A Comparative Study of Start Up Funding & The Use of Artificial Intelligence Powered by Machine Learning Algorithms to Predict Possible IPO vs Non-IPO Outcomes

This research work has been created as a partial fulfilment of the requirements for the master's in finance & data analytics at the University of Stirling. I hereby declare that all the works contained in this research are original and have been produced by me as part of my academic pursuits. Special appreciation & gratitude to Crunchbase for their generous provision of data, a crucial element in the success of this study. I would also like to express my gratitude to everyone who contributed in various ways toward the outcome and completion of this research. Any external sources used in this research have been appropriately cited and credited in accordance with academic standards and guidelines. Should there be any concerns or inquiries, please feel free to contact me at the provided contact details herein. Thank you.

Said Olanrewaju – September 2023
E-mail addresses: olar.said@icloud.com,
ols00082@students.stir.ac.uk

Table of Contents

Abstract

The start-up ecosystem is an ever-evolving industry for both the business owners/founders and the investors alike, every year, thousands of new businesses and innovations are created at the same time, several billions of dollars are invested by private equity firms, venture capitals and several other investors. In this research we explore the transformative potential of machine learning and artificial intelligence in predicting the probability of a possible IPO using binary classification by analyzing key features/variables from data provided by CrunchBase. The positive class represents companies that successfully undertake an IPO while the negative class represents the companies that did not.

In this research we adopt an ensemble approach to model development, utilizing a combination of Logistic Regression, XGBoost, Random Forests, K-Nearest Neighbors and Neural Networks to make predictions alongside with SHapley Additive explanations (SHAP) to address the concerns of explainability and transparency and to justify the predictions. The outcome of the research suggests that of all the different models trained and tuned for this experiment, Gradient Boosting (XGBoost) performed the best with a 91.6% recall rate, minimizing false results while logistic regression was the least performing model with about 16.8% recall rate. Although SHAP is useful for explaining predictions and providing insights into individual predictions, it however does not imply causality.

Keywords:
Start-up Companies, Venture Capital, Machine Learning, Model Training, Prediction, Explainability, Exploratory Data Analysis, Start Up Funding, XGBoost, SHAPley, Neural Networks, Random Forests, K-Nearest Neighbors, Exit Prediction, Private Equity, Crunchbase, Investment, Artificial Intelligence.

Comment by the writer: In the following research paper, words such as the United Kingdom and the United States of America, artificial intelligence, and machine learning will be shortened to UK, US, AI and ML respectively.

Introduction

As the global business landscape continues to evolve over the years, the role and impact of start-ups in the modern economy continues to hold a transformative position with respect to its contribution to technological innovation, job creation and most importantly economic growth. In the United Kingdom, thousands of new companies emerge each year and with the rapidly growing field of innovation spreading across industries - from self-driving cars to payment systems, medical diagnosis tools, artificial intelligence and blockchain innovations experienced in recent years, start-ups and emerging companies no doubt have the potential to not only revolutionise different aspects of our lives and standards of living but also have a huge impact on economic growth and industry performance. From a recent publication by the US International Trade Administration (2022), The UK AI market is worth $21 billion and is expected to grow to add $1 trillion to the UK economy by 2035. Tech Nation reports the UK rank second globally for start-up funding in 2022 (Tech Nation, 2022).

As the industry evolves, with start-up companies in the fore front, recent data from Companies House reveals that between 2021 and 2022, there were over 753,168 company incorporations in the UK alone which is a decrease of 7.1% when compared to the period between 2020 and 2021 – (Companies House, 2022). The report further reveals that, despite fluctuations, the number of incorporations over time have increased at a steady rate.

If we look at the ecosystem from the investment perspective and investors point of view, According to a report published by Dealroom - an online portal that provides data and intelligence on high growth companies, UK start-ups have raised about $8.9 billion so far and currently dominate the European start up and investment landscape with London fintech start-ups raising more investment than any other global hub in 2022 ahead of New York and the Bay Area. However, while this further indicates a vibrant entrepreneurial landscape considering that 22% was raised by fintech start-ups, 14% raised by energy, 12% by health care and the rest coming from across other sectors, such as transportation, real estate, robotics, security and food amongst several others – (Dealroom, 2023).

With about 60% of start-ups expected to fail in first 3 years and more than 90% of start-ups ending as failure (Beauhurst, 2022) understanding the factors that contribute to these different outcomes and being able to predict a company's possible outcome becomes of immense value to investors, stakeholders, policy makers and even to entrepreneurs & business owners.

From inception to viability of the business to successful exits by investors, the journey of a start-up can lead to diverse destinations, including, but not limited to successful acquisitions by larger companies and initial public offerings. Meanwhile, other start-ups may face challenges that ultimately results in their disappearance & failure.

As such, this dissertation explores the potentials of machine learning to calculate the probability of an IPO with high levels of confidence leveraging on publicly available data and focusing on key features.

In recent times, technological advancements in the realm of artificial intelligence (AI) and machine learning (ML) have demonstrated significant potential to revolutionize various industries, including finance and investment.

In this project, we focus on predicting the probability of start-up IPO using machine learning algorithms. The input features for our predictive models encompass text features, such as industry category lists and location, and numerical features, including the amount of funding a company has raised. Leveraging on supervised machine learning our models are trained to be able to predict the probability of a possible IPO using binary classification by analysing key features/variables from data provided by CrunchBase. The positive class represents companies that successfully undertake an IPO while the negative class represents the companies that did not.

In this research, we adopt an ensemble approach to model development, utilizing a combination of Logistic Regressions, XGBoost, Random Forests, K-Nearest Neighbors & Neural Networks. The process involves exploratory data analysis, feature extraction, model training, and evaluation based on known events. To cover explainability and transparency of our machine learning models, we employ SHAP (SHapley Additive exPlanations) to justify the predictions. This method enables us to perform a detailed analysis of each prediction, focusing on specific data points related to an organization. SHAP values are rooted in cooperative game theory and provide a way to distribute "credit" or "blame" for a prediction among the different features. It helps us understand how these data features influenced the predicted outcomes of our machine learning models - (Lundberg & Lee 2018)

The primary objectives of this study are twofold: First, to develop predictive machine learning models that predicts IPO Status as Yes or No, and second, to assess the accuracy and performance of the machine learning models in predicting these outcomes by using SHAP (SHapley Additive exPlanations) to explain the predictions.

To achieve this, we work with data provided by Crunchbase – a digital platform that provides business information about private and public companies. Originally founded in 2007, over the years, Crunchbase has been known as a leading destination for company insights providing data from early-stage start-ups to fortune 500 companies and often served as the number one go to site for investors and business owners alike – (Crunchbase Wikipedia, 2023)

Data provided by Crunchbase for this research includes, but not limited to, organization data, people data, investors, funding rounds, IPOs, acquisitions, and investments, amongst others. As part of the methodology used in this research, which will be outlined in subsequent chapters, the data obtained were split into 3 at the ratio of 60-20-20 to ensure that a balanced approached is implemented. 60% of the data will be used to train the model, 20% for validation and development, while the remaining 20% will be used for testing and to evaluate the model's final performance and ability to work with unseen data - (Ismail, 2022)

The implications of such predictive models are far-reaching. On one hand, investors can make more informed decisions by gauging a start-up's likelihood of certain outcomes based on data-driven insights. On the other hand, policymakers can utilize these models to foster a conducive ecosystem for entrepreneurial growth, driving economic progress and job creation.

While this study aims to provide valuable insights into the use of AI-powered ML algorithms in start-up funding evaluation, certain limitations should be acknowledged. The availability of comprehensive and reliable start-up data, as well as potential biases in the historical data, may impact the accuracy of predictions. Additionally, the success of a start-up may be influenced by various unpredictable external factors beyond the scope of this research.

In conclusion, this introduction establishes the foundation for a research project that integrates machine learning and start-up financing in the United Kingdom. By incorporating insights from the provided examples and templates, our research endeavours to contribute to the understanding of key factors influencing start-up outcomes.

As we proceed with subsequent chapters, we will delve into the methodology, data analysis, results, and discussion, presenting the findings and implications for the UK start-up ecosystem and financial landscape.

The rest of the paper proceeds as follows. In Chapter 2, we briefly discuss related literature. Chapter 3 covers the research design including the data and methodology used during the research. In Chapter 4, we discuss the results and outcomes while chapter 5 covers summary and conclusions.

Chapter 2: Literature Review

With respect to start-up funding and the exploration of machine learning algorithms to predict possible outcomes, previous studies, and literature from this academic body of work reveals the vitality and significance of various funding rounds and their impact on start-up growth trajectory. However, as we review past literature, several key points to take note for the purpose of this research includes the fact that our exploration is more focused on the use of datasets to carry out predictive analysis and most important is understanding the evolution of this space over the years. From availability of data to technological innovation, changes to the financial system and the VC/private equity landscape, the emergence of digital money, decentralized venture capital firms, CBDCs and other financial instruments as well as the advancements in AI and ML technologies and how predictive, generative AI and several other new innovations have transformed financial decision making in recent years both coming together to impact how start-ups are affected.

2019, Hengstberger conducted a comprehensive study titled "Increasing Venture Capital Investment Success Rates Through Machine Learning" for his master's thesis in Mathematics and Finance at Imperial college London Department of Mathematics. In his research, Hengstberger (2019) explored the evolution of data availability and reliability for venture capital. The publication further reveals that the research of quantitative methods to predict company events up until about 2015 relied mostly on self-created survey data which were limited in size and scope. In the venture capital setting, which this research focuses on, large scale databases for this topic have only been established in the last 13 years and have only been able to mature in scale and scope over the last 5 years. From Hengstberger's (2019) contributions, we learn that Crunchbase so far is still the most used data source in recent start-up related research.

Much as there is a growing literature on analysing venture capital investments and predicting possible outcomes of start-ups, Ross et al.'s 2021 study explores the application of machine learning in predicting start-up success and exit outcomes and reveals that the literature may be broken down into two broad strands: (i) Identifying successful investors and (ii) identifying successful start-up investments. From the research, we see how these two strands contribute to defining what features and criteria to look out for while carrying out this research as we further delve deeper into the data structure, data sets and variables used in training the machine models and how they can contribute to the eventual outcome.

Understanding these two sides of the coin i.e., start-ups possible outcome from the organization/founder's perspective and from the investors/investment perspective creates the necessity to explore the possible impact of other external factors, such as behavioural patterns, reputations, traditional methods, and other factors common to both sides of the spectrum, as well as how these could affect possible outcomes. To give an example, Dellermann et al. (2017) combined machine learning with the traditional human approach and in their work, they had explored unknowable risk and observe that humans fill in the gaps for where machine fails. By using a hybrid intelligence model, which is a combination of machine intelligence and collective human intelligence i.e. using people's expertise and judgement to make decisions without relying heavily on machines, they concluded that by combining both machine and human aspects in making predictions for possible outcomes, investors will not only make better decisions but can also minimize the possible chances of losses and errors by considering machine intelligence to be leveraged as feedback for the crowd. For the use of machine

intelligence, they had used the following learning algorithms: Logistic regression, naive bayes, support vector machine, artificial neural networks, and random forests. While their research might have accounted for unknowable risk and considered human impact on possible outcomes, they had only used a data sample that consists of 1,500 start-ups from different industries and split the data into two - training and testing using a 10-fold cross validation approach. However, due to the limitation of data used, possibilities of machine learning errors like high Variance, overfitting, underfitting, difficulty in hyper parameter tuning amongst several others could arise, thus impacting final outcomes.

2020, Ang, Chia, and Saghafia conducted a very similar research work to ours, where they used start-up data provided by Crunchbase too. In their study however, they review the full entrepreneurial ecosystem by taking stock of the key fundraising activities in major cities around the world. According to the publication, the success of a start-up is more complicated from the founder's perspective, therefore raising the question - What does start-up success mean? What exactly are possible outcomes of start-ups? Do all start-ups have to go through the same process and routes? Is it possible that an outcome of the start-up could mean success from the investor perspective and failure from the founder's perspective? From their report, findings suggest that, in some cases, success could mean contributing to social good and in some cases, interestingly, there could be a very long time between the founding of a start-up and the direction it takes. Thus, it is important to define outcomes from both sides of the spectrum without defining any particular outcome as success or not. Understanding this aspect thus further lays the foundation for what variables are to be considered in training the learning algorithm. Additionally, Ang et. al (2020) suggest that labelling start-ups that are on their way to success but have not achieved it yet as unsuccessful is often an unfair assessment of their potential.

For this research and dissertation, we shall therefore be focusing mostly on predicting yes or no outcomes. Therefore, once our machine algorithms are trained and can accurately predict possible outcomes of a start-up, it is best to not conclude that as either successful or not as this will be based on the perspectives of either the investor or the founder.

Previous literature has also suggested certain variables and criteria that could contribute to possible outcomes for start-ups. For instance, tech start-ups often have a different path compared to medical and health related start-ups. To further explain, Ajit et al. 2022 used 10 unique datasets obtained from two open-source repositories - Github and Kaggle. In their research, they found that features such as location are essential for the financial performance of a start-up, as they determine the proximity to potential investors, thus suggesting that, if a start-up is at a prominent location with proper communication and resources, or if it is located at a business centre closer to investors, the probability of investments are higher. Moreover, they found that features, such as categories, broaden the start-ups perspective by including diversified products that can attract potential investors. While a start-up with machine learning technology domain is expected to receive higher investment from investors, it is possible to consider the possibilities of how trends might affect fund raising chances, considering that the report was published in 2022, a period where machine learning and AI innovations are a trend amongst investors and across the global investment landscape, thus bringing to our attention how chances of possible outcomes could be affected by current industry trends.

(Bernstein et al. 2016) show that VCs who are connected to their start-ups through close physical proximity, for instance, through direct flights, could tightly monitor their investments, which leads to better outcomes further supporting the hypothesis suggested by Ajit et al (2020).

While these features should be considered, to the best of our knowledge besides having a robust data set. Identifying the variables to consider is equally as important as the machine learning algorithm to be trained. Some of the most used models are logistic regression, k-nearest neighbour, decision tree, random forests and XGBoost.

Although using a different model, Sharchilev et al. 2018 were able to achieve accuracy in predicting follow on funding using Crunchbase data, as well. J. Arroyo et. al (2019) tried to predict a start-up status by splitting up into two categories, such as start-ups that had funding events and start-ups that did not, using a support vector machine.

Overall, when compared to previous literature, this dissertation and research work contributes to the academic body of work by not only using a much more evolved database and datasets but also considering the limitations and missing factors from various research in the past. Most importantly we find that the definition of start-up outcomes/success differs per scholar, and we find that by using an ensemble method, several common errors encountered in previous literatures can be reduced and the possibility for ease of integration for real life applications can be improved. With the buzz around machine learning and algorithms and how it can contribute to the start-up ecosystem, are there any practical implementation of these models by any organizations or start-ups across the world currently or are these theoretical concepts with little to no practical use cases?

To answer this question, here are some examples of how machine learning models and algorithms are being used practically by some organizations & start-ups across the world:

1. **Netflix** – This is a great example of how machine learning was used. The company originally launched in 2007 uses data from its customer viewing history amongst other sources to personalize users experience and increase screen time and revenue by using data to construct a universe of the user profile index – Allen Yu, 2019. One of the most prominent use and application of machine learning is with the Skip Intro button instantly converting a TV experience ever more like an app – The Atlantic, 2017
2. **ICU Intervene** – Created by the MIT's Computer Science & Artificial Intelligence Laboratory (CSAIL), the ICU intervene uses machine learning to predict possible treatments by learning from massive amounts of intensive care unit data to make real time decisions with reasons behind the decisions it makes ultimately improving patient care and a huge difference in the quality of care received – [MIT News, 2017]

While these are some practical examples of application of machine learning in the entertainment & healthcare industry respectively, below we look at some practical applications of these algorithms in the financial sector around which this dissertation is most relevant.

3. **Paypal** – Founded in 1998, Paypal uses machine learning to prevent fraud. With access to data on more than 350 million consumers and merchants in over 200 markets, Paypal trains complex and intelligent models to help spot suspicious online behaviour that human eyes may miss. The company uses machine learning to help access in real time if an individual is a legitimate customer or not thus further advancing its cyber security measures which is vital for such business model – Paypal, 2021

A much similar practical application of machine learning model closer to the scope of this dissertation is with start-up investing and decision making as recently implemented by Pitchbook across its digital platform thus making this organization the last but not the least example on our list.

4. **Pitchbook:** Quite like Crunchbase – the data provider used for this dissertation, Pitchbook founded in 2006 is a company that provides comprehensive data, research and insights spanning the global capital markets that serves investors and business owners across the world. The company collects and analyses information on the entire investment and business lifecycle. According to McKinley McGrinn – product manager of market intelligence at pitchbook, the new Pitchbook AI tool called The VC Exit Predictor was developed using a proprietary machine learning algorithm however to ensure accuracy, predictions are made for venture backed companies that received at least two rounds of venture financing deals – TechCrunch 2023.

While no predictive tool is perfect, the application of machine learning models and artificial intelligence in contributing to a better decision making for humans alongside its possible impact on economic growth is no longer deniable in fact in a survey of over 1,000 businesses across nine sectors, 86% of them plan to implement AI as a mainstream technology. The ability of machine learning to boost efficiency, productivity for businesses is expected to add 415, trillion to the world economy – PWC 2022.

In the upcoming chapter, we discuss a comprehensive exploration of our research methodology and data framework. We will cover data sources, their inter-relationships, and the methodology employed to execute this project. This chapter not only serves as an introduction to our methodology but also underscores the potential contributions to the academic body of knowledge and practical applications, aiming to catalyse economic growth through our research endeavours.

Chapter 3: Data & Methodology

Data

For this research, a comprehensive and reliable source of business and industry data would be required which led us to reaching out to Crunchbase – a popular online platform that provides data on start-ups, businesses, and key individuals. The Crunchbase platform is a renowned and reputable industry portal that contains key datasets including but not limited to company details, funding rounds, leadership profiles amongst several other valuable resources for investors, entrepreneurs, researchers, and users. To access the data used for this project, a formal application for database access was submitted and follow up e-mail was sent to the support team. A positive response was received from the Crunchbase team thus marking the beginning of the data collection process.

The data used for this research and subsequent experimentation was extracted on August 7, 2023. This date serves as a data snapshot and represents a specific moment in the ever-changing business landscape. In this section, further details on the data will be provided as well as details on the nature, structure of the data. An entity relationship diagram that visually represents the web of relationships between the various entities contained in the extracted data is also illustrated herein. In the next section under methodology, explanations on how the data was prepared and transformed into a usable format for the machine learning model development will be provided alongside details on the methods & strategies implemented.

Following the collection of the data via the api key provided by the Crunchbase team, an explorative data analysis was carried out on the datasets to gain further insights and understanding of the data, some of the outcomes are further discussed in the results section contained in this dissertation. Also, during the exploratory data analysis, it was observed that the datasets provided were made up of several different csv files that could be joined together by unique identifiers. This unique identifier served as the central data for merging the different datasets into one. Below is a simplified ERD diagram illustrating the inter-relationship between the different datasets obtained from Crunchbase:

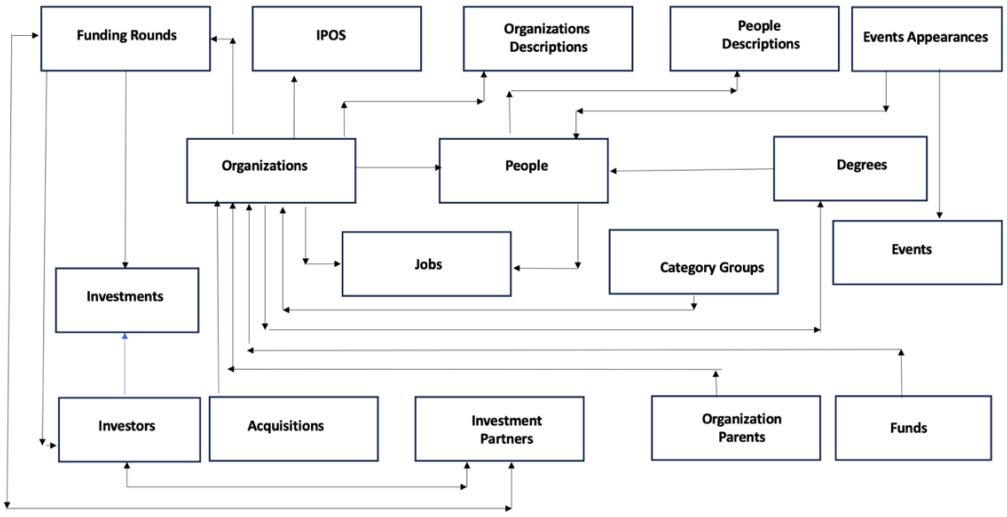

Fig 1. Entity-Relationship Diagram

From the data, we found that organizations.csv serves as the main central hub for everything else. All other files contain one data column or the other that can be traced back to the organization data and vice-versa. It is this interconnectedness via unique identifiers present in all datasets and across the structure of the data that makes it possible to be able to merge them all together as one – considering that CrunchBase is a platform that is built around organizations, their activities and data points connected/related to them.

Let's explore and discuss the datasets contained herein and as used for the research while in the next section we cover how it was used.

As mentioned earlier, organization data (*organizations.csv*) being the central point of all other data provides detailed profiles of companies from their legal entities to funding trajectories. IPOs (*ipos.csv*) delve into the realm of Initial Public Offerings and provide insights on stock prices, valuations, and the transition to public ownership. Data related to investments (*investments.csv*), investors (*investors.csv*), funding rounds (*funding_rounds.csv*) and funds (*funds.csv*) can be found in the respective dataset as named. There is a dataset on People (*people.csv*) that contains insights on individuals and their relationship to organization including companies, investment partners and academic records which could be found under the degree's dataset (*degrees.csv*) and their respective jobs and positions which could be found in the job's dataset (*jobs.csv*) as well as how they are associated with these companies.

Another vital dataset is funds and funding rounds, they provide insights into funding events connected to organizations and the individual investors and their respective participations and impacts. Acquisitions are described in acquisitions (*acquisitions.csv*) while Events (*events.csv*) showcases corporate events as respectively linked to different entities. Another dataset that is unique is the Category groups (*category_groups.csv*) as this dataset serves as a classifier grouping organizations into categories while datasets like Organization Descriptions (*organization_descriptions.csv*), Organization Parents (*org_parents.csv*), Event Appearances (*event_appearances.csv*), and Investment Partners (*investment_partners.csv*) provide additional details and contextual information about organizations, their parent relationships, event appearances and investment partnerships respectively.

The ERD Diagram serves to visually illustrate this web of inter-relationships and dependencies across these datasets towards enriching our understanding of the data.

In the next section we shall explore methodologies employed during this project and how this data was worked upon during its implementation and use for the training of our machine learning models. Next, we also cover some ethical implications and some possible limitations/assumptions before going further to discuss the results and outcomes.

Methodology

Dataset Creation & Target Variables

In this section our objectives are divided into multiple folds. First and foremost is the cleaning and preparation of the data for the machine models. We shall also be covering the different models employed and how the experiment was set up.

For this research, two major environments were utilized which are the R-Studio & The Orange Data Mining Studio. We worked with R-Studio using R-language for data cleaning and pre-processing activities alongside with handling missing values, data labelling and the merging of the appropriate datasets into one dataset which was then used in Orange Studio for the machine learning model development and other aspects of this project.

The use of R-Studio and r-language for data science and machine learning related projects has been highly applauded by academic scholars in the past. In a research work by Hafiz et al. 2020, a comparative study of data science tools was carried out, R-language was described as one of the most important tools for performing calculations, different types of analysis, handling big datasets and calculation of clusters. It was also chosen for this project because it is simple, easily readable, and understandable for users.

The image below illustrates the various variables used for the training of the machines and how the different datasets used were merged to create one dataset.

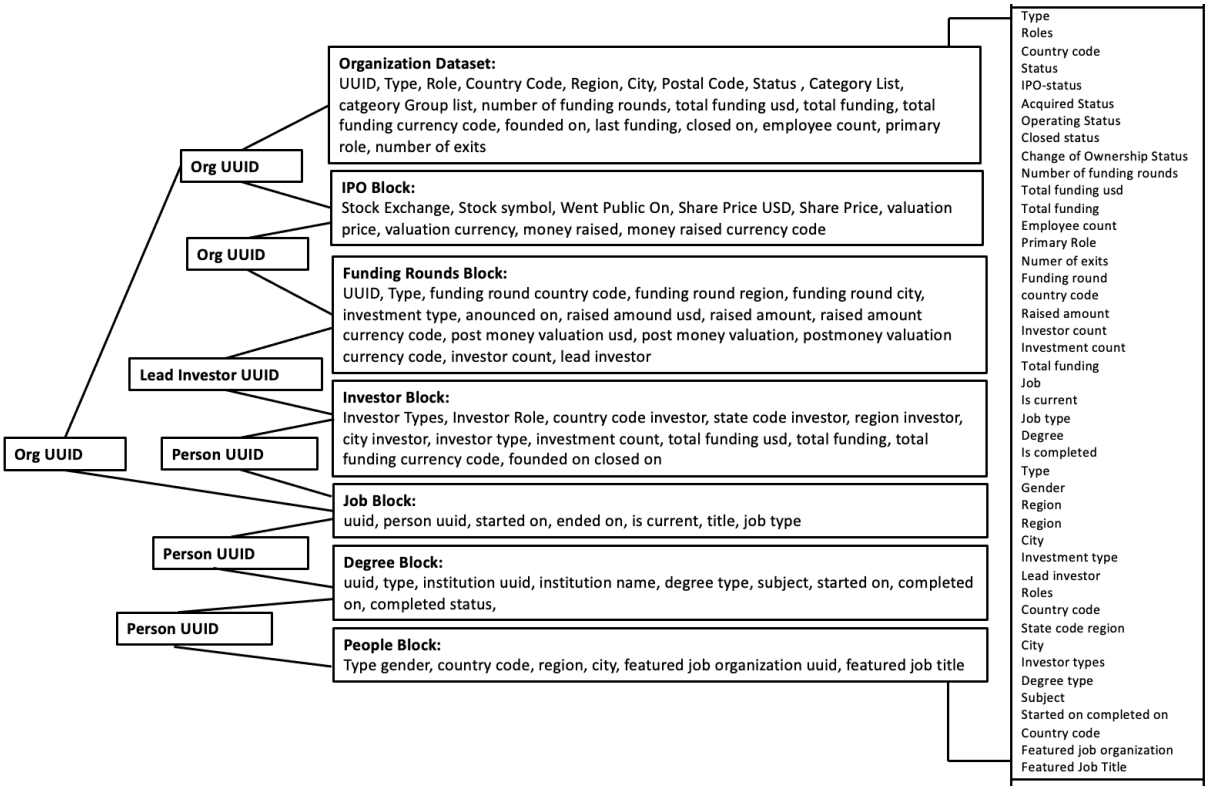

Fig 2. Variable Selection & Mapping

For this experiment, there are 7 primary datasets merged to create the final data used for training the models. The Organization dataset (*organization.csv*), IPO (*ipos.csv*), Funding Rounds (*funding_rounds.csv*), Investor (*investors.csv*), Jobs (*jobs.csv*), Degree (*degrees.csv*), and People (*people.csv*). The reasons for these criteria and selections are because the objective of this project is not only to predict the probability of possible IPO outcomes but also explore how certain criteria such as funding rounds, investor relationship, job titles, academic/degree could play a role in possible outcome of a start-up journey. The organization dataset was merged with IPO dataset via organization uuid. Some of the vital metrics and columns contained in the organization datasets include geographical data such as regions and cities, some financial data such as total funding and some operational details like total

number of employees were also included. However, this will not be enough for us to train a model if our objective is to predict whether investors will get to exit via IPO or not. As such we had to merge with the IPO dataset which contains IPO related data such as stock exchange the organizations are listed on, the date of going public, money raised amongst several other variables. The funding rounds dataset gives us more insights and data on the different funding rounds, amount raised and investment types such as pre-seed, seed, angel investments, series a, series b, equity crowdfunding, amongst several others. This dataset was merged with the previously existing one using the organizations UUID thus ensuring the funding records of each organization was correctly populated on the new dataset.

While the IPO dataset provides insights on successful exits via IPOs, the funding rounds data set provides details and insights on funding activities of each organizations including the lead investors. With the organization UUID and lead investor UUID this dataset was merged with the investors dataset which provided data on investor types, investment counts and total funding. This dataset also provides geographical data on the investors such as the region, city and state code which is quite a vital variable in training the machine models in the event where there is any relationship between the location of the organization and the geographical location of the investor as contained in the academic publication by 2016, Bernstein et al.

Next is the addition of the job dataset which was merged via the person UUID to provide occupational data on all the individuals currently connected to the organizations in the data set followed by the Degrees dataset which provides educational data and finally the people dataset providing mostly geographical data on the individuals connected to the dataset. At the end of the merging of these datasets we had a total of over 3,014,538 records and about 316 columns and variables.

The final dataset used for the training of the models contained only organizations located in the United Kingdom and created within 1999 - 2020, spanning two decades. At the end of filtering the dataset based on these criteria, the final dataset contained 415,530 observations and over 90 variables with approximately 45 total variables used for model training process.

As part of the feature engineering procedure, On the final dataset, a new column/variable was created that uses a binary classification to determine the IPO status of a company as either a "***yes***" or a "***no***". The positive class represents companies that successfully undertake an IPO while the negative class represents the companies that did not. This binary classification is a fundamental part of this experiment as it would not only serve as the target variable for predictions by the machine learning model but also for explanations too. The interpretation of the model outcome and explain prediction models used for this experiment will be carried out using the SHAPley Additive Explanations. SHAP explains the predictions for an instance x by computing the contribution of each feature to the prediction. The explanations are done by computing the SHAPley values from coalitional game theory – (Shapley, L 1953).

The explanation of the model predictions using SHAP is further discussed in subsequent sections in this dissertation.

Ethical Implications:

During the set up and design of the machine learning model, a comprehensive feature engineering strategy was implemented that included an intentional removal of certain columns and variables as part of the process of selecting features that would be included in the final dataset. Datasets that were intentionally excluded include *names, urls, permalinks, social media urls* and *crunchbase ranking* metrics.

The deliberate exclusion of personally identifiable information such as names embodies the principles of privacy-by-design, a core principle of data ethics. This approach not only upholds robust privacy standards but also proactively mitigates potential algorithmic bias thus promoting a model that prioritizes fairness and equity. Simultaneously, the removal of URLS, permalinks and social media links aligns with established data security and privacy standards to guard against inadvertent data leakage – Gabriel et al 2007.

The decision to exclude the CrunchBase ranking columns in rooted in the effort to ensure model transparency and impartiality. This strategy further aligns with the principle of fairness and transparency in machine learning. By eliminating external ranking influences, the model's prediction is predicated solely on intrinsic data characteristics thus enhancing model fairness and interpretability.

UUIDs within the dataset were incorporate to signify commitment to maintaining data integrity and to foster data linkage in the datasets. These are very vital components of data science best practices as it serves as both a pivotal tool in thwarting data duplication as well as streamlining data integration across the diverse datasets used in the training of the models thus ensuring data quality, consistency, robustness of the predictive models which further enhances its credibility.

Limitations & Assumptions

A subset of dataset was excluded from the training of the models due to limited time available for the experimentation and computational resources required to handle the substantial size of the final dataset. Specifically, the following datasets were omitted from the model training process: *Organization Descriptions, People Descriptions, Events Appearances, Category Groups, Events, Organization Parents, Investment Partners, Funds, and Investments.*

Datasets such as *Organization Descriptions, People Descriptions, Events Appearances, Category Groups, Events, Organization Parents,* predominantly consists of textual data rendering them less suitable for inclusion in this experiment and for the objective we are trying to achieve. Additionally, datasets like *Investment Partners, Funds, Investments and Acquisitions* were excluded because they contain information that would not have been available at the outset of an organization's journey, potentially introducing a look-ahead bias which often occurs when the trained model is exposed to information from the future which it would not have had in a real-world scenario. – Anand and Prakash, 2019

Also, to ensure the models adherence to real-world conditions and for further avoidance of incorporating future information, columns/variables from subsequent data were removed for example from the IPO dataset, Funding rounds dataset and Investor dataset, columns/variables such as Stock exchange symbol, stock price, post money valuation amongst several other future information were removed from the dataset used to train the models.

It is important to note that this study is also limited to start ups within the United Kingdom, indicated by the country code GBR as such might not be applicable or account for start-ups located in other jurisdiction or take into consideration any possible impact those data might have on the results assuming there is a cross economic impact or relationship between the data sets say as collective global ecosystem.

Furthermore, an underlying assumption of this research is that all organizations in the final dataset either began their journey as start-ups or align with the characteristics of start-up entities and with the assumption that it takes at least 3 years before a start-up can be fully evolved enough or ready to be acquired/consider an IPO. Thus, dataset from the last 3 years from the date of this publication were also not included in the training of the models.

After merging the data sets as one, the next steps which was a pre-requisite/first step to the experiment set up to be described next was data pre-processing. As mentioned in previous sections, one of the activities carried out was to filter out organizations located in the United Kingdom via the country column with the United Kingdom represented as GBR.

Experiment Set up & Architecture.

Fig 3 illustrates the architecture behind the set-up of the experiment.

First and foremost, it is important to know that this experiment was carried out using the Orange Data Mining toolkit developed by the University of Ljubljana, Slovenia in 1996 with a stable release updated on 5th, March 2023. Version 3.35.0 was the version used for this experiment. The Orange toolkit features a visual programming front-end for explorative data analysis and interactive data visualization.

Some of the reasons why the orange data mining toolkit was chosen for this experiment is because much as it is a component based visual programming software package for machine learning and data sciences, it also allows to write custom scripts in python which builds upon C++ implementations of computationally intensive tasks making it suitable for both experienced users and programmers and well as students of data mining – Demsar, Curk, Erjavec et al (2013).

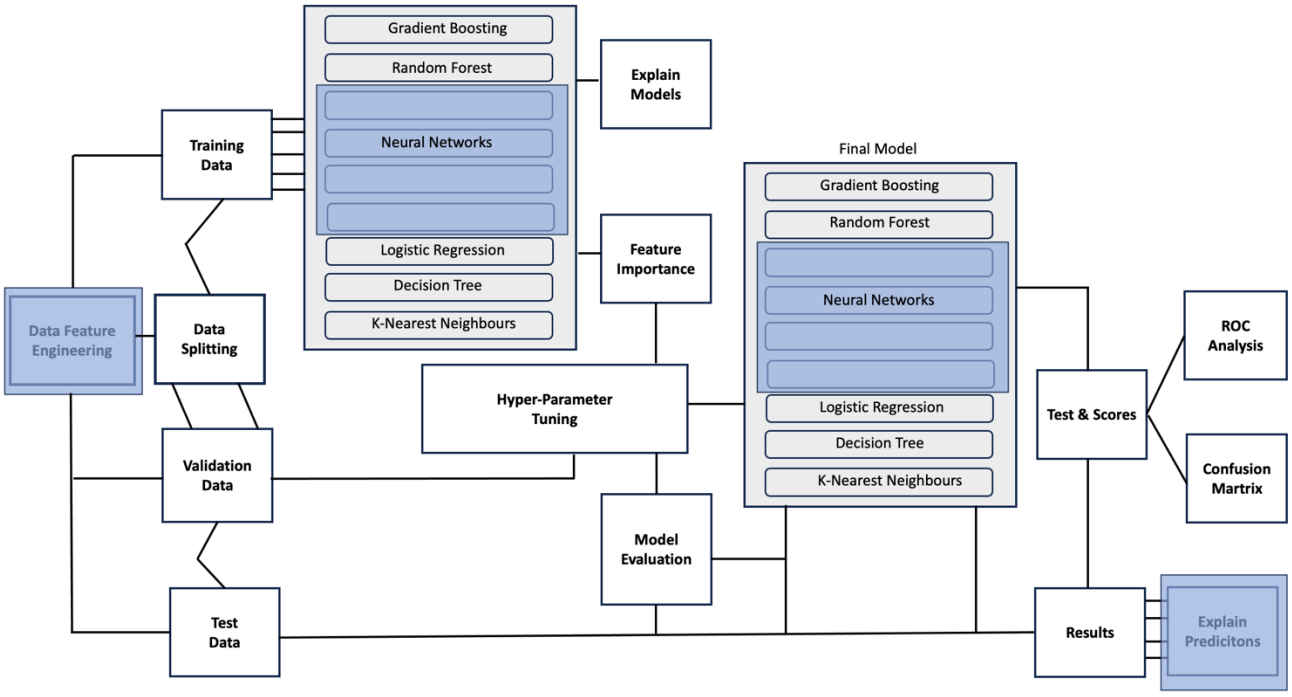

Fig 3. Experiment Set Up

In the workflow implemented for this experiment, we employ six different models: Gradient Boosting, Random Forest, Neural Networks, Logistic Regression, Decision Tree & K-Nearest Neighbours. The workflow begins with further cleaning, pre-processing of the data, data feature engineering and ends with the explanation of the results/predictions.

In this section, we shall cover an overview of the experimental configuration, the methodology employed for this study and a discussion on the 5 different phases with reference to the illustration in Fig 3, exploring both the technical and non-technical aspects of the set-up. The five phases as below are further discussed herein:

1. Data Feature Engineering and Splitting
2. Model Selection
3. Hyper-Parameter Tuning
4. Model Evaluation, ROC Analysis & Confusion Matrix
5. Predictions, Explanations & Interpretations

Data Feature Engineering & Splitting:

This step involves the transforming of the raw dataset into a format suitable for the machine learning models to work effectively. This process is simplified in orange data studio because it offers a visual and intuitive option to categorize the type and role for each column within the dataset. Two options are provided during the loading of the dataset which are roles and types thus defining how the data should be treated. The data types allow to define if the data in the column are categorical data, numeric data, date/time, or text data thus specifying the nature of the data contained in the column. Specifying the correct type is essential as it

impacts tasks such as encoding, scaling and model building. While roles specify how these columns should be used during data analysis and machine learning task. Orange uses types and roles to guide the software in understanding the purpose and nature of each column. There are four primary roles in the orange data studio which are *Feature, Target, Meta, and Skip* – (Orange Documentation 2015).

Fig. 4 below illustrates how the data sets used in this experiment are encoded for the use on the training of the machine models by Orange Data studio.

Subsequently, the datasets are divided into 3 using a 60-20-20 ratio with 60% of data to be used for training the model, 20% representing the validation dataset used for hyperparameter tuning while the remaining 20% is the testing dataset to assess the overall model. Improper splitting could lead to overfitting or underfitting thus harming the model performance. Some of the other major reasons for splitting the data at 60-20-20 is to avoid model selection bias, to optimize hyper parameters on the development test and this choice was made based on the size of the data ranging between 100 – 1,000,000 – (Muraina, Ismail 2022)

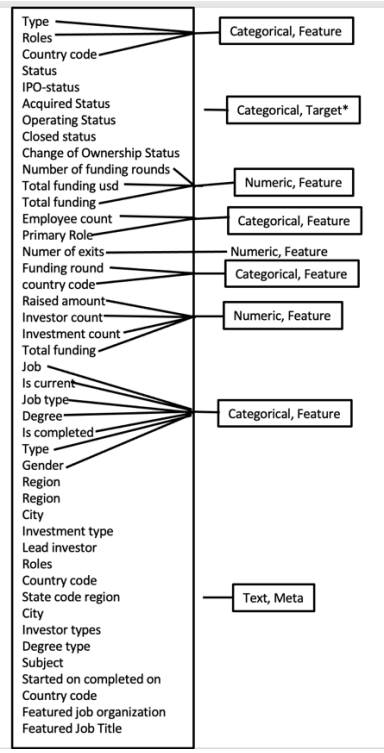

Fig 4 Data Feature Engineering

Model Selection:

For this experiment, 6 different models were trained – Gradient Boosting, Random Forest, Neural Network, Logistic Regression, Decision Tree, and K-Nearest Neighbours. Logistic Regression, Decision Trees and Random Forests are wildly recognized algorithms that have been used extensively in many prior studies involving Crunchbase data (Ross et al 2021, Krishna et al 2016).

Logistic Regression stands out as one of the simplest and quickest to train algorithm and often serving as an initial source for tackling various classification tasks. The primary strength of the logistic regression model lies in its low variance rendering it less susceptible to overfitting especially when dealing with binary classification problems that exhibit a distinct separation between classes. However as discussed by (Bento 2017), Logistic regression can be vulnerable to overfitting when trained on datasets with numerous correlated features. It measures the relationship between a categorically dependent variable and independent variables by using probability scores as predicted values of the dependent variable and supports problems with high dimension data.

Decision trees, random forests and gradient boosting possess a feature selection mechanism that makes them adept and suitable for handling datasets with numerous variables and varying degrees of importance. Secondly, these classifiers are considered white box - meaning their decisions can be interpreted thus facilitating the measurement of feature relevance in the classification process. The interpretability is particularly variable considering our objectives as part of this experiment is to also identify factors contributing to success so while Logistic regression as mentioned earlier models the probability of a binary outcome and is known for its simplicity, Decision trees offer a rule based approach to classification making them useful for explaining the decision making process while Random Forests on the other hand harnesses the power of ensemble learning by combining multiple decision trees to enhance predictive accuracy and resilience against overfitting – J. Arroyo et al. 2019

Hyperparameter Tuning:

For this dissertation, recall the experiment set up and model training was carried out using Orange Data Mining studio where each different parts of the architectures are connected using widgets. These widgets however come with default settings often not tailored to the specific type of dataset we are working with thus making it necessary to fine-tune our models to enhance performance. The process of automating this optimization is called hyperparameter tuning for example k in nearest neighbours and the learning rates for neural networks – Chang (2020). For this experiment, the models we used are supervised learning models. In supervised learning we have available both the input represented as x and the corresponding output or target denoted y. The primary objective is to find the best possible predictive model function often represented as f. Basically the model that minimizes the specific cost function that quantifies the disparity between the predicted outputs generated by our model and the actual ground-truth labels. The choice of the predictive model function f is contingent on the models structure and its hyperparameter settings and depending on these factors we'd have a finite set of model architectures which would restrict our model function f to belong to a predefined set of functions denoted as F thus to find the optimal predictive model involves finetuning of the model parameters and hyper parameters in order to enhance its predictive accuracy – Li Yang (2020). The configuration used for the different models in this experiment are further described in this section below while fig 4 illustrates the mathematical formula used to obtain the optimal predictive model f*:

$$f^* = \arg\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}(f(x_i), y_i)$$

Fig 4 Formula for Optimal Predictive Model

Where n is number of training data points, xi is the feature vector of the i-th instance, yi is the corresponding actual output and L is the cost function value of each sample.

It is important to take note that during this research we found that in the process of hyperparameter tuning in supervised learning models, different loss functions are used across the various algorithms including squared Euclidean distances, cross entropy, and information gain. Conversely, the different algorithms create distinct predictive model structure via specific hyper parameter configurations – Gambella (2019). Some of the different configurations that contributed to the hyper parameter tuning process for this experiment are described below for each model:

- Model: Extreme Gradient Boosting (XGBoost) – Number of Trees = 100, Learning Rate = 0.3, Regularization Lambda = 1. For growth Control, Limit Depth of Individual Trees = 6
- Model: Random Forest: Number of Trees = 10, Subset Not Allowed to Split smaller than 5
- Model: Neural Networks. Here we had 4 different neural networks (See Fig 3) For Neural Network 1: Neurons in hidden layers = 10, Activation = ReLu, Solver = Adam, Regularization Alpha = 0.001, Maximum Number of Iterations = 1,000. For neural Network 2: Neurons in Hidden Layer = 5, Activation = Identity, Solver = Adam, Regularization Alpha = 0.001, Maximum Number of Iterations = 1,000. For Neural Network 3: Neurons in Hidden Layers = 5, Activation=TanH, Solver = SGD, Regularization Alpha = 0.01 and Maximum Number of Iterations = 1,000. For Neural Network 4: Neurons in Hidden Layers = 5, Activation = Identity, Solver = LBFGS-B, Regularization Alpha = 0.001 and Maximum Number of Iteration = 200
- Logistic Regression = Regularization Type = Ridge L2, Strentgth = C1
- KNN – Number of Neighbours = 3, Metric = Euclidean, Weight = Uniform
- Tree = Induce Binary Tree, Min Number of Instances in Leaves = 3, Do Not Split Subset Smaller than 5, Limit the Maximal Tree Depth to 100

Model Evaluation, ROC Analysis & Confusion Matrix:

ROC Analysis provides a powerful alternative to additional traditional model performance assessment using confusion Matrix. During this research work, after hyper parameter tuning and a final model has been trained, ROC graphs were used to organize the classifiers and visualize their performance. ROC which stands for Receiver Operating Characteristics is a graphical representation that helps us understand how well a binary classification model distinguishes between true positives and false positives and false positives under different thresholds. They are very useful tool for evaluating classifiers – Fawcett (2005) The confusion matrix is also a vital part of model evaluation as it provides a comprehensive break down of predictions to categories such as true positives, true negatives, false positives and false negatives which are crucial metrics required for the calculation of other performance indicators like the precision, recall, F1 Score and accuracy. Both the ROC Analysis and Confusion Matrix offer a detailed perspective on how well our models are performing –

Ghani et al (2015). In Orange Data Studio both ROC Analysis and Confusion Matrix usually gets the evaluation results from the test and scores widgets (See Fig 3). In the results section of this dissertation, we shall further cover how these model evaluation techniques were used.

Predictions, Explanations & Interpretations:

The final phase of this experiment which makes up for the second core major objective of this dissertation is the expandability of the predictions. As illustrated in Fig 3, After the validation set is employed to fine-tune hyperparameters and assess model performance, the test set serves as an unbiased evaluation of the final models' predictive capabilities. The test dataset contains only data that would have been available at the start of the analysis and by avoiding the use of future information, look ahead bias could be prevented. To quantify the contribution of each feature to the predictions we used SHAP which is based on Shapley Values. It is easy to confuse one for the other or to mix them both – However while Shapley values deal with fair distribution of values among players in a cooperative game, SHAP values are a technique for explaining the contributions of individual features – Christoph (2022). For this research, we had used SHAP (Shapley Additive exPlanations) and this explains the prediction of an instance x by computing the contribution of each feature to the prediction as specified in the formula below:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

Where g is the explanation model and $z' \in \{0,1\}^M$ is the coalition vector, M is the maximum coalition size and $\phi_j \in R$ is the feature attribution for a feature j, the Shapley values – Christoph Molnar (2022).

Fig 5 Formula for SHAP Explanation

For this experiment, the explain prediction widget was connected to the result and shows what features affect the predictions – Orange (2015). However, SHAP cannot replace performance metrics. The mean squared error, mean absolute error, F1 Score and the areas under the ROC curves amongst other performance metrics evaluate a models' goodness of fit while SHAP on the other hand yields explanation for a model's prediction. Explanations solely tell why a certain prediction was made and do not necessarily imply causality - Winter (2002). In the next section we shall look at the results and outcome of the experiment.

Results

Fig 6 and Fig 7 shows the outcome of the machine models. While Fig 6 contains the results from the ROC analysis, Fig 7 contains the model evaluation results. The ROC curve is a plot of the true positive rate against the false positive rate at various threshold settings. A classifier is considered more accurate when its curve closely traces the left-had top borders of the ROC Space. The x axis represents the False Positive Rate (FPR) which is a proportion of actual negatives that the model incorrectly predicts as positives while the Y-axis represents the true positive rate or sensitivity which is the proportion of actual positives that the model incorrectly identifies. In the next paragraph we shall look at the outcome of the model evaluation results as well as the explanation of the predictions (See Fig 8)

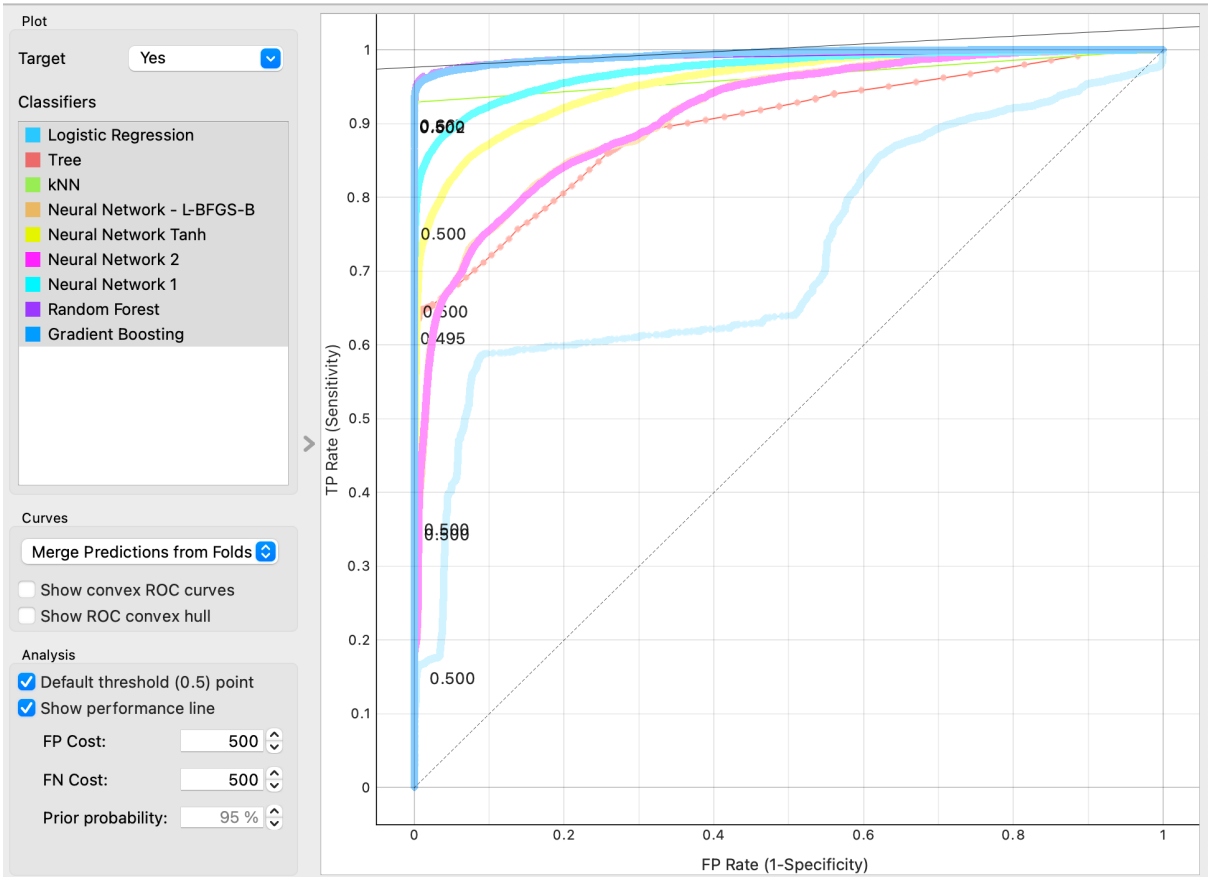

Fig 6 ROC Analysis Curve

Based on the model evaluation results as illustrated in Fig 7 The models trained for this experiment – Gradient Boosting, Random Forests, Logistic Regression, Neural Networks, KNN and Decision Tree were expected to predict the IPO status of a company as either Yes or No Outcomes. From the results, Gradient Boosting performed best compared to the others while Logistic Regression had the least performance. Gradient boosting achieved high values for its metrics – AUC (0.993), Correct Classification Rate (0.996), F1 Score (0.954), Precision (0.996), Recall (0.916) and Matthews Correlation Coefficient (0.953). On the other hand, Logistic regression metrics reported AUC (0.725), Correct Classification Rate (0.945),

F1 Score (0.236), Precision (0.395), Recall (0.168) and Matthews Correlation Coefficient (0.233). The other models including Random Forest, Neural Networks, Decision Tree & K-Nearest Neighbours achieved performance levels between the two extremes however, Neural Networks had varying performances depending on the specific parameters configured.

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.725	0.945	0.236	0.395	0.168	0.233
Tree	0.896	0.980	0.761	0.964	0.629	0.770
kNN	0.964	0.995	0.947	0.979	0.918	0.945
Neural Network - L-BFGS-B	0.912	0.962	0.498	0.762	0.370	0.515
Neural Network Tanh	0.955	0.979	0.760	0.886	0.665	0.758
Neural Network 2	0.911	0.962	0.490	0.753	0.363	0.507
Neural Network 1	0.973	0.987	0.853	0.955	0.771	0.852
Random Forest	0.990	0.995	0.952	0.994	0.914	0.951
Gradient Boosting	0.993	0.996	0.954	0.996	0.916	0.953

Fig 7 Model Evaluation Metrics

While Gradient boosting excelled in minimizing both False positives and False Negatives, a critical feat in binary classification scenarios, it achieved an impressive 91.6% recall rate indicating its ability to effectively identify positive cases. Random Forest and k-Nearest Neighbours had recall rates of 91.4% and 91.8% respectively alongside Precision scores of 99.4% and 97.9% respectively. The recall rate for the logistic regression model was about 16.8% suggesting a clear inability in identifying true positive instances and a precision of 39.5% which reflects a relatively high rate of false positive predictions. Neural networks in this experiment set up showcase the impact of architecture selection on model performance as they all record varying range of metrics. These results illustrate the varying strengths and weaknesses of each model in handling classification task.

Using the SHAP library to explain the best performing model - Gradient Boosting as illustrated in Fig 8a, recall that SHAP value is a measure of how much each feature affects the model output. A higher SHAP value means the feature has a higher impact on the prediction for the selected class. The positive SHAP values pointing right from the centre are feature values with impact towards the prediction for the selected class while the negative values pointing towards left have an impact against the classification – Orange Explain Model Documentation. From the outcome of the SHAP values for the gradient boosting model we found that features such as Total Funding, Raised Amount had high impact on the performance of the model. Also, Job Type Executive has a high impact on the model however it tends to affect in both ways quite similar to the gender which has a moderate to low impact but also affects almost equally in both ways. This exploration of feature contribution through SHAP values enriches our understanding of the models underlying mechanism thus producing actionable insights for data-driven decision making.

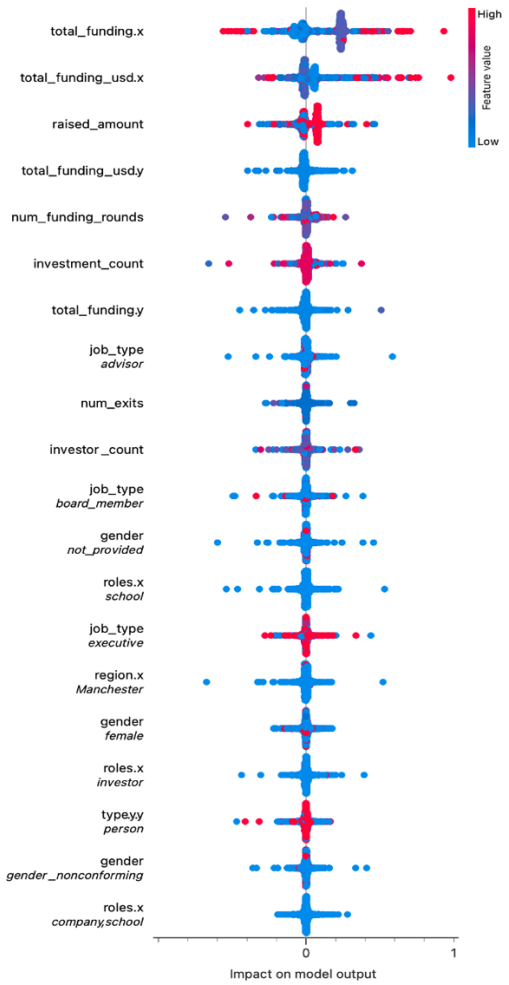

Fig 8 – Model Explanation (Left)

Discussion and Conclusions

While all the models have their unique strengths and weaknesses, from the evaluation of all the models, a common trend identified is that they all have a higher CA (Correct Accuracy) score than F1 Score indicating that they perform relatively better in overall accuracy than in balancing precision and recall. Also, MCC (Matthews Correlation Coefficient) consistently outperforms Recall in these models thus implying their ability to classify both positive and negative cases. Considering the significant impact of specific features revealed through SHAP values, one recommendation for improving performance will be to consider exploring advanced techniques for both feature selection and engineering and of course a high computing power and more advanced algorithm development/architecture. While this experiment reveals Gradient Boosting as the best performing model and logistic regression as the least performing model, future research endeavours could consider a deeper investigation into these aspects alongside with a continuous exploration of machine learning techniques as the industry evolves. Also, important to take note are the limitations and hypothesis assumed during this experiment and most importantly the fact that while no model is extremely

perfect, the objective is to work with data and carry out research work and analysis to at least find which could be the closest. The Crunchbase dataset is an incredibly powerful data resource that could be totally invaluable to fund investors and various stakeholders operating within this ecosystem. The data offers opportunities for gaining insights, developing predictive models, and exploring segmentation which could contribute significantly to this vibrant ecosystem.

Future Works:

For future research and contribution to the existing academic body of knowledge, there are two major propositions to be considered. Firstly, integrating a time series analysis into the existing framework would be highly valuable as this current research does not account for temporal trends which are crucial in the understanding of the ever evolving start up landscape. For instance, the rise of AI start-ups might become prominent trend amongst investors thus leading to a surge in such investments and start-ups but by incorporating time series, these transient patterns could be captured to further improve predictive accuracy.

Another intriguing area to explore would be the impact of behavioural factors on start-ups success and how news headlines, press releases and social media discussions can significantly impact a company's reputation thus by leveraging natural language processing techniques to analyse textual data, more light could be shed on how behavioural factors such as public sentiments could impact a start-ups outcome. This is currently not considered in this research.

In conclusion, this dissertation has endeavoured to explore and contribute to the dynamic intersections of machine learning, start-up funding, predictive analysis, model training and explainability through the examination of various algorithms discussed herein. We have sought to provide insights that illuminate the intricacies of start-up funding and possible IPO status/outcomes.

As the academic documentation of this research work comes to an end, it is sincerely hoped that the contents of this publication prove both valuable and useful to fellow researchers, practitioners, and scholars in the field as it is my aspiration that this work contributes meaningfully to the academic body of knowledge, standing the test of time.

Looking ahead, as mentioned earlier, it is recommended that future research endeavours delve into the evolving landscape of technology by considering the potential implementation of cloud computing and more sophisticated computing environments. The integration of higher computing power coupled with advanced resources has the potential to propel research in these areas towards a more robust and impactful outcome further fostering continued growth and understanding.

Bibliography

- International Trade Administration (2022) *United Kingdom Artificial Intelligence Market.* Available: https://www.trade.gov/market-intelligence/united-kingdom-artificial-intelligence-market-0#:~:text=The%20UK's%20AI%20market%20is,the%20UK%20economy%20by%202035 [Accessed: 27 July 2023].

- Tech Nation (2022) *UK Tech Ecosystem Update: London Tech Week 2022.* Available: https://technation.io/uk-tech-ecosystem-update-2022/ [Accessed: 27 July 2023].

- Companies House (2022) *Companies register activities: 2021 to 2022.* Available: https://www.gov.uk/government/statistics/companies-register-activities-statistical-release-2021-to-2022/companies-register-activities-2021-to-2022#:~:text=During%202021%20to%202022%2C%20there,was%20in%202020%20to%202021 [Accessed: 27 July 2023].

- Dealroom (2023) *UK tech update – London Tech Week 2023.* Available: https://dealroom.co/reports/uk-tech-update-london-tech-week-2023?utm_medium=email&_hsmi=262617897&_hsenc=p2ANqtz--P1-ZL6Gy1Y15rZuLfy4URIcUozGrZYqB6BXn7ME9mOH4XZnmtM9AUrPSHb44i0m-dR7UiasXrdnn3v3w7fU_9Mc6jEHs_JIWV2rX2hl4qvkb2UDQ&utm_content=262617897&utm_source=hs_email. [Accessed: 27 July 2023]

- Beauhurst Start up Fail, Scale & Exit Rates In The – 2022. Available: https://www.beauhurst.com/blog/startup-fail-scale-exit/ [Accessed: 29, June 2023]

- Lundberg, S. M., & Lee, S. I. (2018). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems. (pp. 4765-4774).

- Crunchbase Wikipedia, 2023 . Last Updated On 2 March 2023. Cunchase, Wikipedia. Available: https://en.wikipedia.org/wiki/Crunchbase. [Accessed: 19, July 2023].

- Ismail, 2022. Ideal Dataset Splitting Ratios In Machine Learning Algorithms: General Concerns For Data Scientist And Data Analytics. In Proceedings of the 7th International Mardin Artuklu Scientific Researches Conference. February 2022 at Mardin, Turkey. (pp. 496-497)

- (Hengstberger, 2019) - Increasing Venture Capital Investment Success Rates Through Machine Learning. Msc in Mathematics & Finance. Imperial College of London. Available: https://www.imperial.ac.uk/media/imperial-college/faculty-of-natural-sciences/department-of-mathematics/math-finance/HENGSTBERGER_THOMAS_01822754.pdf. [Addressed: 22, July, 2023]

- Ross et al 2021: Greg Ross, Sanjib Das, Daniel Sciro, Hussain Raza (2021) Capital VX: A machine learning model for startup selection and exit prediction. The Journal of Finance & Data Science. Volume 7, pp 94-114

- Dellermann et al 2017: Dominic Dellermann, Nikolaus Lipusch, Phillipp Ebel, Karl Poop (2021) Finding the Unicorn: Predicting early stage startup success through a hybrid intelligence method. Available: https://www.researchgate.net/publication/351449034_Finding_the_unicorn_Predicting_early_stage_startup_success_through_a_hybrid_intelligence_method. [Accessed On: July 22, 2023]

- 2020, Ang, Chia, and Saghafia: Yu Qian Ang, Andrew Chia & Souroush Sahgafian (2021) Using Machine Learning To Demystify Startups' Funding, Post Money Valuation, and Success. Part of the Springer Series in Supply Chain Management book series. Volume 11. pp 271-296

- Ajit et al. 2022 - Ajit Kumar Pasayat, Adway Mitra and Bhaskar Bhowmick (2022): Determination of essential features for predicting start-up success: An empirical approach using machine learning, Technology Analysis & Strategic Management. Available: https://doi.org/10.1080/09537325.2022.2116569. [Accessed: July 22, 2023]

- Bernstein et al. 2016 - Shai Bernstein, Xavier Giroud and Richard R. Townsend (2016) The I pact of Venture Capital Monitoring. The Journal of The American Finance Association. Vol. 71. Issue No. 4, pp. 1603-1620

- Sharchilev and Michael, 2018 - Boris Sharchilev, Michael Roister, Andrey Rumyantsev, Denis Ozornin, Pavel Serdyukov and Maarten de Rijke (2018) Web-based Startup Success Prediction In CIKM'18: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. October 2018. Pp 2283-2291.

- J. Arroyo et al. 2019 – Javier Arroyo, Francesco Cora, Guillermo Jimenez-Diaz and Juan A. Recio-Garcia (2019). Assessment of Machine learning Performance For Decision Support In Venture Capital Investments. Vol 7. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8821312 [Accessed: July 30, 2023]

- Allen Yu, 2019 – How Netflix Uses AI, Data Science and Machine Learning – From A Product Perspective. Available: https://becominghuman.ai/how-netflix-uses-ai-and-machine-learning-a087614630fe [Accessed August 16, 2023]

- The Atlantic – 2017 – Netflix's 'Skip Intro' Button Makes TV Ever More like An App. Available: https://www.theatlantic.com/technology/archive/2017/10/netflixs-skip-intro-button-makes-tv-ever-more-like-an-app/544427/ [Accessed: August 16, 2023]

- MIT News 2017 – Using Machine Learning to Improve Patient Care. Available: https://news.mit.edu/2017/using-machine-learning-improve-patient-care-0821 [accessed: August 14, 2023]

- Paypal 2021 – The Power of Data: How Paypal Leverages Machine Learning to Tackle Fraud. Available: https://www.paypal.com/us/brc/article/paypal-machine-learning-stop-fraud [Accessed: August 16th, 2023]

- PWC 2022 – PWC 2022 AI Business Survey. Available: https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-business-survey.html [Accessed: August 16th, 2023]

- Hafiz et al (2020) – Hafiz Burhan Ul Haq, Haroon Ur Rashid Kayani, Saba Khalil Toor, Sadia Zafar, Imran Khalid (2020) The Popular Tools of Data Sciences, Benefits, Challenges and Applications. International Journal of Computer Science and Network Security. Vol.20 No.5, May 2020. pp 64 — 72.

- Shapley, L. 1953. 17. A Value for n-Person Games. In: Kuhn, H. and Tucker, A. ed. *Contributions to the Theory of Games (AM-28), Volume II*. Princeton: Princeton University Press, pp. 307-318. Available: https://doi.org/10.1515/9781400881970-018 [Accessed: August 31, 2023]

- Gabriel et al 2007 - Gabriel Ghinita, Panagiotis Karrass, Panos Kalnis, Nikos Mamoulis (2007). Fast Data Anonymisation with Low Information Loss. 33rd International Conference on Very Large Databases, University of Vienna, Austria, September 23-27, 2007. Pp 758 - 769

- Anand and Prakash, 2019 - Anand Iyer and Aditya Prakash (2019). Chapter 10. Controlling Biases. In Igor Tulchinsky et al. Finding Alphas: A Quantitative Approach to Building Trading Strategies. 2nd Edition. Available: https://doi.org/10.1002/9781119571278.ch10 [Accessed: August 31, 2023]

- Demsar, Curk, Erjavec et al. (2013) – Janez Demsar, Tomaz Curk, Ales Erjavec, Crt Gorup, Tomaz Hocevar, Mitar Milutinovic, Martin Molina, Matija Polajnar, Marko Toplak, Anze Staric, Mita Stajdohar, Lan Umek, Lan Zagar, Jure Zbontar, Marinka Zitnik and Boaz Zupan (2013) Orange Data Mining Toolbox in Python. Journal of Machine Learning Research. 14. pp 2349 - 2353. Available: https://www.jmlr.org/papers/volume14/demsar13a/demsar13a.pdf [Accessed: September 2, 2023]

- Orange (2015) – Orange Documentation. Available: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/loading-your-data/index.html [Accessed: September 2, 2023]

- Muraina, Ismail 2022 - Ismail Olaniyi Muraina (2022) Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns For Data Scientist And Data Analytics. In Proceedings of the 7th International Mardin Artuklu Scientific Research Conference. February 2022 at Mardin, Turkey. (pp. 496-497)

- Krishna et al 2016 - Khrisna Amar, Agrawal Ankit, Choudhary Alok (2016) Predicting the Outcome of Start-ups: Less Failure, More Success. 16th International Conference on Data Mining Workshops. 2, July 2016. Volume 0, pp 798 - 805.

- Bento 2017 - Bento, F.R.D.S.R., 2017. *Predicting start-up success with machine learning* (Doctoral dissertation), NOVA Information Management School

- Chang (2020) – Anthony C. Chang (2020). Chapter 5 – Machine & Deep Learning. In: Intelligence Based Medicine. Artificial Intelligence & Human Cognition in Clinical Medicine & Healthcare – pp 67 – 140,

- Li Yang, Abdallah Shami – Neurocomputing volume 415, 20 November 2020, pp 295-316 Available: https://www.sciencedirect.com/science/article/pii/S0925231220311693#b0165. [Accessed: September 5, 2023]

- Ghani et al (2015) – Accuracy Assessment of Urban Growth pATTERN classification Methods Using Confusion Matrix & ROC Analysis

- Fawcett (2005) – Introduction to ROC Analysis – pp 861 -874 - https://www.researchgate.net/publication/222511520_Introduction_to_ROC_analysis

- Winter (2002) – Handbook of Game Theory With Economic Applications pp 2025-2054 - https://www.sciencedirect.com/science/article/abs/pii/S1574000502030163