	0.7947	0.7522	0.7728	Series C
RF2_Multilabel	0.7321	0.7338	0.7329	Series B
RF3_Multilabel	0.6000	0.6185	0.6091	Series A

Table 13 – Random Forest performance for predicting both success metrics across all funding rounds

The combined models also achieved high metrics therefore the combined investment strategy can be beneficial. Portfolio metrics would stay above 60% for successful exit by predicting for both minority classes. We can observe the residual gains for each metric separately in the following table:

	% of residual gains		Dataset ratio
Model	IPO to Acquisition	Acquisition to IPO	IPO to Acquisition
RF1M	55.29%	3.25%	18:82
RF1M	48.56%	3.07%	16:84
RF1M	44.48%	2.46%	18:82

Table 14 – Residual gains for Random Forest

Combining the residual gains with an all IPO portfolio would yield exit rates from 67.7% up to 88.26%, investing from Series A to Series C deals. We can also see the IPO to Acquisition ratio within the dataset as an indicator of actual real-life occurrences.

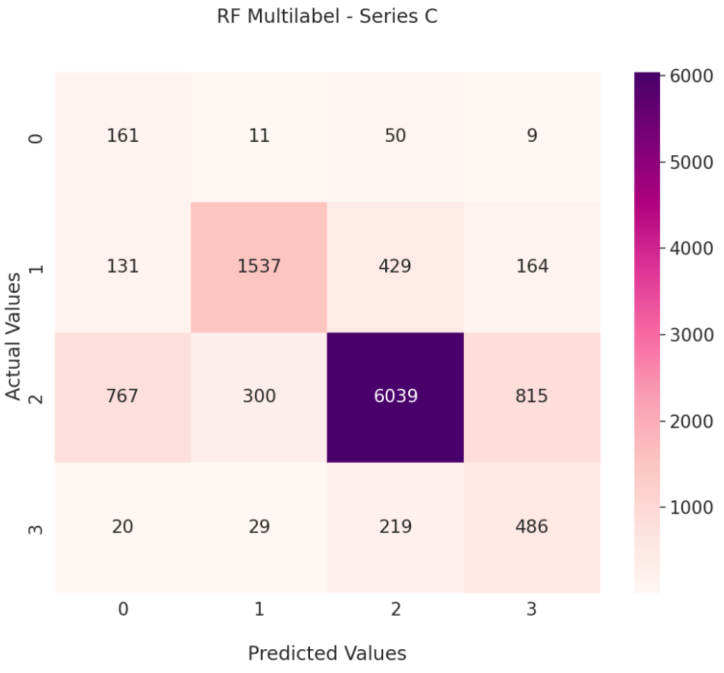


Figure 13 – RF Multilabel Confusion Matrix for Series C dataset

Multilayer Perceptron

	IPO			
Model Name	Precision	Recall	F1-score	Dataset
MLP1_Multilabel	0.4193	0.4504	0.4343	Series C
MLP1_Binary	0.3392	0.4845	0.3990	Series C
MLP2_Binary	0.3311	0.4795	0.3918	Series B
MLP2_Multilabel	0.3127	0.4902	0.3819	Series B
MLP3_Multilabel	0.2847	0.3881	0.3284	Series A
MLP3_Binary	0.2194	0.3956	0.2823	Series A

Table 15 – Multilayer Perceptron performance for predicting IPOs across all funding rounds

Multilayer perceptron has the best performance out of all the models tested especially. The MLP1 multilabel classification model achieved an outstanding 0.4193 precision, which would mean a portfolio where 4 out of 10 companies have an IPO. The multilabel classification approach performs better by a large margin with data up to Series B and Seed, but it has seen no gain on the best post-Series A model. The MLP2 Binary was also built using only 7 layers of 32 neurons, while the MLP2 Multilbale was with 10 layers of 32 neurons. One explanation for this would be that the dataset for Series A and Series B have more data points that benefit the multilabel classification approach.

Overall MLP still has a clear linear trend across the funding stages and multilabel classification performs better overall. The almost 7% gain for Series A investments and the 8% gain for Series C investments over the binary classification models is outstanding compared to the other ML methods explored.

	Acquisition			
Model Name	Precision	Recall	F1-score	Dataset
MLP1_Multilabel	0.8998	0.6271	0.7391	Series C
MLP1_Binary	0.6136	0.8023	0.6954	Series C
MLP2_Multilabel	0.6255	0.7796	0.6941	Series B
MLP2_Binary	0.5896	0.7801	0.6716	Series B
MLP3_Multilabel	0.6250	0.6347	0.6298	Series A
MLP3_Binary	0.5639	0.6391	0.5991	Series A

Table 16 - Multilayer Perceptron performance for predicting acquisitions across all funding rounds

MLP kept the trend over funding rounds for acquisitions but shifted to the same trend we observed for XGBoost. Multilabel outperforms Binary at every stage using the F1-score as a measure.

	IPO_Acq combined			
Model Name	Precision	Recall	F1-score	Dataset
MLP_Success	0.8504	0.6700	0.7495	Pre-Seed, Success
MLP1_Multilabel	0.8135	0.6051	0.6940	Series C
MLP2_Multilabel	0.5679	0.7355	0.641	Series B
MLP3_Multilabel	0.5635	0.5999	0.5812	Series A

Table 17 – MLP performance for predicting both success metrics across all funding rounds

The combined two minority class metrics are similar to XGBoost metrics on precision but lower on F1-score. Despite that, it still shows that a combination can be beneficial as a strategy.

Here we can also see the results for the MLP_Success binary classification model. This data combined IPOs and acquisitions into one outcome dummy indicator variable. The results show that this model can compete on metrics with our later-stage models despite using only early indicators. It has one limitation, it only includes closed startups as the other class. When including operating the metrics drop to a similar level as XGB4, this is an indication that real-life use-case might be limited.

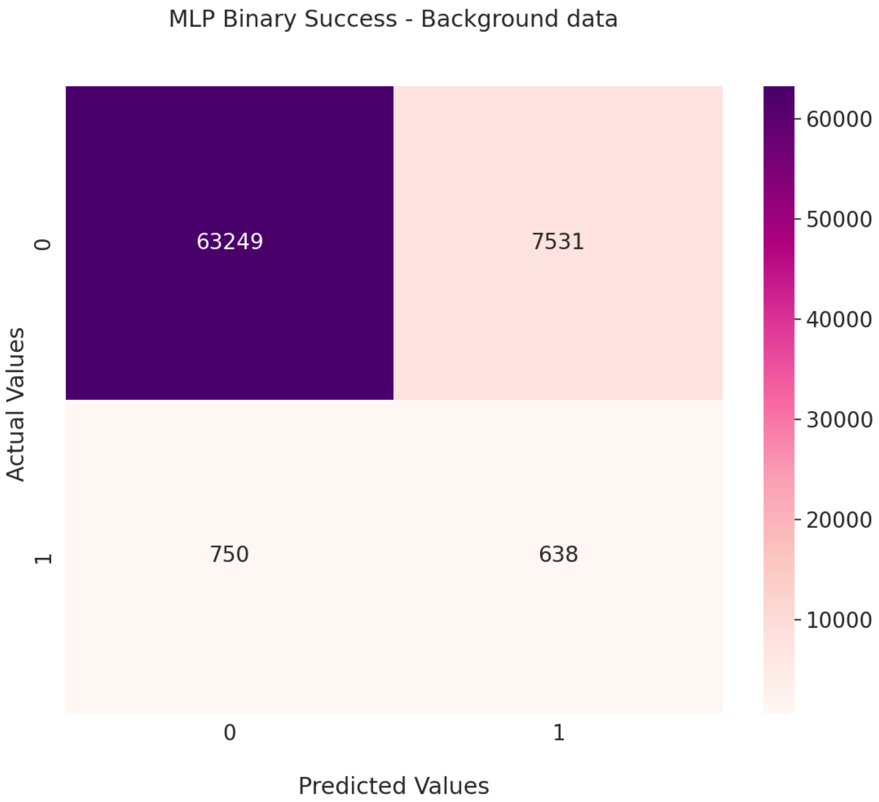


Figure 14 – Pre-Seed MLP Model combining M&As and IPOs into one success metric

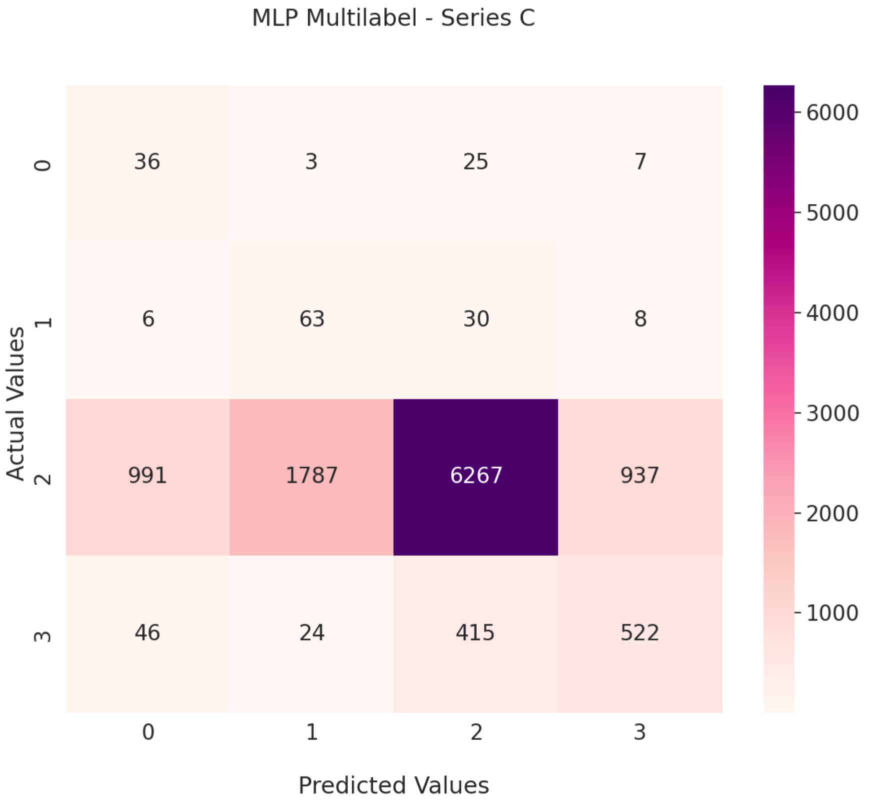


Figure 15 - MLP1 Multilabel Classification confusion matrix

When analyzing more deeply the MLP confusion matrixes we can see that the investment strategy here would have performed well due to the high precision of the IPO prediction.

Adding in our residual gains to the mix, a portfolio investing only in IPOs achieved a staggering 98.4% exits if we add up both acquisitions and IPOs. In the following table, we can see this metric for the other stages:

	% of residual gains		Dataset ratio
Model	IPO to Acquisition	Acquisition to IPO	IPO to Acquisition
MLP1	56.65%	9.51%	18:82
MLP2	37.84%	5.10%	16:84
MLP3	36.27%	5.22%	18:82

Table 18 – Residual gains for MLP

MLP2 and MLP3 have a decreased upside compared to MLP1 and also compared to XGBoost results. Despite this fact, the two possible portfolios would have had a 69.11% and 65.01% exit ratio respectively.

Analysis and implications

This research built well-functioning models using 3 different types of machine learning methods. All models had metrics that made them suitable for different use cases. The Naive Bayes model failed to capture the most relevant outcome, IPO, therefore they are discarded as a relevant tool in this research. Comparison with available metrics to compare the method to non-ML VC investments showed an above-market performance.

Our chosen method for investment is the simplest one, invest in all predicted IPOs in Series A, B, and C funding rounds. Using the multilabel classification method we can evaluate on the test the residual Acquisition gain from this investment strategy to see our portfolio metrics. The following table shows the portfolio performances. To compare we have the following two baseline metrics:

- 20% IPO ratio in the portfolio of the top 5% of VCs (Ross, Das, Sciro, & Raza, 2021.) →
- 24% exit ratio (M&A + IPO) for top10% of VC (Hong, Serfes, & Thiele, 2020)
- 17% exit ratio for below 90th percentile VCs (Hong, Serfes, & Thiele, 2020)

Model Name	IPO	Residual gain	Total	Stage
MLP1_Multilabel	0.4193	0.5665	0.9858	Series C
XGB1_Multilabel	0.3094	0.6038	0.9132	Series C
RF1_Multilabel	0.3297	0.5529	0.8826	Series C
XGB2_Multilabel	0.2924	0.5447	0.8371	Series B
RF2_Multilabel	0.3246	0.4856	0.8102	Series B
XGB3_Multilabel	0.1884	0.5532	0.7416	Series A
MLP2_Multilabel	0.3127	0.3784	0.6911	Series B
RF3_Multilabel	0.2322	0.4448	0.6770	Series A
MLP3_Multilabel	0.2847	0.3627	0.6474	Series A

Table 19 – Portfolio performance with the residual gain of acquisitions

We can see that besides XGB3 all of the multilabel classification approaches outperform all 3 of the baseline metrics found in previous literature. This research shows that a more data-driven approach to Venture Capital investment would be beneficial for all VCs in the industry. Limitations to replicating this performance in real-life are the following:

- Source enough potential IPOs and Acquisitions in the deal flow.
  This is an important aspect, but VCs who are unable to do this are going to generate losses regardless of the investment approach chosen.

- Close the signaled IPO deals.

  This factor introduces a limit because the best deals usually have intensive competition. Therefore the complementing capabilities of the VC firms are still a crucial aspect.

One solution for the competition aspect is to invest in an earlier stage. That is hindered by the general underperformance of models that rely only on early indicators. This research still has one model that combines success metrics into one category and uses binary classification with that. The model MLP Success had good metrics but required a modified dataset that removed one observed category. Therefore the model signals that earlier investment prediction models can use this approach for model building but it is also important to note that the model in this research had limitations that signal it is not yet feasible for real-life use.

Analyzing social media

This research analyzed the effect of social media analysis on both approaches, value prediction, and outcome prediction. Findings showed that even with the limited data that was extracted from Twitter and the Wayback Machine performance of certain models improved. This shows that there is a possibility to utilize further social media analysis in the VC industry, but a careful selection of the approaches might be necessary. This might be different when using more accurate social media datasets.

Value prediction

Regression	Original	Social	Social (W:0.8;1;0.23)
Voting Regressor	0.5205	0.5235	0.5281
Gradient Boosting Regressor	0.4496	0.4575	0.4575
Extra Trees Regressor	0.4607	0.4912	0.4912
LGBM Regressor	0.4382	0.3763	0.3763

Table 20 – Comparing regressions’ performance with the addition of social media data

As you can see from the results value prediction improved for the Extra Trees Regressor by 0.035 and for the Gradient Boosting Regressor by 0.0079 while decreasing by 0.0619 for the LGBM Regressor. The overall performance slightly increased, being able to explain 0.3% more of the observations. The change in the component regression pointed toward a need to change the weights, compared to the original model. A change in the weights resulted in an increase of 0.76% compared to the original regression. Considering the limitations these results implicate that analyzing social media explains the observations better, but it also significantly changes the right approaches for model engineering.

Comparing the first 1000 predictions we can also observe that the model was predicting smaller negative values less often while predicting higher values more often. This behavior change can explain the differences in the adjusted R2 scores. The following plot shows the first 1000 predictions with the best social media regression, the corresponding Figure 11 can be found in the original regressions analysis.

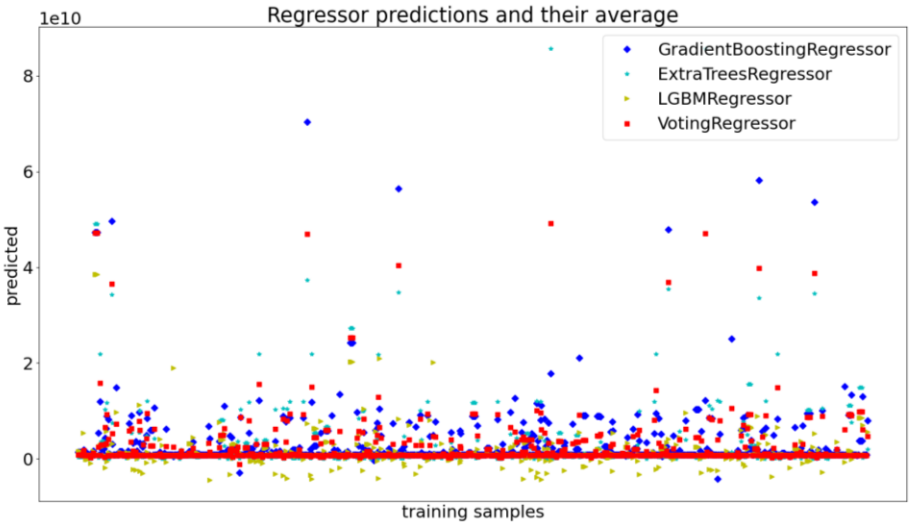


Figure 16 – Plotting the first 1000 predictions of the regressions with social media data.

Outcome prediction

	IPO			Delta	
Model Name	Precision	Recall	F1-score	Precision	F1-Score
RF1_Multilabel	0.3297	0.6765	0.4363	⬆	⬆
RFS_Multilabel	0.3366	0.6561	0.4449		
RF1_Binary	0.2965	0.6765	0.4123	⬆	⬆
RFS_Binary	0.3001	0.6710	0.4148		
MLP1_Multilabel	0.4193	0.4504	0.4343	⬇	⬇
MLPS_Multilabel	0.3301	0.5158	0.4025		
MLP1_Binary	0.3392	0.4845	0.3990	⬇	⬆
MLPS_Binary	0.3158	0.5900	0.4114		

Table 21 – Random Forest performance for predicting IPOs across all funding rounds

The results show that slight improvements were observable with the addition of social media data across both the Binary and the Multilabel approach for the Random Forest models for Series C investors. The MLP approach only saw improvements for the F1 score for the Binary approach. This observation is in line with the results for the value prediction, the most improvements were achieved with a decision tree approach. This leads to a similar conclusion, the addition of social media to models predicting startup outcomes is a useful addition to the models, but its value of it depends on the approach chosen. The limitations of the social media data prevent us from reaching a clear conclusion on the approach but the results point in the direction of decision trees as the best method.

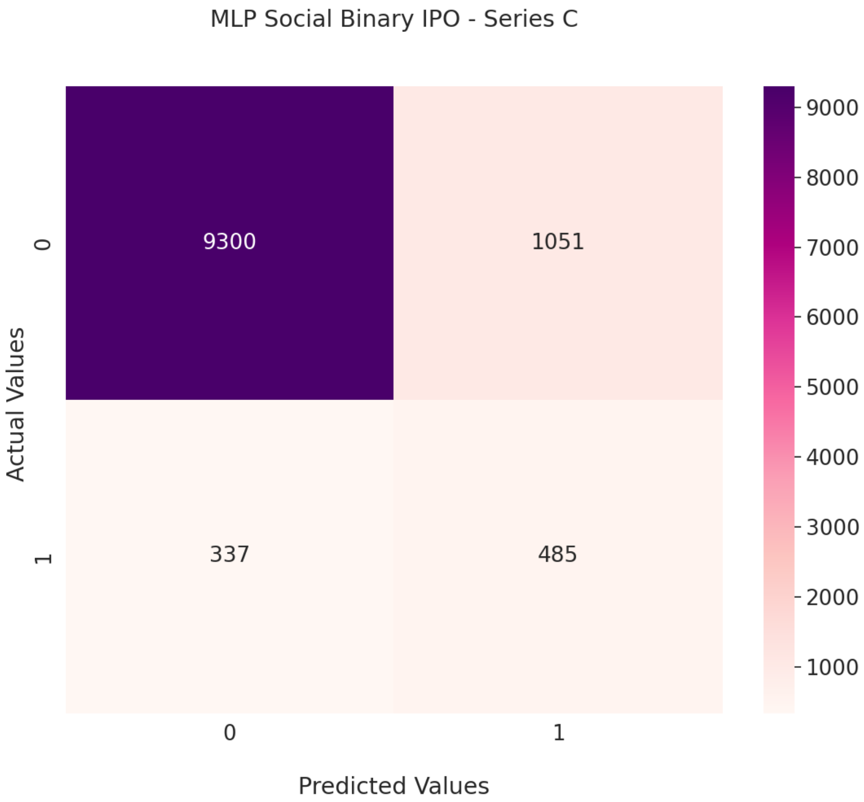


Figure 17 – MLP Social Binary – Series C dataset

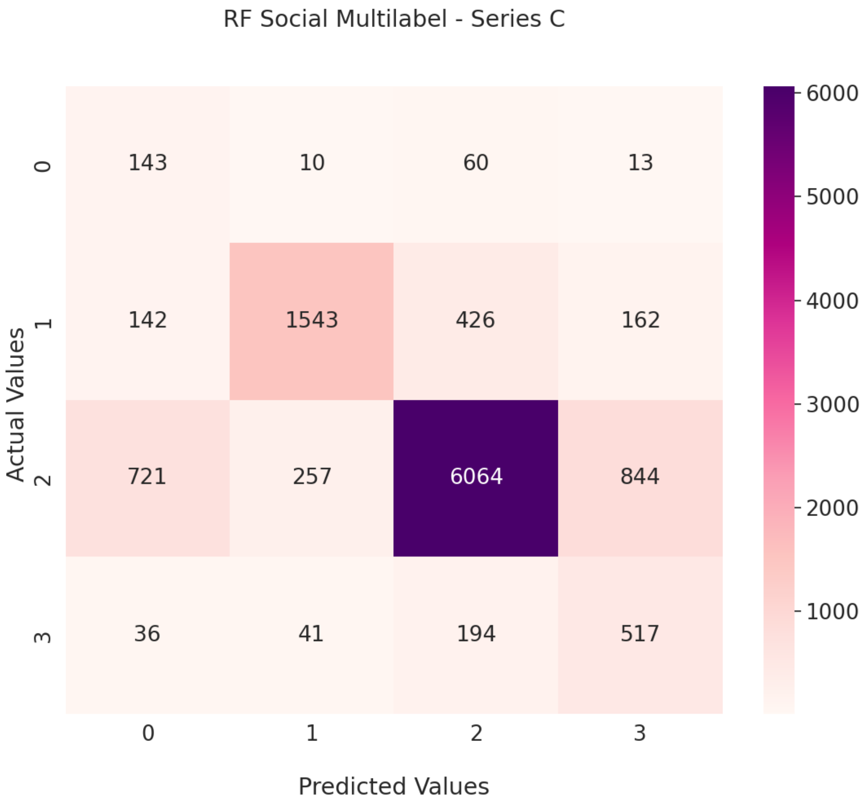


Figure 18 – RF Social Multilabel – Series C dataset

Conclusion

This research shows that utilizing machine learning for predicting startup exit values can be possible by using additional data. The voting regressor had an adjusted R^2 of 0.5205. This in itself is not high enough for immediate real-life applications, but the potential of adding further data to it means this has a high potential for practical applications. Potential data that can be used could be financial data of the companies that were not acquirable within the scope of this research and additional and improved social media data.

This research also shows that outcome prediction could be utilized for a better investment portfolio that could outperform the market by a margin. Because the real-life application of such a tool could potentially suffer a decrease in performance the margin it was able to outperform top VC firms validates its potential, even with a loss in performance it would still be able to outperform the market. The potential to use it based on these results is from Series A to Series C funding rounds, improving in accuracy at the later stages. Using the multilabel approach the models were compared to the three baseline metrics found in the literature. It showed that, besides one early model (Series A), all of them performed better on the dataset than the best on the market. The large margin enables the conclusion that this could be replicated in real-life circumstances, and during the duration of the research one AI-investment fund has been launched (Harley-McKeown, 2022) (QuantumLightCapital, 2022.). On Quantum Light Capital's website, they also promise similar improvements in performance for investors with their new approach. This leads to the conclusion that in the future VC funds that utilize machine learning will be able to outperform traditional approaches from Series A onwards and that more and more funds will invest in capabilities to support its investment decision-making.

Lastly, this research analyzed social media data as potential data to use for models. Both the value and the outcome predictions saw improvements, most improvements were observed for decision tree approaches on both predictions. The performance observed for decision trees and MLP models show that the correct approach for future applications is most likely with these two. This leads to the conclusion that social media analysis is useful and could be the main focus of future research.

To conclude this research shows a new approach to decision-making for investment strategy in a field characterized by qualitative rather than quantitative data. This new approach could benefit the field with better valuations that would be beneficial after the dot-com crisis and the current crisis saw overvaluations for startups, and better portfolios would benefit investors who are investing in a volatile and turbulent environment that is characterized by high-risk high-reward.

Limitations and future research

Limitations to this research were characterized by available data that can be easily extracted. The research relied on Crunchbase and USPTO for the first part, while social media relied on Twitter. Twitter bottlenecks included the lack of access for European students and the lack of historical follower numbers available. To circumvent this Wayback machine was used, but the amount of data acquired was smaller than the data from Crunchbase. Future research therefore could explore the availability of further data on the social media aspect, either from Twitter or from other sources (such as Instagram, Reddit, and Discord).

Crunchbase also suffers from missing data points, future research could add value by exploring sources to extract more data for education and founder track record. Crunchbase also has information about first employees, future research could explore this further as well, first employees are important for startups' success and they also capture the talent attraction ability of the founders.

The research was also limited by the lack of financial data publicly available for such startups. Although they are not available publicly, VC firms do get access to such by potential investment cases. The bottleneck here is the ability for future research to acquire such in large volume to benefit the models, but this could also be explored further using government datasets as a potential source.

References

Arora, A., Fosfuri, A., & Rønda, T. (2021.). Waiting for the Payday? The Market for Startups and the Timing of Entrepreneurial Exit. *Management Science, 67*(3), 1453-1467. Retrieved from https://doi.org/10.1287/mnsc.2020.3627

Bai, S., & Zhao, Y. (2021). Startup Investment Decision Support: Application of Venture. *Systems*(9), 55. doi:https://doi.org/10.3390/

Chen, T., & Guestrin, C. (2016.). XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 785-794.). New York: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2939672.2939785

Cotei, C., Farhat, J., & Khurana, I. (2021.). The impact of policy uncertainty on the M&A exit of startup firms. *Journal of Economics and Finance*. Retrieved from https://doi.org/10.1007/s12197-021-09553-9

Gloor, P. A., Colladon, A. F., Grippa, F., & Hadley, B. M. (2020.). The impact of social media presence and board member composition on new venture success: Evidences from VC-backed U.S. startups. *Technological Forecasting & Social Change, 157*. Retrieved from https://doi.org/10.1016/j.techfore.2020.120098

Harley-McKeown, L. (2022, May 17). *Revolut founder set to launch venture capital fund powered by artificial intelligence*. Retrieved from THE BLOCK: https://www.theblockcrypto.com/linked/147286/revolut-founder-set-to-launch-venture-capital-fund-powered-by-artificial-intelligence

Ho, T. K. (1995.). *Random Decision Forests.* Retrieved from NC State University: https://www4.stat.ncsu.edu/~lu/ST7901/reading%20materials/Ho1995.pdf

Hong, S., Serfes, K., & Thiele, V. (2020). Competition in the venture capital market and the success of startup companies: Theory and evidence. *Journal of Economics & Management Strategy, 29*(4), 741-791. Retrieved from https://doi.org/10.1111/jems.12394

Hunter, D. S., Saini, A., & Zaman, T. (2017.). Picking Winnders: A Data Driven Approach to Evaluating the Quality of Startup Companies. *arXiv.org, Papers 1706.04229*. Retrieved December 12., 2021., from https://ideas.repec.org/p/arx/papers/1706.04229.html

Jalbert, T., Jalbert, M., & Furumo, K. (2013.). The Relationship Between CEO Gender, FInancial Performance, And Financial Management. *Journal of Business & Economics Research, Vol. 11(1)*, 25-33. Retrieved January 3., 2022., from https://ssrn.com/abstract=2218859

Jiao, P., Veiga, A., & Walther, A. (2020.). Social media, news media and the stock market. *Journal of Economic Behaviour and Organization, 176*, 63-90. Retrieved from https://doi.org/10.1016/j.jebo.2020.03.002

Kaplan, S. N., & Lerner, J. (2010). It Ain't Broke: The Past, Present, and Future of Venture Capital. *Journal of Applied Corporate Finance*, 36-47.

Kim, J., Jeon, W., & Geum, Y. (2021). Industry Convergence for Startup Businesses: Dynamic Trend Analysis Using Merger and Acquisition Information. *IEEE Transactions of Engineering Management*, 1-22. Retrieved from http://dx.doi.org/10.1016/j.jcorpfin.2014.10.017

Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the Outcome of Startups: Less Failure, More Success. *IEEE 16th International Conference on Data Mining Workshops* (pp. 798-805). IEEE Computer Society. doi:DOI 10.1109/ICDMW.2016.103

Kurshunova, E., Tiberius, V., Cesinger, B., & Bouncken, R. (2021.). Potential pitfalls of startup integrations: An exploratory study. *Journal of Business Venturing Insights, 15*. Retrieved from https://doi.org/10.1016/j.jbvi.2021.e00237

Nicholas, T. (2019). *VC: An American History.* Harvarad University Press.

Ogane, Y. (2015.). Competition among Financial Institutions and Startup Company Exit. *International Journal of Business, 20(2)*, 128-143. Retrieved from http://www.jstor.org/stable/40282379

Pedregose, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830. Retrieved from https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pineiro-Chousa, J., Vizcaíno-González, M., & Pérez-Pico, A. M. (2017.). Influence of Social Media over the Stock Market. *Psychology & Marketing, 34*(1), 101-108. Retrieved from https://doi.org/10.1002/mar.20976

Pisoni, A., & Onetti, A. (2018.). When startups exit: comparing strategies in Europe and the USA. *Journal of Business Strategy, 39*(3), 26-33. Retrieved from https://doi.org/10.1108/JBS-02-2017-0022

QuantumLightCapital. (2022., 13. 06.). *QuantumLight*. Retrieved from QuantumLight: https://quantumlightcapital.com/

Roche, M. P., Conti, A., & Rothaermel, F. T. (2020). Different founders, different venture outcomes: A comparative analysis of academic and non-academic startups. *Research Policy, 49*(10). Retrieved from https://doi.org/10.1016/j.respol.2020.104062

Ross, G., Das, S., Sciro, D., & Raza, H. (2021.). CapitalVX: A machine learning model for startup selection and exit prediction. *The Journal of Finance and Data Science 7, 7*, 94-114. Retrieved from https://doi.org/10.1016/j.jfds.2021.04.001

Sage, Verve Search. (2019, May 23). *Unicorn League*. Retrieved January 14, 2022, from Worderist: https://worderist.com/portfolio/unicorn-league/

scikit-learn. (2022). *scikit-learn: machine learning in Python — scikit-learn 1.0.2 documentation*. Retrieved January 12, 2022, from scikit-learn: https://scikit-learn.org/stable/index.html

Silver, D., Hubert, T., Schrittweiser, J., Antonoglou, I., Lai, M., Guez, A., . . . Hassabis, D. (2017.). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *CoRR, abs/1712.01815*. Retrieved January 4., 2022., from https://arxiv.org/pdf/1712.01815.pdf

Slama, R. B., Ajina, A., & Lakhal, F. (2019.). Board gender diversity and firm financial performance in France: Empirical evidence using quantile differene-in-differences and dose-response models. *Cogent Economics & Finance, 7(1)*. Retrieved December 29., 2021., from http://dx.doi.org/10.1080/23322039.2019.1626526

Wennberg, K., & DiTienne, D. R. (2014.). What do we really mean when we talk about 'exit'? A critical review of research on entrepreneurial exit. *International Small Business Journal, 32*(1), 4-16. Retrieved from https://doi.org/10.1177/0266242613517126

Wennberg, K., Wiklund, J., DeTienne, D. R., & Cardon, M. S. (2010.). Reconceptualizing entrepreneurial exit: Divergent exit routes and their drivers. *Journal of Business Venturing, 25*, 361-375. Retrieved from https://doi.org/10.1016/j.jbusvent.2009.01.001

Wise, S., Yeganegi, S., & Laplume, A. O. (2022, June). Startup team ethnic diversity and investment capital raised. *Journal of Business Venturing Insights, 17*. doi:https://doi.org/10.1016/j.jbvi.2022.e00314

Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C., & Liu, C. (2012.). A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch. *Sixth International AAAI Conference on Weblogs and Social Media* (pp. 607-610.). Association for the Advancement of Artificial Intelligence.

Zhang, H. (2004, January). The Optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS, 2*. Retrieved January 8, 2022, from https://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf

Zhong, H., Liu, C., Zhong, J., & Xiong, H. (2018.). Which startup to invest in: a personalized portfolio strategy. *Annals of Operations Research, 263*, 339-360. Retrieved from https://doi.org/10.1007/s10479-016-2316-z

Appendices

Table 22 – List of all features after data engineering besides educational and geographical features (due to length) [Series C dataset]

	mean	std	min	25%	50%	75%	max
status	1.7734	0.7974	0	1	2	2	3
Administrative Services	0.0314	0.1743	0	0	0	0	1
Advertising	0.0536	0.2252	0	0	0	0	1
Agriculture and Farming	0.0099	0.0988	0	0	0	0	1
Apps	0.0481	0.2140	0	0	0	0	1
Artificial Intelligence	0.0428	0.2024	0	0	0	0	1
Biotechnology	0.0876	0.2828	0	0	0	0	1
Clothing and Apparel	0.0180	0.1331	0	0	0	0	1
Commerce and Shopping	0.1206	0.3256	0	0	0	0	1
Community and Lifestyle	0.0415	0.1994	0	0	0	0	1
Consumer Electronics	0.0604	0.2383	0	0	0	0	1
Consumer Goods	0.0301	0.1710	0	0	0	0	1
Content and Publishing	0.0382	0.1916	0	0	0	0	1
Data and Analytics	0.1147	0.3187	0	0	0	0	1
Design	0.0441	0.2054	0	0	0	0	1
Education	0.0345	0.1825	0	0	0	0	1
Energy	0.0357	0.1855	0	0	0	0	1
Events	0.0109	0.1036	0	0	0	0	1
Financial Services	0.1094	0.3122	0	0	0	0	1
Food and Beverage	0.0448	0.2069	0	0	0	0	1
Gaming	0.0229	0.1497	0	0	0	0	1
Government and Military	0.0094	0.0963	0	0	0	0	1
Hardware	0.1534	0.3603	0	0	0	0	1
Health Care	0.1706	0.3762	0	0	0	0	1
Information Technology	0.2010	0.4008	0	0	0	0	1
Internet Services	0.1834	0.3870	0	0	0	0	1
Lending and Investments	0.0274	0.1633	0	0	0	0	1
Manufacturing	0.1006	0.3008	0	0	0	0	1
Media and Entertainment	0.1190	0.3238	0	0	0	0	1
Messaging and Telecommunications	0.0186	0.1351	0	0	0	0	1
Mobile	0.0946	0.2927	0	0	0	0	1
Music and Audio	0.0137	0.1164	0	0	0	0	1
Natural Resources	0.0264	0.1604	0	0	0	0	1

Navigation and Mapping	0.0088	0.0936	0	0	0	0	1
Other	0.1031	0.3041	0	0	0	0	1
Payments	0.0266	0.1608	0	0	0	0	1
Platforms	0.0135	0.1156	0	0	0	0	1
Privacy and Security	0.0430	0.2029	0	0	0	0	1
Professional Services	0.0828	0.2756	0	0	0	0	1
Real Estate	0.0454	0.2081	0	0	0	0	1
Sales and Marketing	0.1044	0.3057	0	0	0	0	1
Science and Engineering	0.1640	0.3703	0	0	0	0	1
Software	0.3774	0.4847	0	0	0	1	1
Sports	0.0223	0.1476	0	0	0	0	1
Sustainability	0.0296	0.1694	0	0	0	0	1
Transportation	0.0627	0.2424	0	0	0	0	1
Travel and Tourism	0.0277	0.1640	0	0	0	0	1
Video	0.0328	0.1780	0	0	0	0	1
has_patent	0.1043	0.3057	0	0	0	0	1
money_raised_series_a	3456347.3466	18018173.9384	0	0	0	1534278	1.56E+09
post_series_a_valuation	2072195.4284	81582051.6326	0	0	0	0	1.11E+10
series_a_investor_count	0.8322	1.9201	0	0	0	1	46
top_investor_on_board_A	0.0264	0.1602	0	0	0	0	1
money_raised_angel	50208.8241	1136435.1860	0	0	0	0	2E+08
post_angel_valuation	12469.5594	483634.8366	0	0	0	0	59701059
angel_investor_count	0.0729	0.6373	0	0	0	0	42
top_investor_on_board_angel	0.0008	0.0274	0	0	0	0	1
money_raised_series_b	7200642.0925	29425677.3663	0	0	0	2122400	2.26E+09
post_series_b_valuation	9535414.8813	179769186.9829	0	0	0	0	2.09E+10
series_b_investor_count	0.9670	2.1697	0	0	0	1	69
top_investor_on_board_B	0.0401	0.1962	0	0	0	0	1
money_raised_seed	351555.1649	1583327.7048	0	0	0	0	1.24E+08
post_seed_valuation	78628.3577	2052552.0953	0	0	0	0	3.15E+08
seed_investor_count	0.5430	1.7992	0	0	0	0	84

top_investor_on_board_seed	0.0068	0.0823	0	0	0	0	1
money_raised_pre_seed	6862.8826	142769.2967	0	0	0	0	22109171
pre_seed_valuation	4806.2727	290973.3860	0	0	0	0	45465982
pre_seed_investor_count	0.0346	0.3632	0	0	0	0	26
top_investor_on_board_preseed	0.0001	0.0095	0	0	0	0	1
money_raised_grant	247846.7479	6446352.7777	0	0	0	0	5.06E+08
money_raised_equity_crowdfunding	9696.9816	537624.4762	0	0	0	0	1E+08
post_equity_crowdfunding_valuation	65065.6863	4400291.8162	0	0	0	0	8.5E+08
money_raised_product_crowdfunding	2501.6571	184496.6593	0	0	0	0	31896712
post_product_crowdfunding_valuation	0.0000	0.0000	0	0	0	0	0
not_first_venture	0.1198	0.3247	0	0	0	0	1
successful_IPO_before	0.0291	0.1680	0	0	0	0	1
successful_sale_before	0.0351	0.1840	0	0	0	0	1
has_female_founder	0.0993	0.2991	0	0	0	0	1

Table 23 – XGB1 model multilabel approach feature importance

	Feature	Importance
1	Administrative Services	0.001726
2	Advertising	0.007409
3	Agriculture and Farming	0.004851
4	Apps	0.001465
5	Artificial Intelligence	0.007157
6	Biotechnology	0.017239
7	Clothing and Apparel	0.001706
8	Commerce and Shopping	0.001729
9	Community and Lifestyle	0.001041
10	Consumer Electronics	0.002035
11	Consumer Goods	0.002509
12	Content and Publishing	0.002896
13	Data and Analytics	0.003311
14	Design	0.002272
15	Education	0.001952

16	Energy	0.003311
17	Events	0.003097
18	Financial Services	0.00261
19	Food and Beverage	0.002156
20	Gaming	0.002905
21	Government and Military	0.000362
22	Hardware	0.001543
23	Health Care	0.004052
24	Information Technology	0.002755
25	Internet Services	0.006443
26	Lending and Investments	0.010623
27	Manufacturing	0.011204
28	Media and Entertainment	0.003971
29	Messaging and Telecommunications	0.000956
30	Mobile	0.005276
31	Music and Audio	0.001199
32	Natural Resources	0.031713
33	Navigation and Mapping	0.000418
34	Other	0.001773
35	Payments	0.001673
36	Platforms	0.00349
37	Privacy and Security	0.001948
38	Professional Services	0.003553
39	Real Estate	0.007826
40	Sales and Marketing	0.003007
41	Science and Engineering	0.012505
42	Software	0.021847
43	Sports	0.001399
44	Sustainability	0.00304
45	Transportation	0.002029
46	Travel and Tourism	0.00105
47	Video	0.000766
48	has_patent	0.009834
49	money_raised_series_a	0.002083
50	post_series_a_valuation	0.000597
51	series_a_investor_count	0.002827
52	top_investor_on_board_A	0.00283
53	money_raised_angel	0.001536
54	angel_investor_count	0.001628

55	money_raised_series_b	0.008643
56	post_series_b_valuation	0.004831
57	series_b_investor_count	0.180673
58	top_investor_on_board_B	0.003554
59	money_raised_seed	0.006095
60	post_seed_valuation	0.000967
61	seed_investor_count	0.00831
62	top_investor_on_board_seed	0.00129
63	money_raised_pre_seed	0.000637
64	pre_seed_valuation	0.001527
65	pre_seed_investor_count	0.001341
66	money_raised_grant	0.004505
67	money_raised_equity_crowdfunding	0.00018
68	money_raised_product_crowdfunding	4.67E-05
69	not_first_venture	0.00379
70	successful_IPO_before	0.019801
71	successful_sale_before	0.00623
72	Andhra University	0.000564
73	Bar-Ilan University	8.78E-07
74	Baylor University	0.000368
75	Ben-Gurion University of the Negev	0.003752
76	Boston University	0.000319
77	Brigham Young University (BYU)	0.000581
78	Brown University	6.59E-05
79	California State University	0.000449
80	Carnegie Mellon University	0.000496
81	Cheung Kong Graduate School of Business	0.005119
82	Columbia Business School	0.001629
83	Columbia University	0.000204
84	Cornell University	0.002049
85	Duke University	0.000266
86	Fudan University - School of Management	0.02745
87	Harvard Business School	0.001282
88	Harvard Medical School	0.00158
89	Harvard University	0.007264
90	Hebrew University of Jerusalem	0.005741
91	IDC Herzliya	0.021445
92	INSEAD	0.002308
93	Indian Institute of Technology Bombay (IIT)	0.004192

94	Indian Institute of Technology Madras	0.009092
95	Kellogg School of Management	0.000994
96	Massachusetts Institute of Technology	0.001412
97	McGill University	0.00383
98	Mines ParisTech	0.000855
99	New York University	0.003555
100	Northeastern University	0.003292
101	Northwestern University	0.000659
102	Open University of Israel	0.000362
103	Princeton University	0.001408
104	Purdue University	0.001578
105	Rensselaer Polytechnic Institute	0.001239
106	Seoul National University	0.000243
107	Shanghai Jiao Tong University	0.005117
108	Stanford Graduate School of Business	0.000695
109	Stanford University	0.002481
110	Stanford University School of Medicine	0.000732
111	Technion	0.000882
112	Tel Aviv University	0.000872
113	The University of Texas at Austin	0.004186
114	Tsinghua University	4.76E-05
115	Tufts University	5.75E-05
116	University of British Columbia	0.000348
117	University of California Berkeley	0.001389
118	University of California Davis	0.002858
119	University of California, Los Angeles	0.002644
120	University of California, San Diego	0.000131
121	University of Cambridge	0.000979
122	University of Chicago	0.001922
123	University of Florida	0.002231
124	University of Illinois at Urbana-Champaign	0.000738
125	University of Melbourne	0.002441
126	University of Michigan	0.000884
127	University of Oxford	0.002851
128	University of Pennsylvania	0.002771
129	University of Southern California	0.000687
130	University of Toronto	0.001053
131	University of Virginia	0.001388
132	University of Washington	0.003791

133	University of Waterloo	0.000127
134	Wharton School of the University of Pennsylvania	0.001035
135	Yale University	0.003461
136	AE	5.63E-05
137	AUS	0.02016
138	BEL	0.000137
139	BRA	0.000252
140	CA	0.018198
141	CAN	0.01437
142	CHN	0.067924
143	CO	0.000715
144	CT	5.35E-05
145	DC	0.000282
146	DEU	0.001304
147	DNK	0.000441
148	ESP	0.000456
149	FL	0.000812
150	FRA	0.001637
151	GA	0.000244
152	GBR	0.001632
153	HKG	0.02361
154	ID	0.005392
155	IDN	0.00811
156	IL	0.002598
157	IND	0.048863
158	IRL	0.003311
159	ISR	0.008533
160	JPN	0.023959
161	KOR	0.007427
162	KS	0.000321
163	MA	0.009259
164	MD	0.000942
165	MEX	0.000385
166	MI	4.78E-07
167	MN	4.14E-05
168	MO	5.17E-05
169	MYS	0.017627
170	NC	0.002918
171	NH	0.00123

172	NJ	0.001253
173	NLD	0.000798
174	NV	0.000793
175	NY	0.003645
176	NZL	0.000116
177	OH	0.002192
178	OR	0.000508
179	PA	0.000468
180	PHL	0.001705
181	SGP	0.001807
182	SWE	0.005649
183	TN	0.00235
184	TWN	0.017864
185	TX	0.004534
186	UT	0.000992
187	VA	0.000526
188	WA	0.002922
189	WI	0.000344
190	ZAF	0.000282

Table 24 – XGB1 Binary IPO feature importance

	Feature	Importance
1	Administrative Services	0.001947
2	Advertising	0.006064
3	Agriculture and Farming	0.01816
4	Apps	0.001459
5	Artificial Intelligence	0.001642
6	Biotechnology	0.024658
7	Clothing and Apparel	0.001982
8	Commerce and Shopping	0.001661
9	Community and Lifestyle	0.000897
10	Consumer Electronics	0.001737
11	Consumer Goods	0.002731
12	Content and Publishing	0.000699
13	Data and Analytics	0.003105
14	Design	0.003251
15	Education	0.000656
16	Energy	0.003367
17	Events	0.002669

18	Financial Services	0.003099
19	Food and Beverage	0.002162
20	Gaming	0.000707
21	Government and Military	0.001454
22	Hardware	0.00114
23	Health Care	0.00388
24	Information Technology	0.001957
25	Internet Services	0.023007
26	Lending and Investments	0.035735
27	Manufacturing	0.027067
28	Media and Entertainment	0.00083
29	Mobile	0.000699
30	Music and Audio	0.000277
31	Natural Resources	0.044371
32	Navigation and Mapping	0.026664
33	Other	0.000588
34	Payments	0.00069
35	Platforms	0.000236
36	Privacy and Security	0.001654
37	Professional Services	0.002109
38	Real Estate	0.013195
39	Sales and Marketing	0.000933
40	Science and Engineering	0.01172
41	Software	0.067395
42	Sports	0.004001
43	Sustainability	0.006262
44	Transportation	0.006753
45	Travel and Tourism	0.001069
46	Video	0.000333
47	has_patent	0.004774
48	money_raised_series_a	0.00185
49	post_series_a_valuation	0.00043
50	series_a_investor_count	0.00442
51	top_investor_on_board_A	0.003258
52	money_raised_angel	0.001724
53	angel_investor_count	0.000939
54	money_raised_series_b	0.004856
55	post_series_b_valuation	0.000637
56	series_b_investor_count	0.019072

57	top_investor_on_board_B	0.000631
58	money_raised_seed	0.002823
59	post_seed_valuation	0.000187
60	seed_investor_count	0.024622
61	money_raised_pre_seed	0.001805
62	pre_seed_investor_count	0.000449
63	money_raised_grant	0.004127
64	not_first_venture	0.009599
65	successful_IPO_before	0.028755
66	successful_sale_before	0.018301
67	HEC Paris	0.002249
68	Harvard Business School	0.000737
69	Harvard Medical School	0.000908
70	Harvard University	0.000688
71	Hebrew University of Jerusalem	0.000693
72	Indiana University	0.000794
73	Massachusetts Institute of Technology	0.004121
74	Shanghai Jiao Tong University	0.00934
75	Stanford University	0.00037
76	Tsinghua University	0.000265
77	University of California Berkeley	0.001645
78	University of California Davis	0.004355
79	University of Illinois at Urbana-Champaign	0.000264
80	University of Washington	0.000294
81	Wharton School of the University of Pennsylvania	0.000742
82	AUS	0.02397
83	BEL	0.000207
84	BRA	0.000232
85	CA	0.007592
86	CAN	0.019119
87	CHN	0.035536
88	CO	0.001226
89	DEU	0.001447
90	DNK	0.000713
91	FL	0.000666
92	FRA	0.014693
93	GA	0.000556
94	GBR	0.001834
95	HKG	0.033973

96	IDN	0.052035
97	IL	0.001412
98	IND	0.043859
99	ISR	0.009734
100	JPN	0.082923
101	KOR	0.000596
102	MA	0.006895
103	MD	0.000266
104	MN	0.000182
105	MYS	0.02831
106	NC	0.001763
107	NLD	0.000389
108	NV	0.003485
109	NY	0.001802
110	NZL	0.000789
111	OH	0.00018
112	PA	0.004278
113	PHL	0.000127
114	SGP	0.003236
115	SWE	0.032931
116	TN	0.005651
117	TWN	0.030712
118	TX	0.01539
119	VA	0.00024
120	WA	0.002351
121	WI	0.000848
122	ZAF	0.000457

Table 25 - TOP 50 Investors on Crunchbase

raised_amount_usd	name	country_code	city
7.67E+10	SoftBank Vision Fund	GBR	London
4.3E+10	Alibaba Group	CHN	Hangzhou
3.82E+10	Tencent	CHN	Shenzhen
3.58E+10	SoftBank	JPN	Tokyo
3.07E+10	Tiger Global Management	USA	New York
3.03E+10	Kohlberg Kravis Roberts	USA	Hudson
2.87E+10	China Development Bank	CHN	Beijing
2.51E+10	Warburg Pincus	USA	New York
2.49E+10	Caisse de Depot et Placement du Quebec	CAN	MontrÃ©al

2.42E+10	Insight Partners	USA	New York
2.14E+10	Goldman Sachs	USA	New York
2.06E+10	Silver Lake	USA	Menlo Park
1.85E+10	General Atlantic	USA	New York
1.76E+10	Blackstone Group	USA	New York
1.68E+10	Temasek Holdings	SGP	Singapore
1.55E+10	Sequoia Capital	USA	Menlo Park
1.46E+10	EIG Global Energy Partners(EIG)	USA	Washington
1.46E+10	Altria	USA	Richmond
1.43E+10	Central Huijin Investment	CHN	Beijing
1.35E+10	Fortum	FIN	Espoo
1.35E+10	T. Rowe Price	USA	Baltimore
1.35E+10	Abu Dhabi Investment Authority	ARE	Abu Dhabi
1.31E+10	Andreessen Horowitz	USA	Menlo Park
1.23E+10	New Enterprise Associates	USA	Menlo Park
1.18E+10	Coatue	USA	New York
1.18E+10	Google	USA	Mountain View
1.17E+10	Apollo	USA	New York
1.11E+10	Glencore	CHE	Baar
1.1E+10	Berkshire Hathaway	USA	Omaha
1.1E+10	Hillhouse Capital Group	CHN	Changyang
1.08E+10	European Investment Bank	LUX	Luxembourg
1.03E+10	CVC Capital Partners	LUX	Luxembourg
1.01E+10	HNA Group	CHN	Haikou
1E+10	Accel	USA	Palo Alto
9.91E+09	TCV	USA	Menlo Park
9.89E+09	Sequoia Capital China	CHN	Beijing
9.85E+09	Federal Government of Germany	DEU	Berlin
9.75E+09	GIC	SGP	Singapore
9.63E+09	DST Global	USA	California
9.37E+09	The Carlyle Group	USA	Washington
9.21E+09	Saudi Arabia's Public Investment Fund	SAU	Riyadh
9.02E+09	JAB Holding Company	LUX	Luxembourg
8.83E+09	TPG	USA	San Francisco
8.44E+09	Baidu	CHN	Beijing
8.33E+09	Apple	USA	Cupertino
8.18E+09	Canada Pension Plan Investment Board	CAN	Toronto
7.95E+09	JP Morgan Chase	USA	New York

7.71E+09	General Motors	USA	Detroit
7.5E+09	IDG Capital	CHN	Beijing

Figure 19 – XGB1 IPO confusion matrix

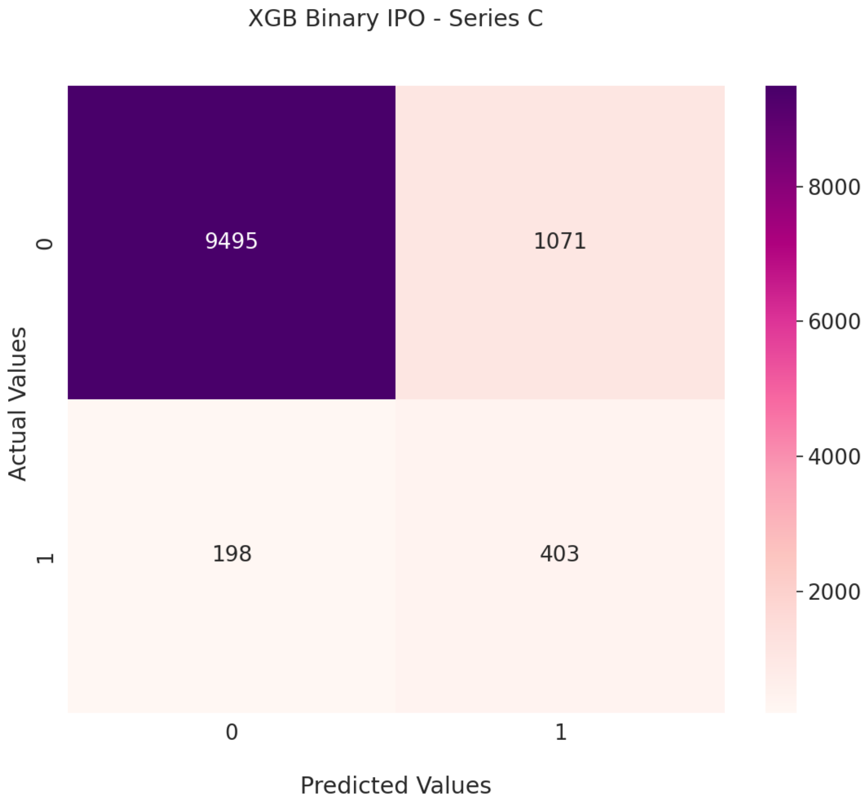


Figure 20 – XGB1 Acquisition confusion matrix

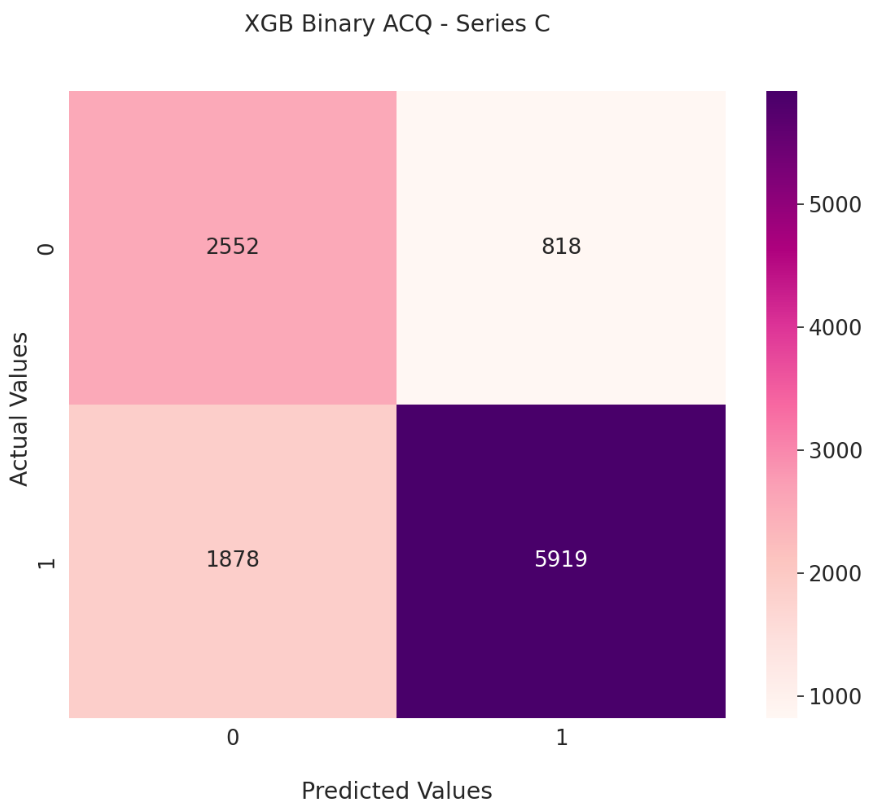


Figure 21 – XGB2 IPO confusion matrix

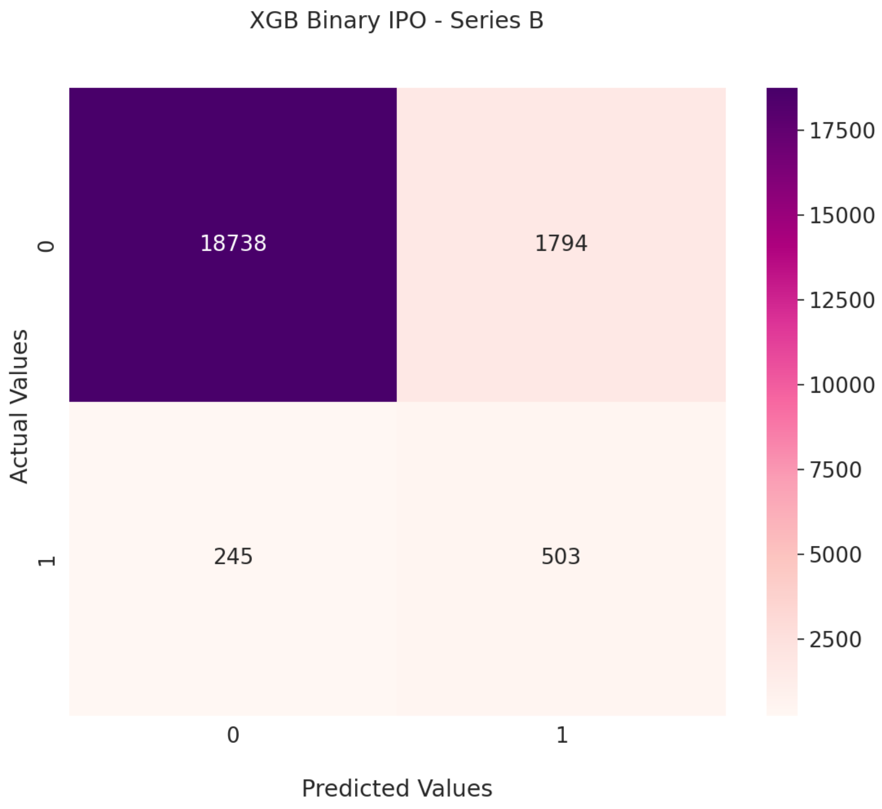


Figure 22 – XGB2 Acquisition confusion matrix

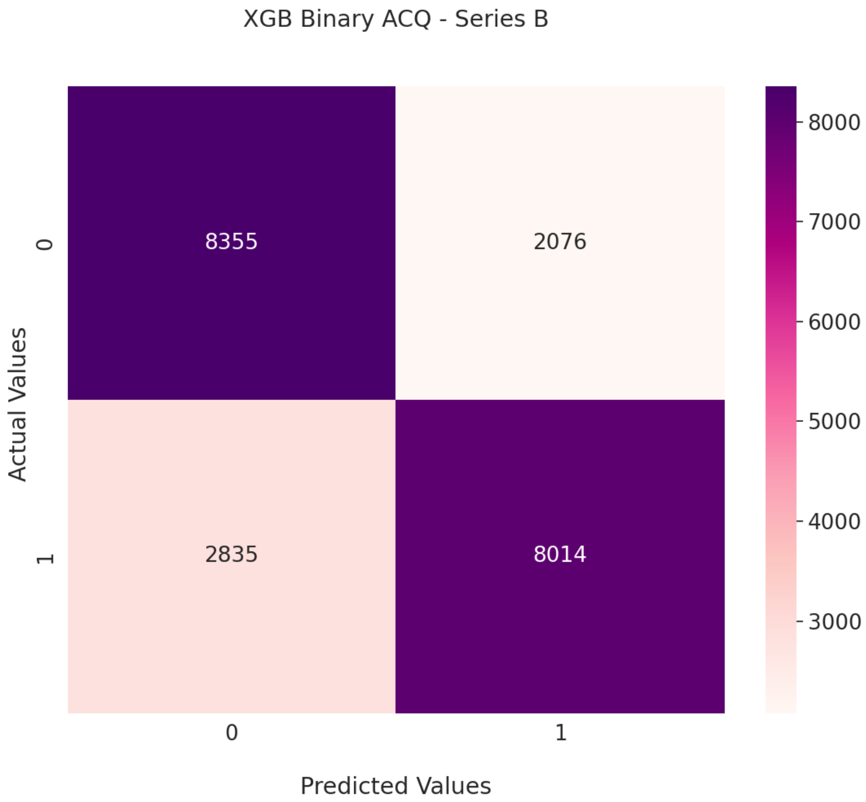


Figure 23 – XGB2 Status confusion matrix

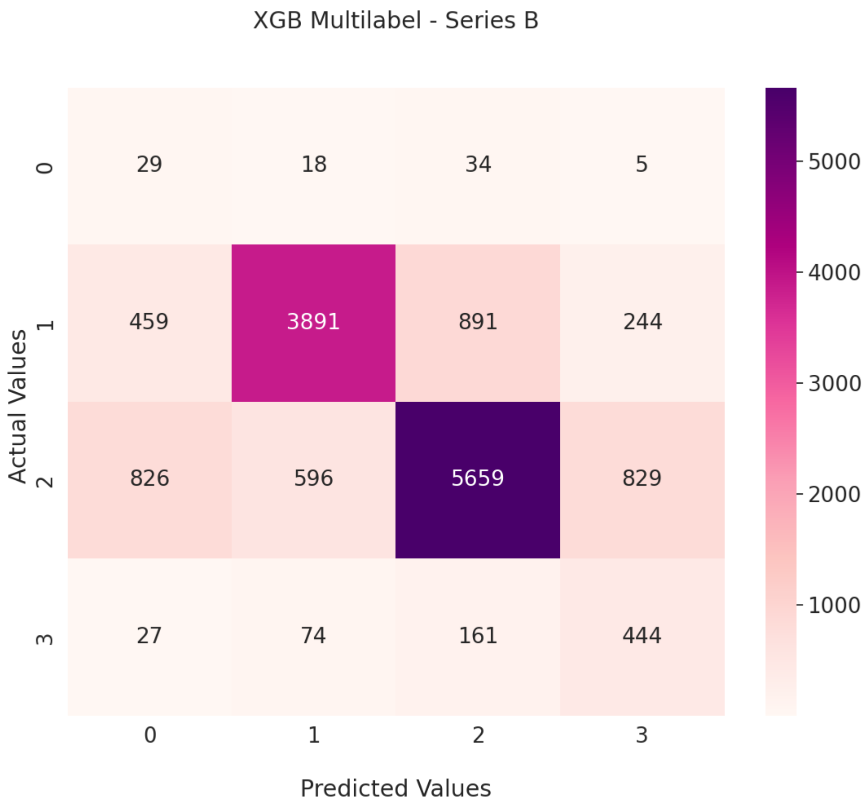


Figure 24 – XGB3 IPO confusion matrix

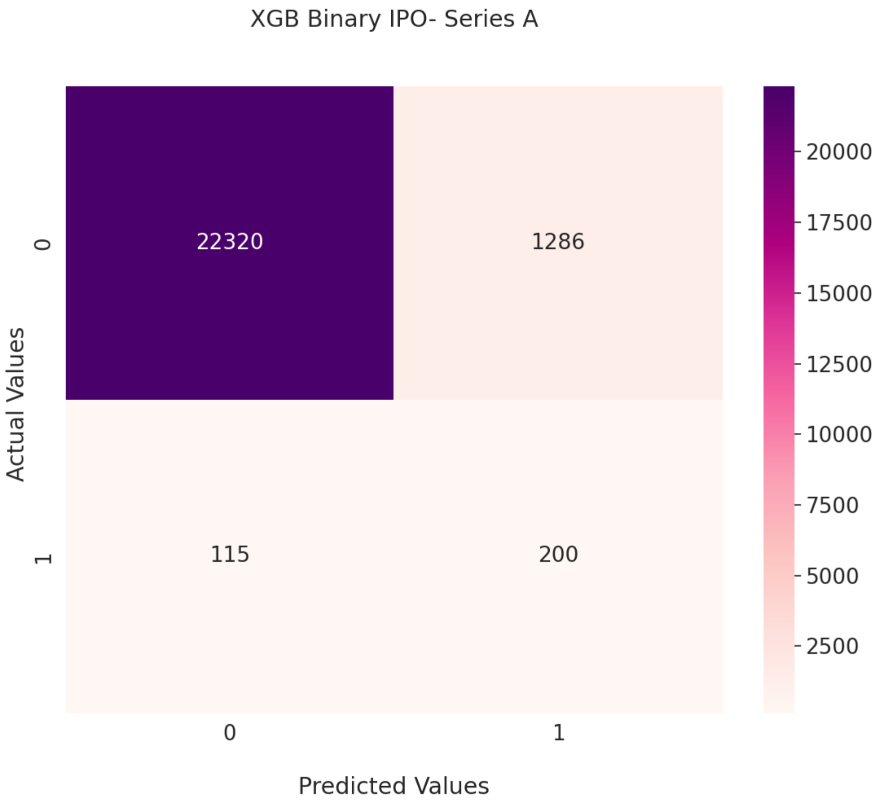


Figure 25 – XGB3 Acquisition confusion matrix

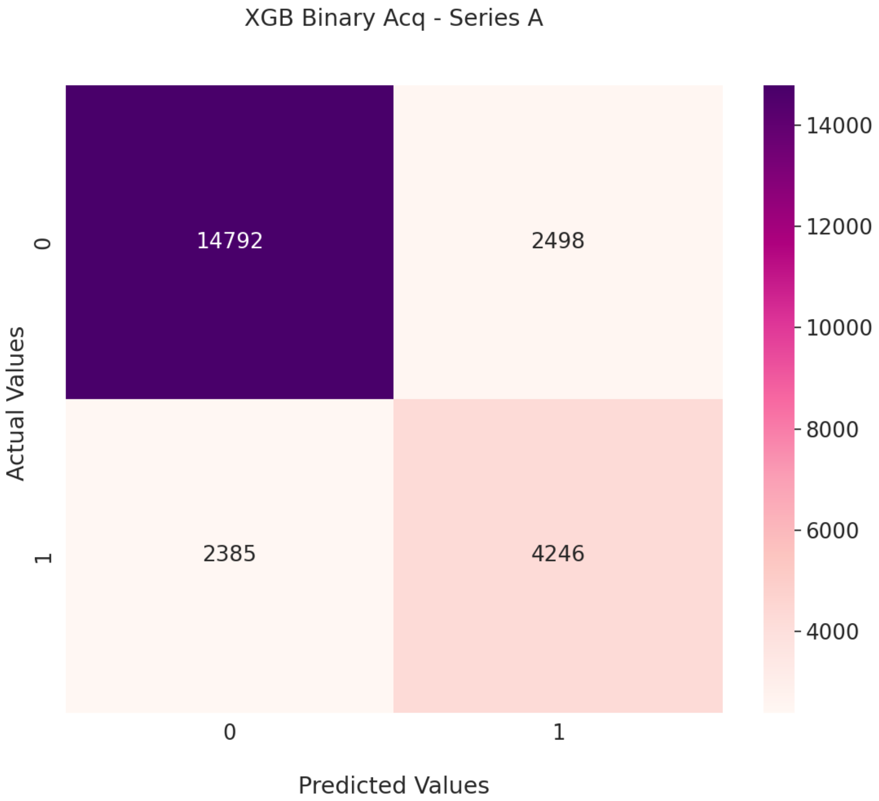


Figure 26 – XGB3 Status confusion matrix

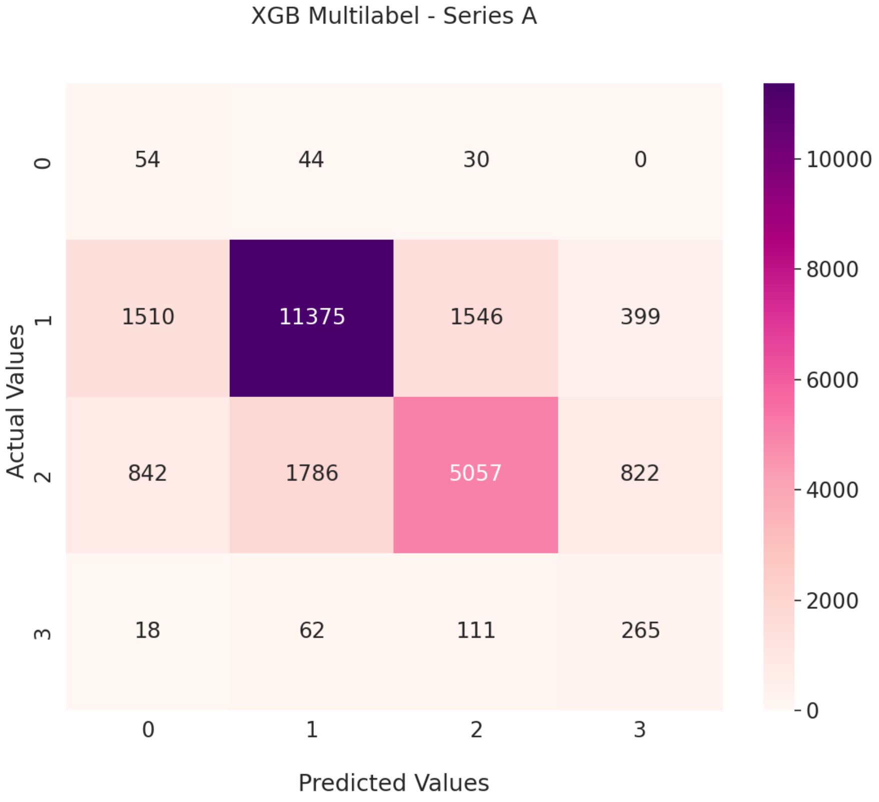


Figure 27 – XGB4 IPO confusion matrix

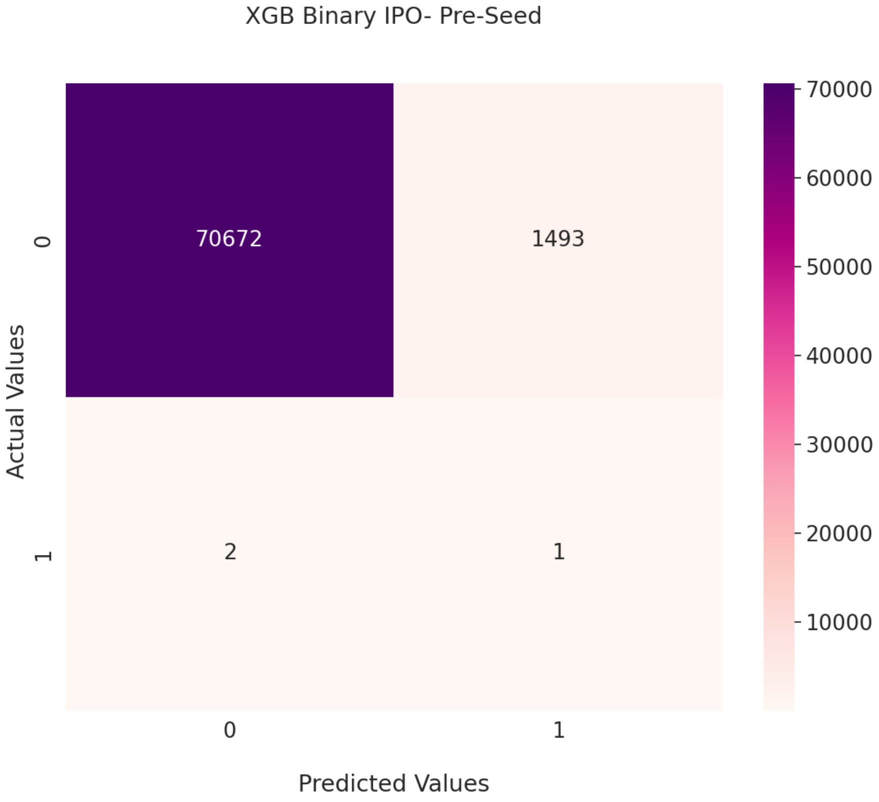


Figure 28 – XGB4 Acquisition confusion matrix

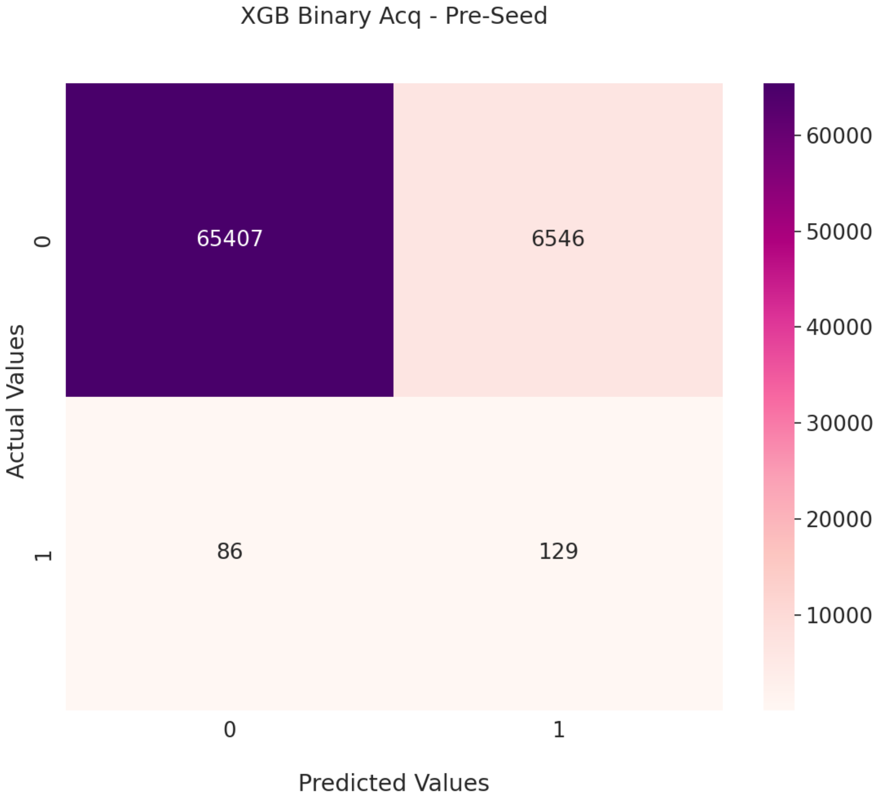


Figure 29 – XGB4 Status confusion matrix

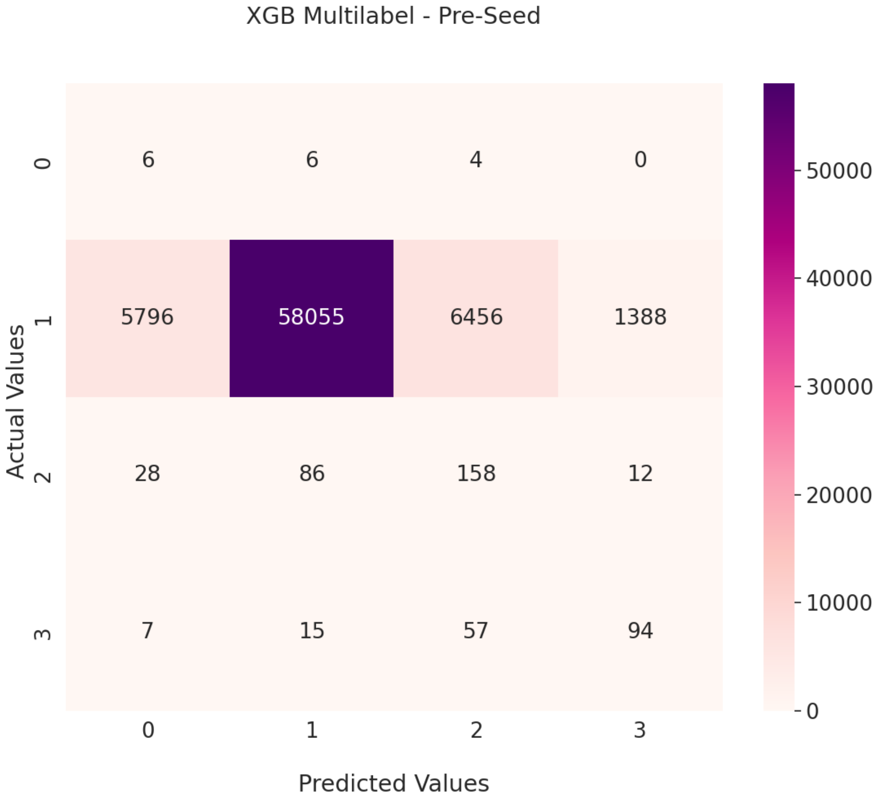


Figure 30 – Naive Bayes IPO confusion matrix

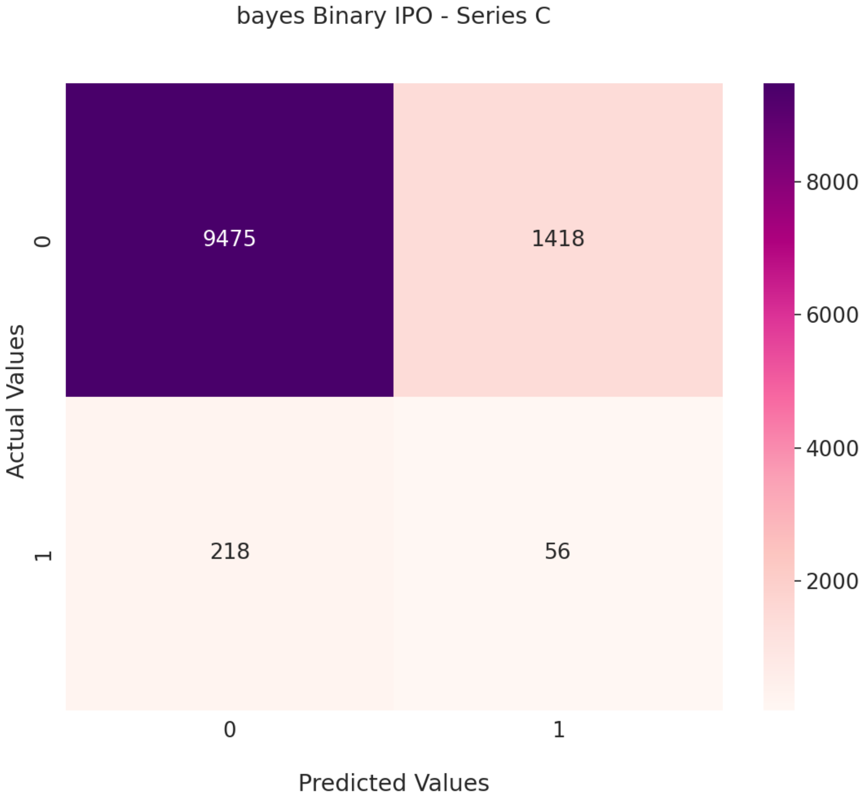


Figure 31 – Naive Bayes Acquisition confusion matrix

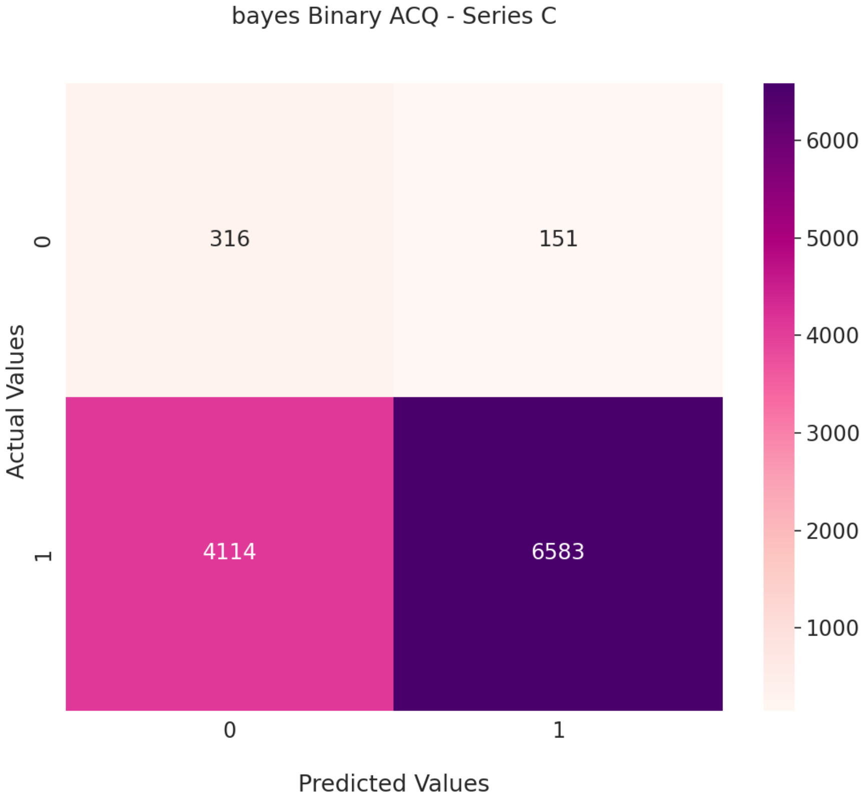


Figure 32 – Naive Bayes Status confusion matrix

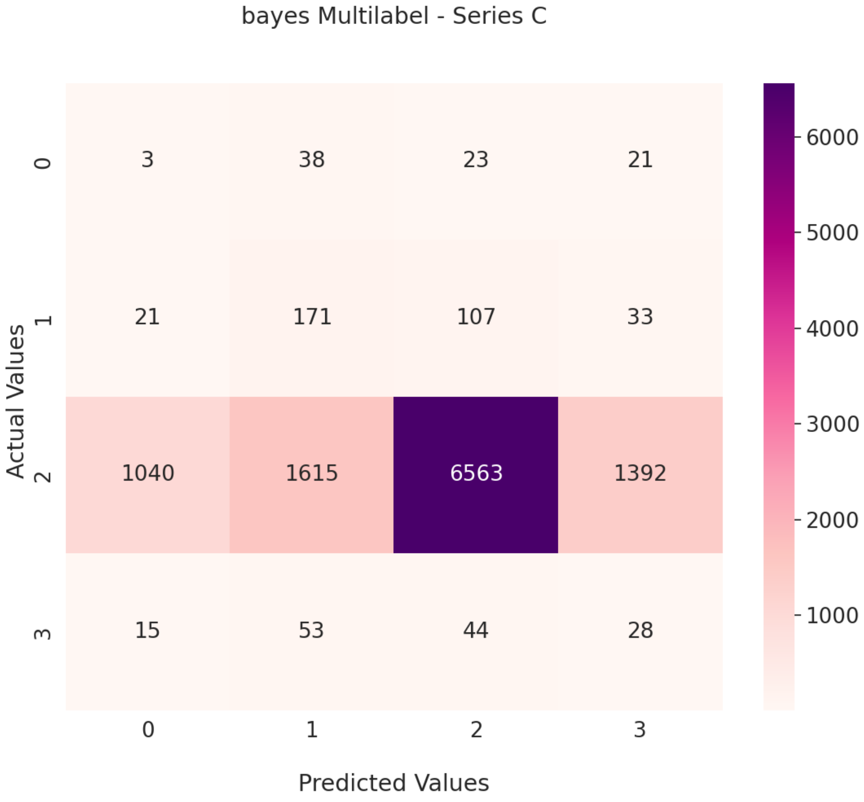


Figure 33 – RF1 IPO confusion matrix

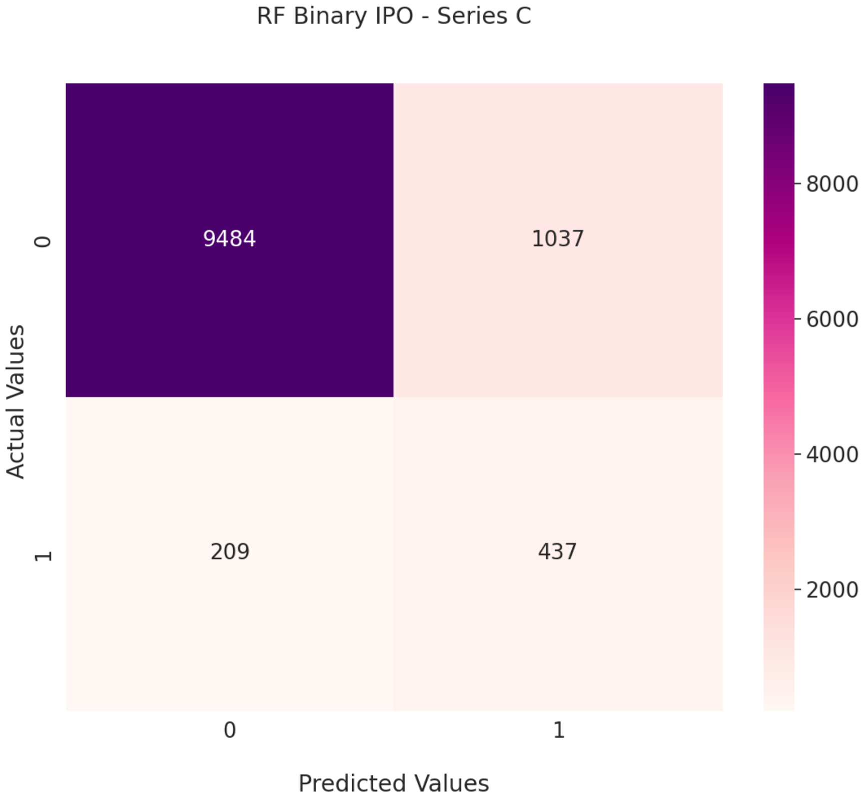


Figure 34 – RF1 Acquisition confusion matrix

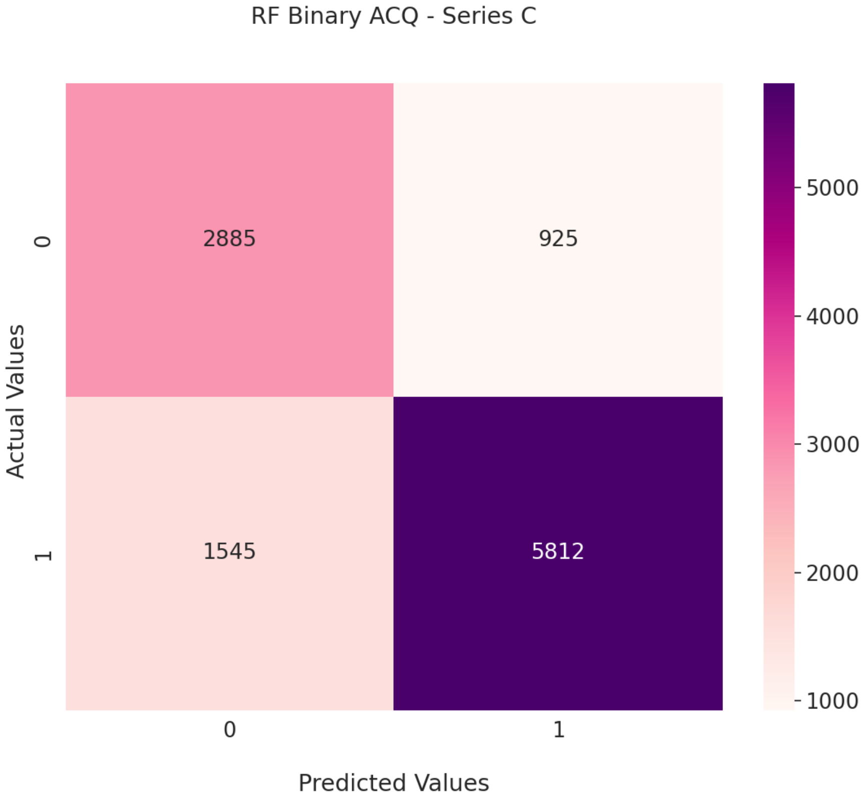


Figure 35 – RF2 IPO confusion matrix

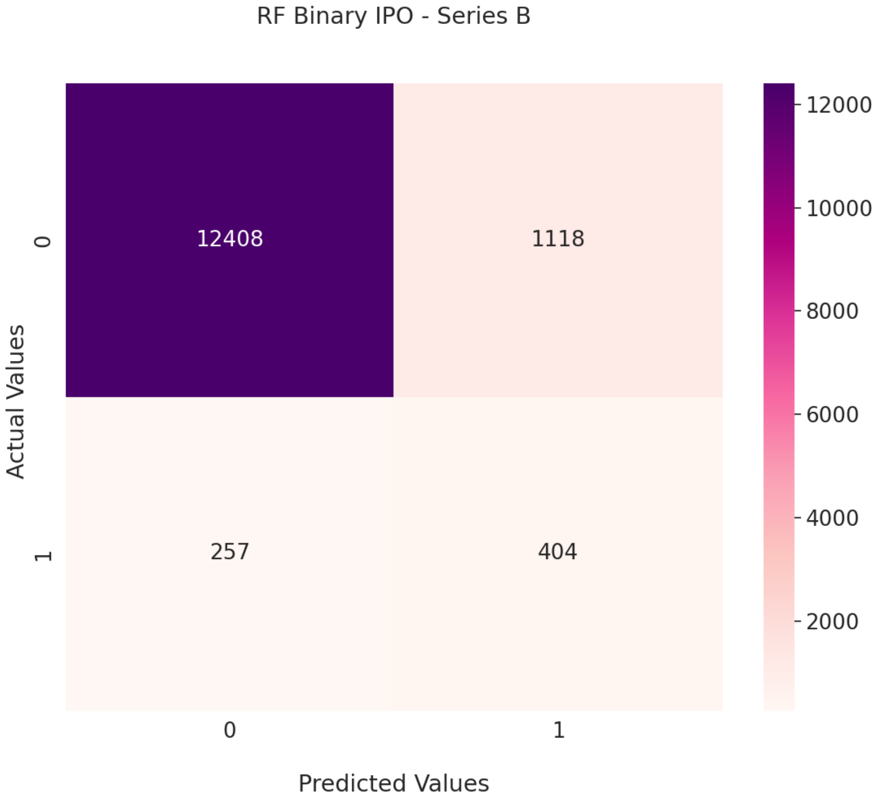


Figure 36 – RF2 Acquisition confusion matrix

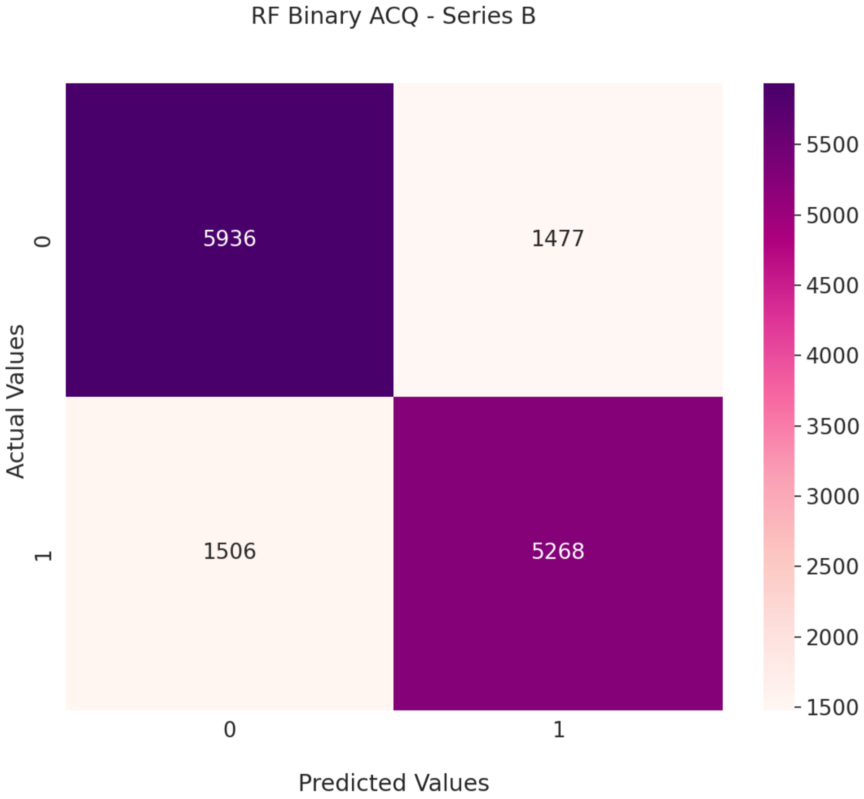


Figure 37 – RF2 Status confusion matrix

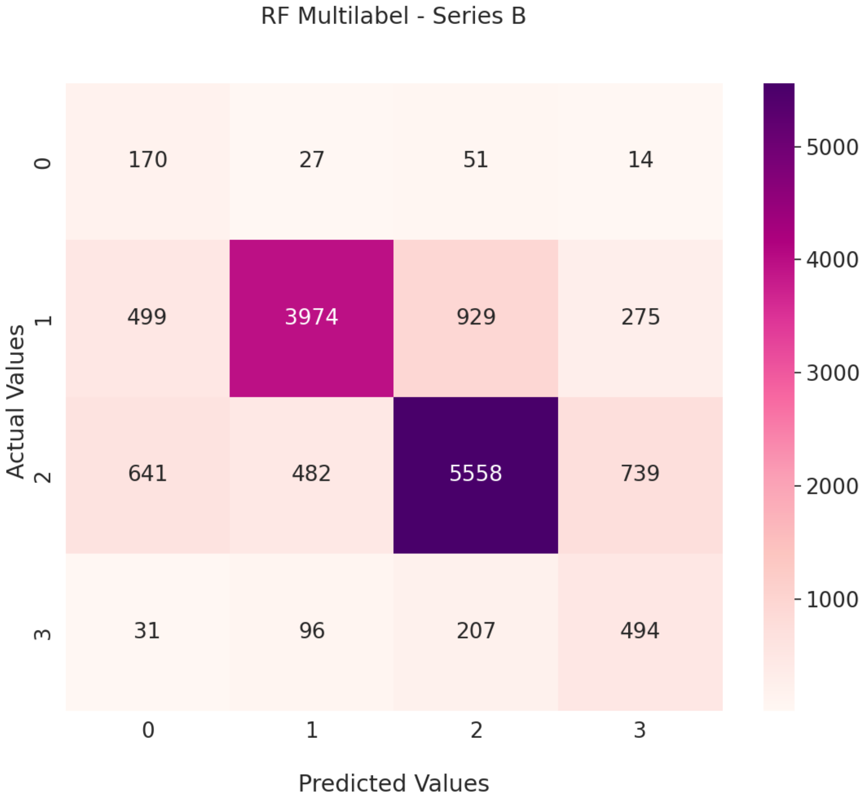


Figure 38 – RF3 IPO confusion matrix

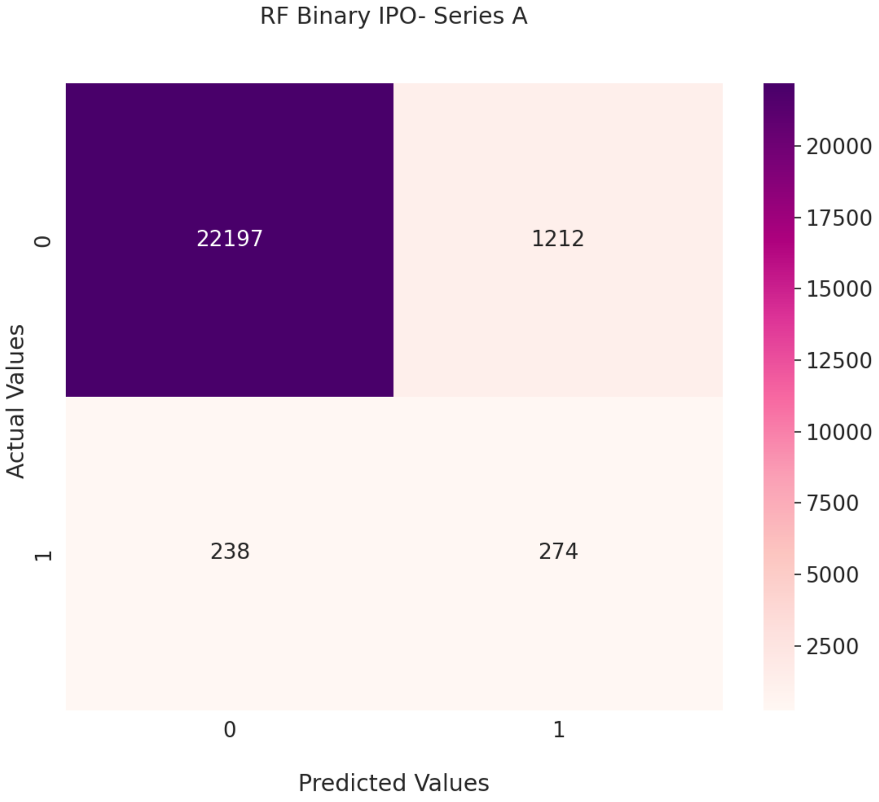


Figure 39 – RF3 Acquisition confusion matrix

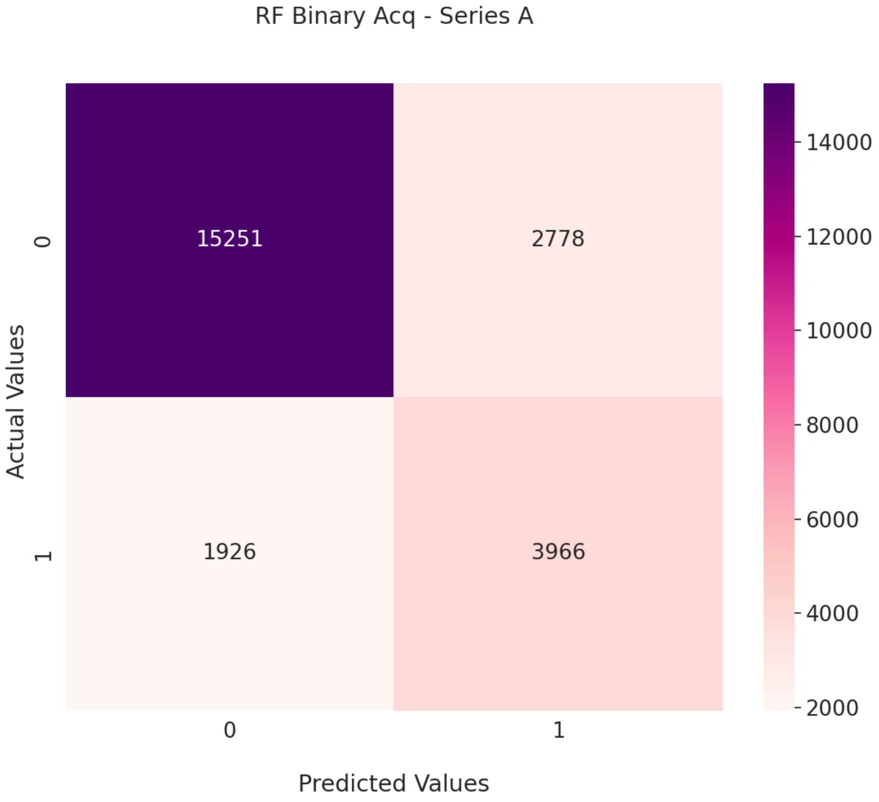


Figure 40 – RF3 Status confusion matrix

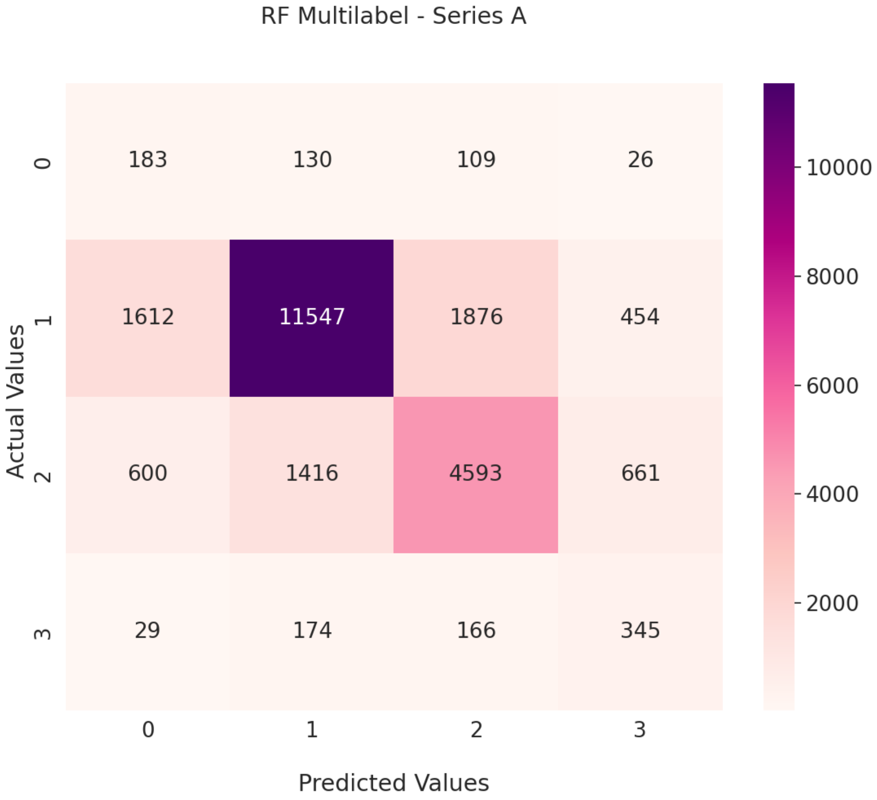


Figure 41 – MLP1 IPO confusion matrix

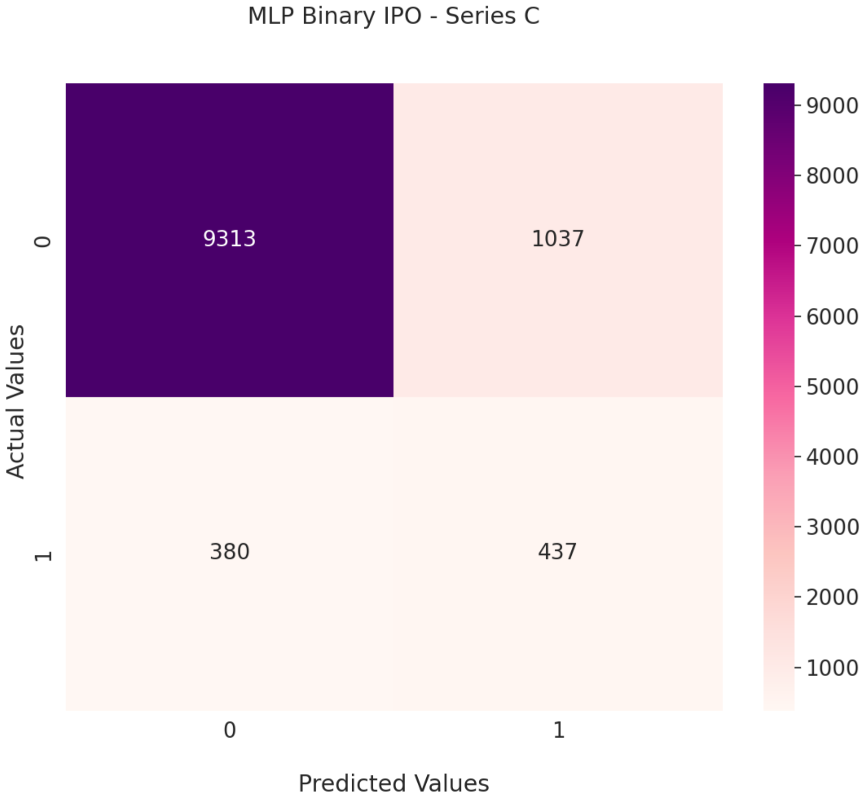


Figure 42 – MLP1 Acquisition confusion matrix

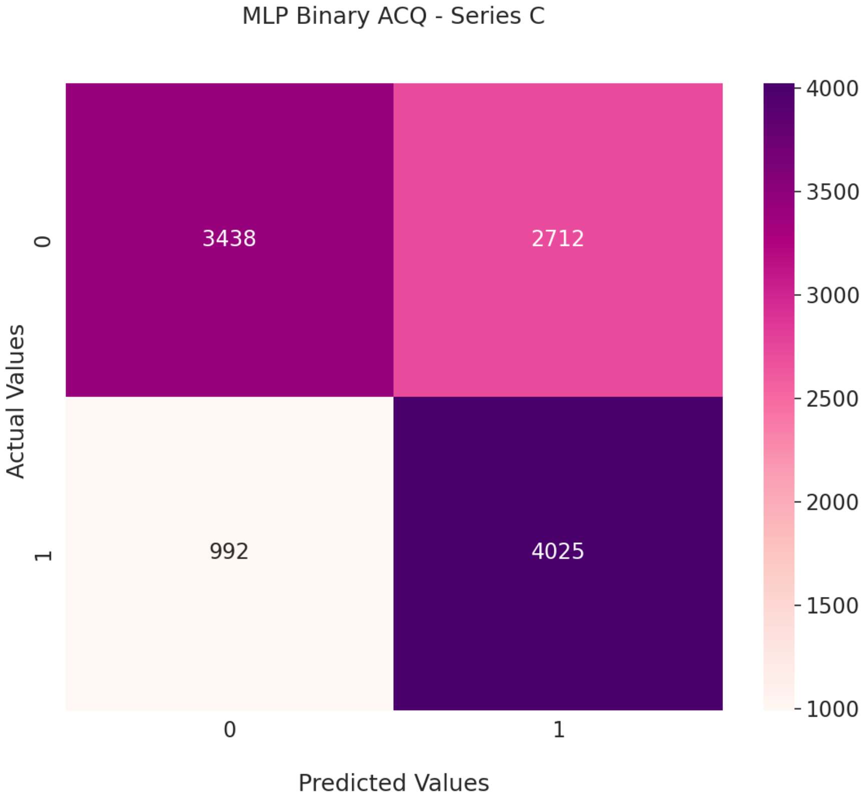


Figure 43 – MLP2 IPO confusion matrix

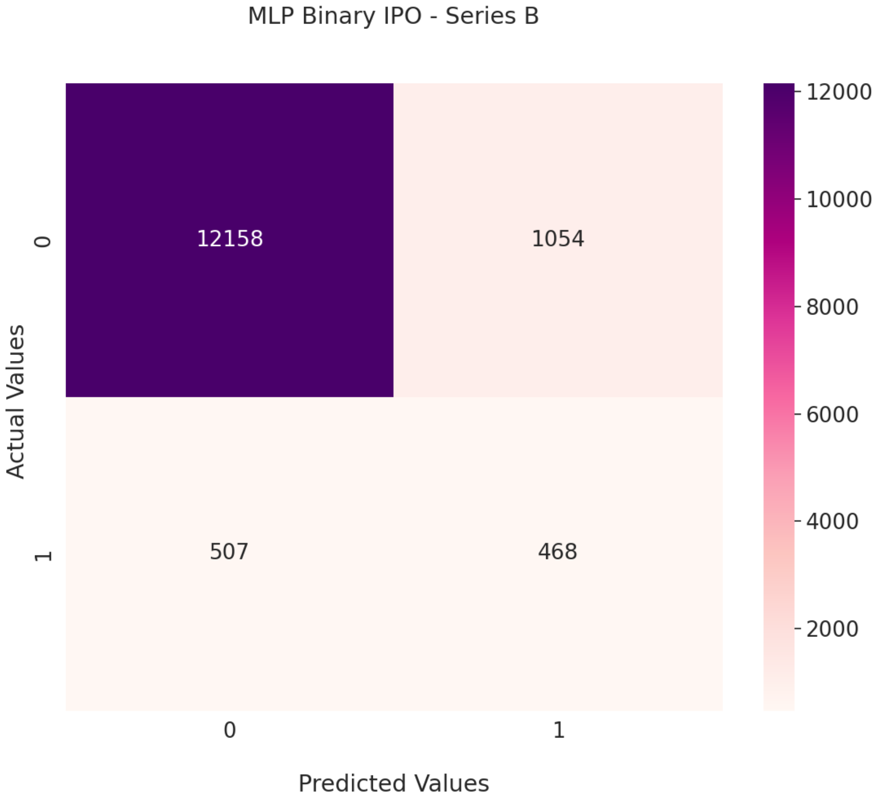


Figure 44 – MLP2 Acquisition confusion matrix

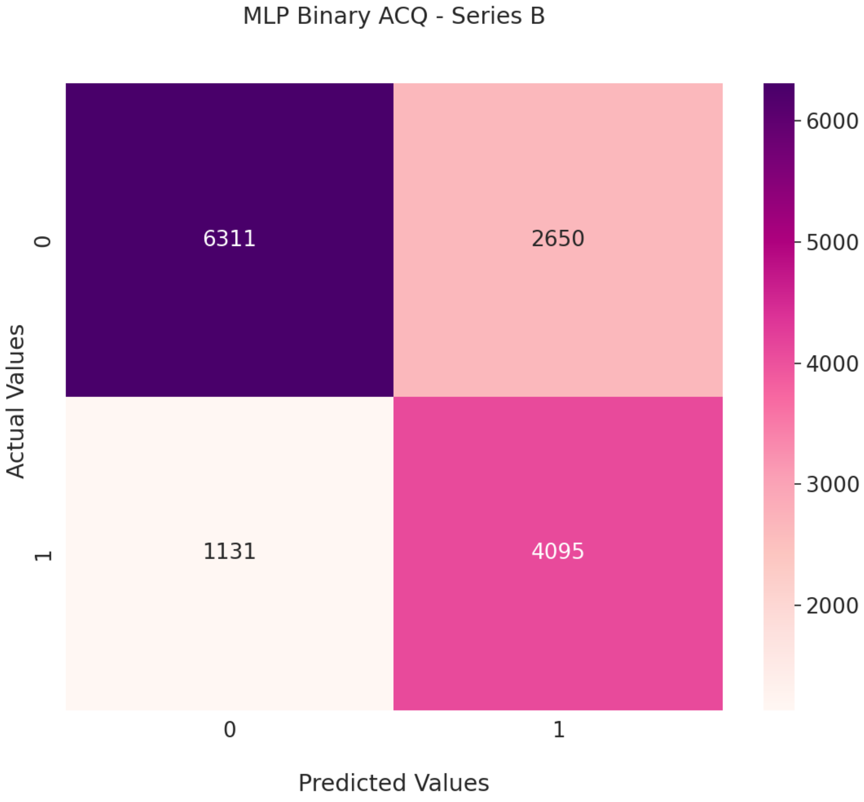


Figure 45 – MLP2 Status confusion matrix

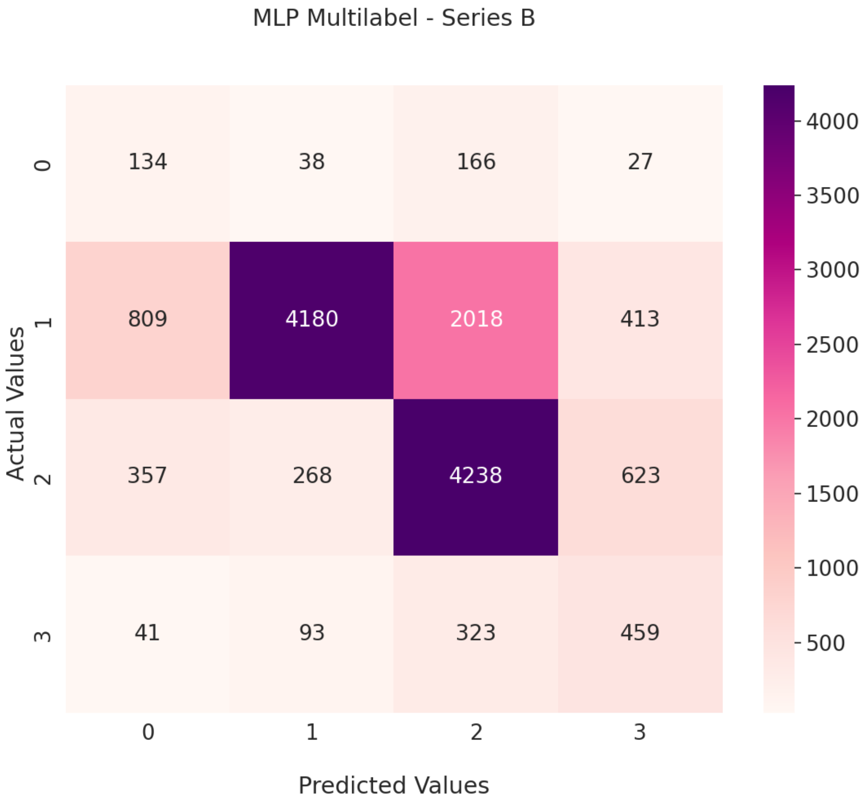


Figure 46 – MLP3 IPO confusion matrix

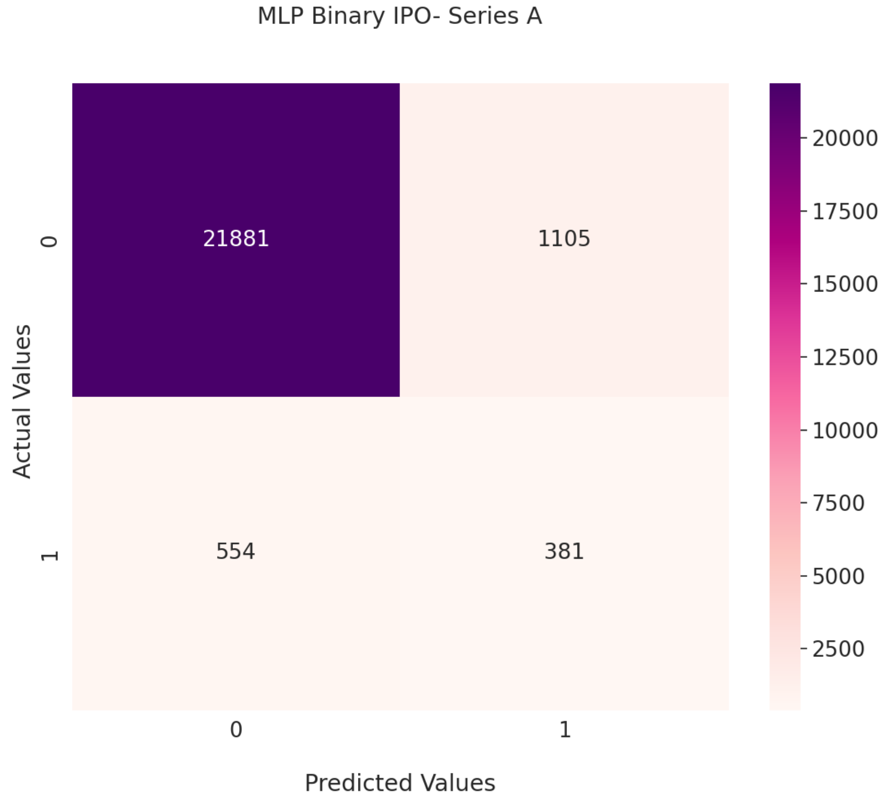


Figure 47 – MLP3 Acquisition confusion matrix

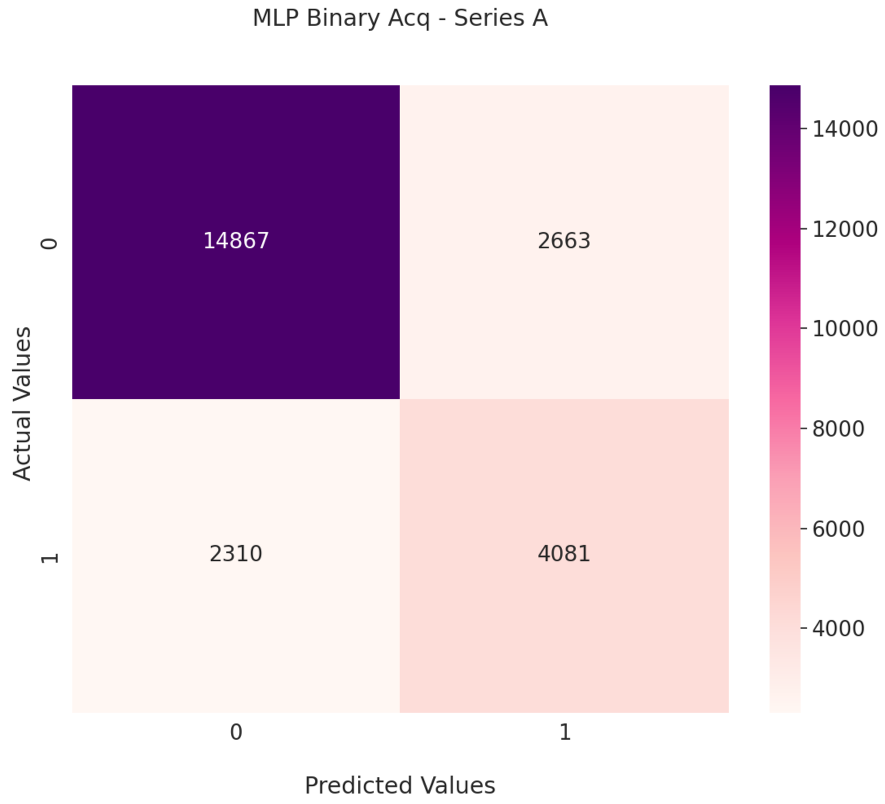


Figure 48 – MLP3 Status confusion matrix

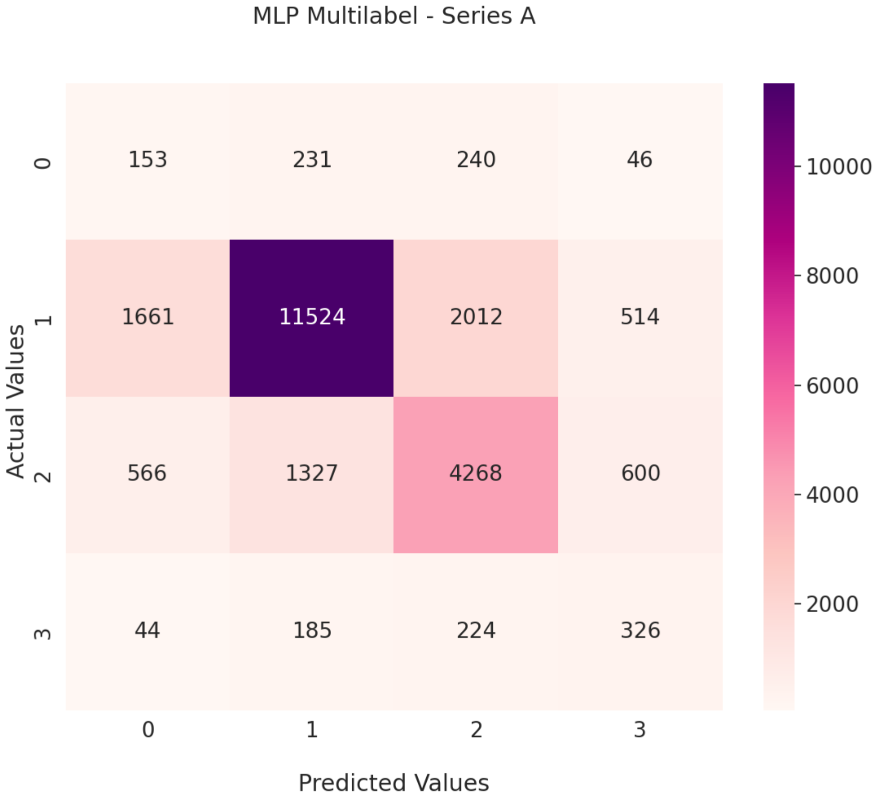


Figure 50 – MLP Social Multilabel – Series C dataset

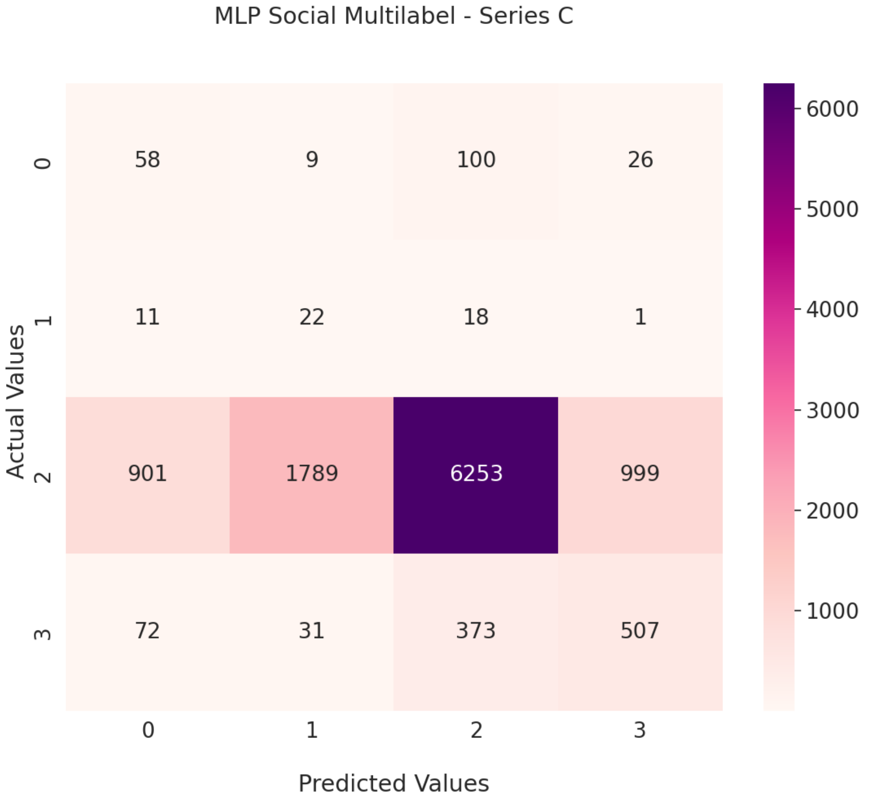
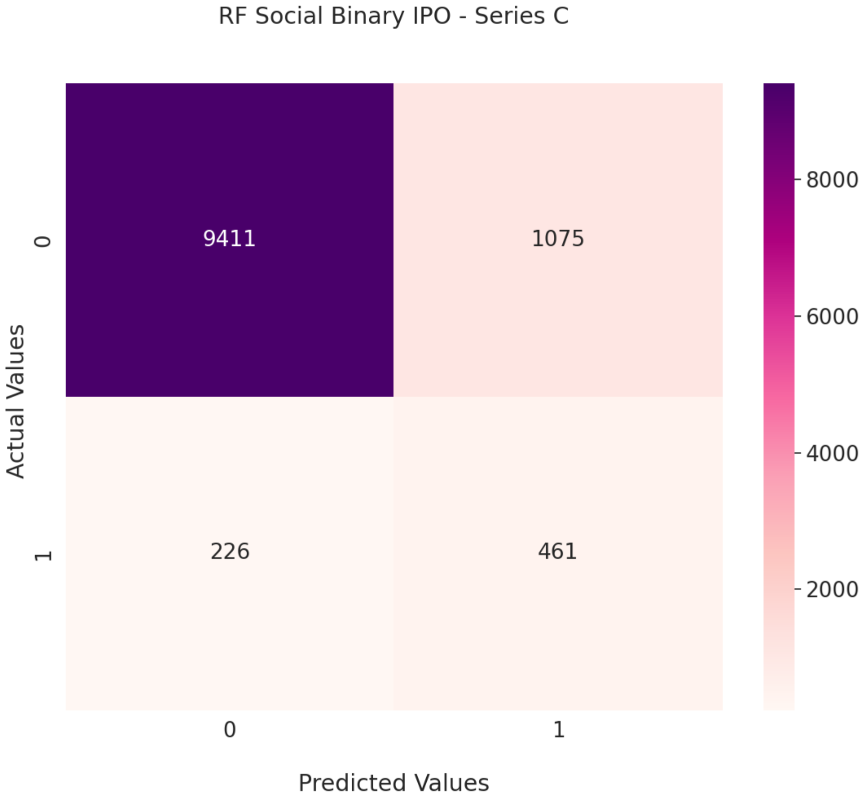


Figure 51 – Random Forest Social Binary – Series C dataset


RF Social Binary IPO - Series C
Actual Values
Predicted Values