The Use of Founder Timeline Information to Predict Startup Success

Tom Marsden

University of Oxford, UK

March 24 2023

Abstract

The aim of this project is to create a computational model to predict the likelihood of a startup being successful based upon the timelines of its founders. Features were extracted from large data files containing the scraped LinkedIn profiles of founders. From this, we were able to extract information on employment and education history, as well as other data such as age. Through data exploration we were able to identify and engineer predictive features. The class of model used to predict success was a neural network. Using these models, we were able to achieve true positive rates of greater than 90% and overall accuracies over 80% using a variety of feature sets.

1 Introduction

The primary objective of Vela Partners is to use machine learning and data analysis to predict startup success rates. Investing in startups is a risky business and the majority of startups end up failing, so most investments venture capital (VC) firms make come to nothing. It is also impossible to use standard financial forecasting techniques due to the lack of financial data and unpredictable growth. Therefore, any edge or insight into the chances of a startup's success is highly valuable, since those companies which do succeed yield a high return on investment.

This is where VenTech comes in. VenTech is the use of data and technology to improve the efficacy of VC firms. With the improvements and increasing adoption of machine learning techniques, it is a logical step to try to apply them to the world of VC, especially given the ability of machine learning to make connections which humans cannot see.

In this version of Moneyball, we explore the predictive potential of data scraped from the LinkedIn profiles of founders, which has thus far not been used. This focuses on information about the timelines of the founders, such as education and employment timelines, along with limited categorisation of the education. In section 2 we will cover the structure of this dataset. Section 3 covers data exploration and feature engineering and in section 4 we use a neural network to predict success from this data.

2 Data Structure

The data consisted of two long .csv files, one for successful companies and one for unsuccessful, with two columns:

- **linkedin_url** contains a URL linking to the founder's LinkedIn page - this can be used to reference data for that founder in the main data file
- **json_string** contains a .json file in the form of a string which contains information from the founder's LinkedIn page.

The data within the .json files is what the whole paper focuses on. The useful data is structured as such:

- **education** contains all educational experiences listed on LinkedIn
  - **institution: name** contains name of the educational institution
  - **major: name** contains name of the subject studied
  - **degree: name** contains type of degree, e.g. Bachelor's, PhD
  - **from** contains start date in a string as well as a timestamp
  - **to** contains end date in a string as well as a timestamp, if not ongoing
- **employment** contains all jobs listed on LinkedIn
  - **employer: name** contains name of the employer
  - **title** contains job title
  - **from** contains start date in a string as well as a timestamp
  - **to** contains end date in a string as well as a timestamp, if not ongoing
- **other data** contains any other useful data such as name, age, etc.

This information can all be extracted using the json Python library. It was discovered that there was a non-negligible amount of missing or bad data. This was due to multiple factors. The main reason was people not uploading their employment and education history on LinkedIn. The data also appeared to be outdated, which could potentially lead to missing information. Some of the data appeared to have been corrupted - some dates went back as far as the 14$^{\text{th}}$ Century. Some jobs and educations were listed as beginning after they ended which had to be thrown out. An investigation into these issues, as well as re-scraping these profiles would be worthwhile.

After removing founders with no data points, we were left with 3865 successful founders and 5659 unsuccessful founders.

3 Data Analysis

These features are the result of brainstorming followed by data exploration which we will now dive into. The first group of features are all based off of temporal data. The second group is based off of classification of education information.

3.1 Age

This is simply just the age of a founder on the founding date of the company. We then plotted a normalised histogram of age for founders of both successful and unsuccessful companies which you can see in figure 1. As you can see, successful founders are slightly more likely than unsuccessful ones to be in the age range 25-40, whereas unsuccessful founders are more likely to be younger. The difference is quite small, but this is just one of many features with minor differences which we will be integrating into the models.

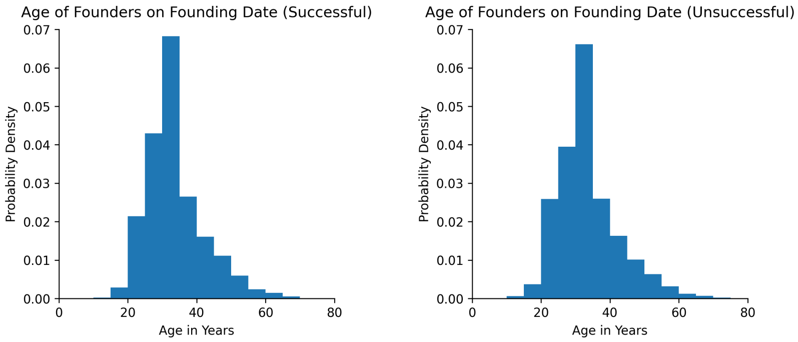


Figure 1: Normalised histograms of founder age on founding date

3.2 Number of Jobs and Educational Experiences

Here we are summing the number of either jobs or educational experiences that each founder has had. We start with the number of jobs of which a normalised histogram is plotted in figure 2. On average, successful founders have held more jobs than unsuccessful ones. This is especially apparent when you look at the skew around the peak on both plots. For education, as can be seen in figure 3, successful founders typically have more educational experiences than unsuccessful ones. The biggest difference is how many have three degrees - this can be interpreted as successful founders being more likely to have a PhD.

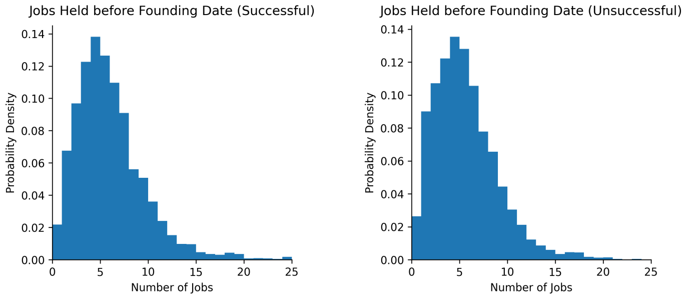


Figure 2: Normalised histograms of average number of jobs before founding date

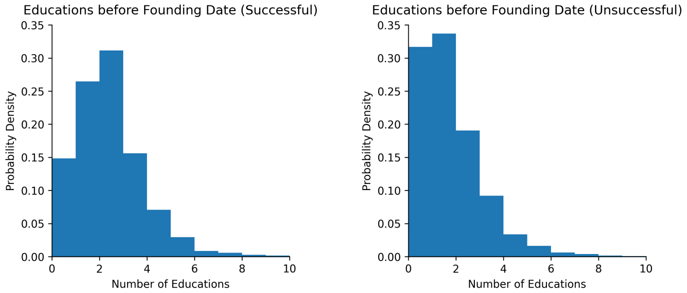


Figure 3: Normalised histograms of average number of educational experiences before founding date

3.3 Job and Educational Experience Length

We now look at each founder's average job length and average 'degree' (not all educational experiences listed are degrees) length. Starting with job length, which is plotted in figure 4, where successful founders tend to have jobs for longer than unsuccessful founders, especially when you focus on the 0-1 year range. Average degree length is plotted in figure 5. Successful founders are more likely to study for 3-4 years and unsuccessful founders are more likely to have studied for 0-1 years.

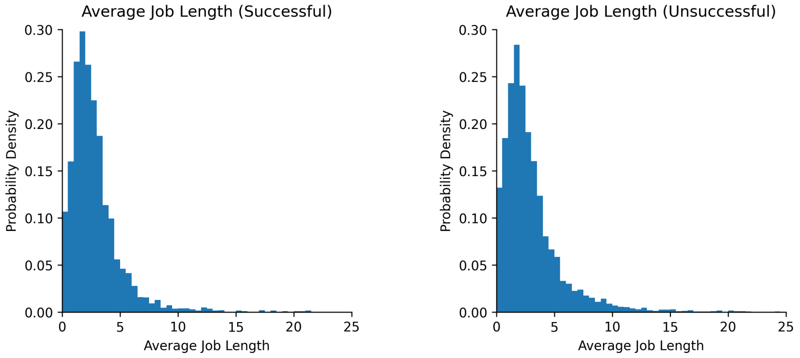


Figure 4: Normalised histograms of each founder's average job length

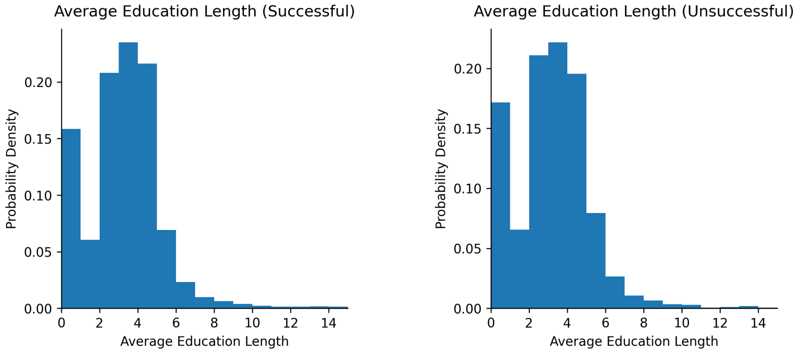
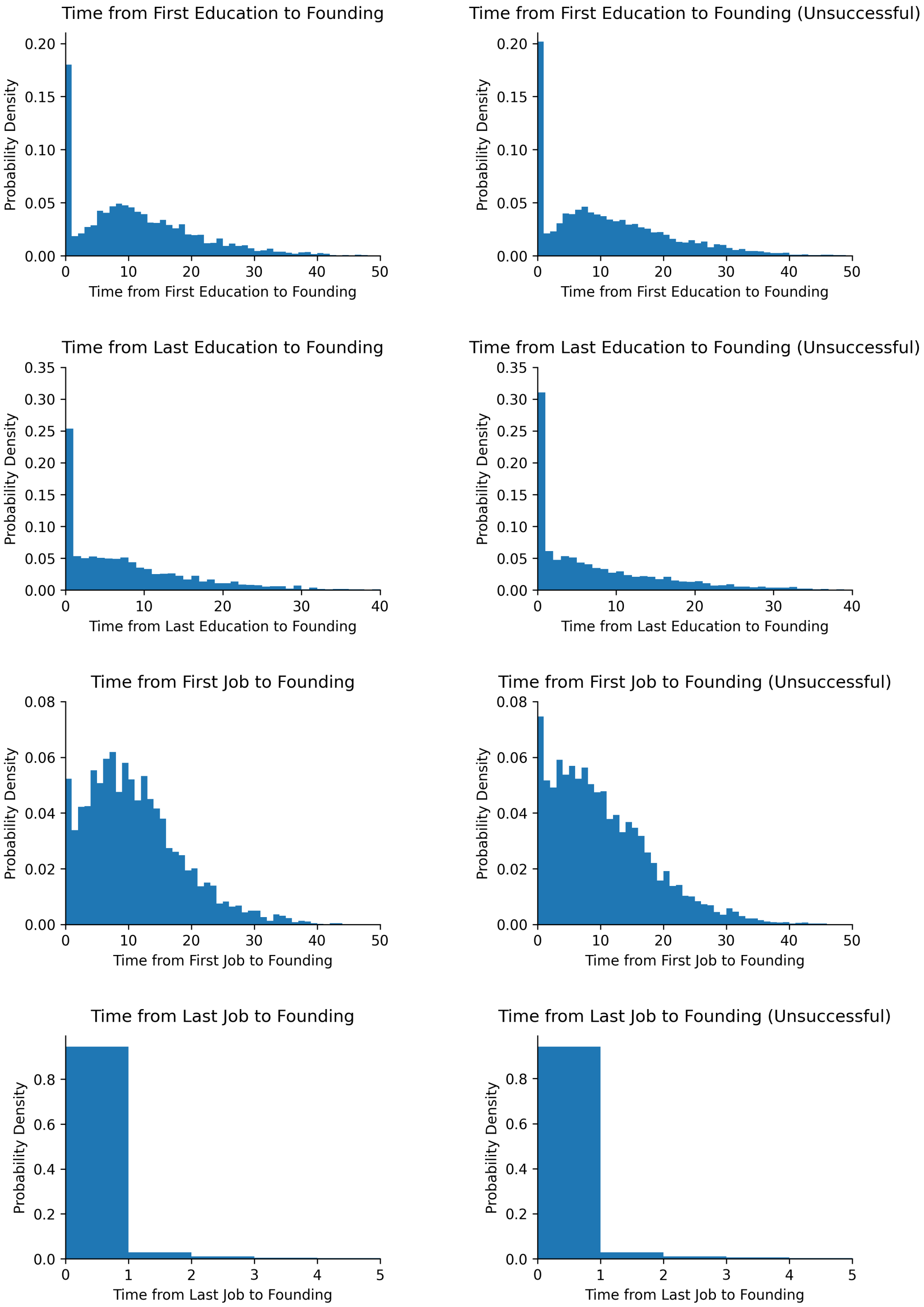


Figure 5: Normalised histograms of each founder's average degree length

3.4 Time before Founding

We now look at how long it takes founders to found each startup from different points on their timeline. Plots of this can be seen in figure 6. In general, successful founders are founding slightly later in their careers than unsuccessful ones. The peaks at zero for first education and first job are amplified since scores are set to zero where data doesn't exist.



Time from First Education to Founding

Time from First Education to Founding (Unsuccessful)

Time from Last Education to Founding

Time from Last Education to Founding (Unsuccessful)

Time from First Job to Founding

Time from First Job to Founding (Unsuccessful)

Time from Last Job to Founding

Time from Last Job to Founding (Unsuccessful)

Figure 6: Normalised histograms of time taken for each founder to found the startup

3.5 Time Unemployed

Time unemployed is found by looking at all of the time between their first listed job or education and founding, where they were neither working or studying. This is plotted in figure 7. Successful founders tend to spend slightly less time unemployed and unsuccessful founders are much more likely to spend a long time unemployed as can be seen from the tails of the histograms.

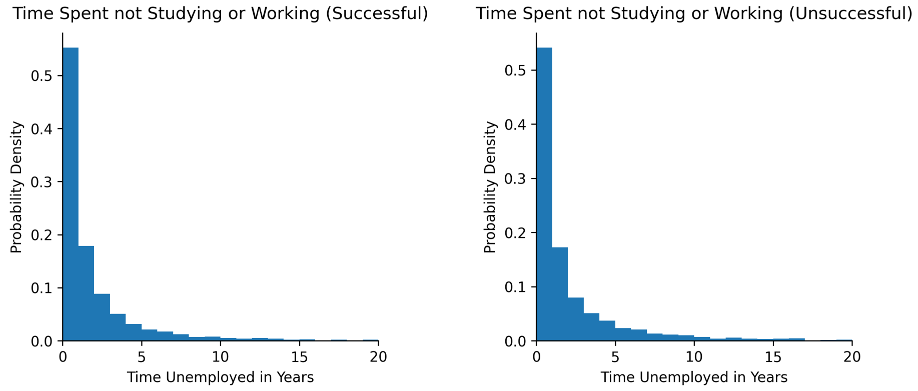


Figure 7: Normalised histograms of time spent unemployed by each founder

3.6 Number of Jobs at the Time of Founding

This feature represents how many ongoing jobs the founder had at the time of founding, including at the new company (can be zero since job at the new company is not always listed). This is plotted in figure 8, where you can see that successful founders are more likely to have just one job than unsuccessful founders who are more likely to have two than successful founders.

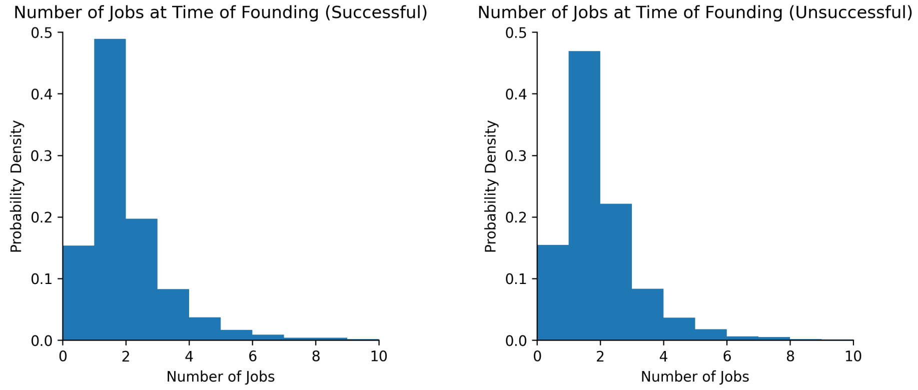


Figure 8: Normalised histograms of number of jobs at the time of founding

3.7 Degree Subject

This feature is a one-hot encode for certain degree types selected for significance, those being Computer Science, Mathematics, Chemistry, Business, Management, Marketing, and Biology. A bar chart of the percentage founders of each type, successful and unsuccessful, studying each subject can be seen in figure 9. Computer Science is clearly the most impactful subject, with a large difference between successful and unsuccessful founders who studied it, as well as a large number studying it overall.

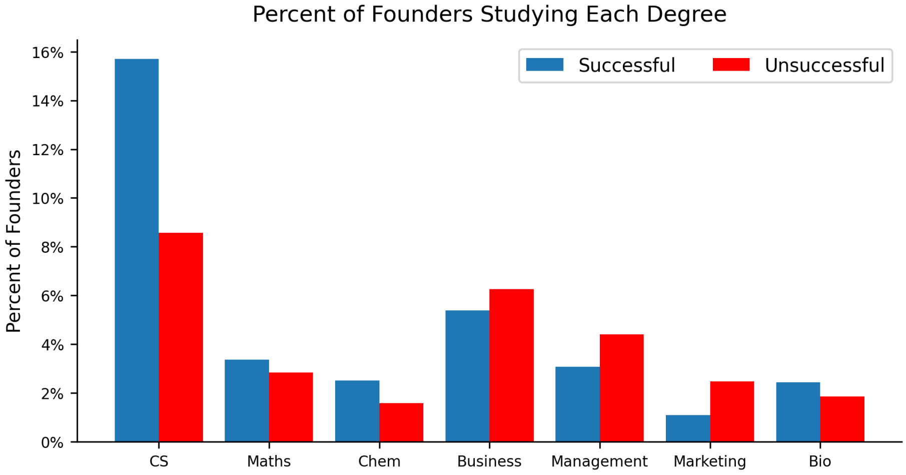


Figure 9: Bar chart of percentage of founders studying each degree subject

3.8 Degree Type

The final feature is degree type, which is a one-hot encode for whether or not the founder has a Master's degree and/or a PhD. This is shown as a bar chart in figure 10. Successful founders are more likely to have a PhD and slightly more likely to have a Master's.

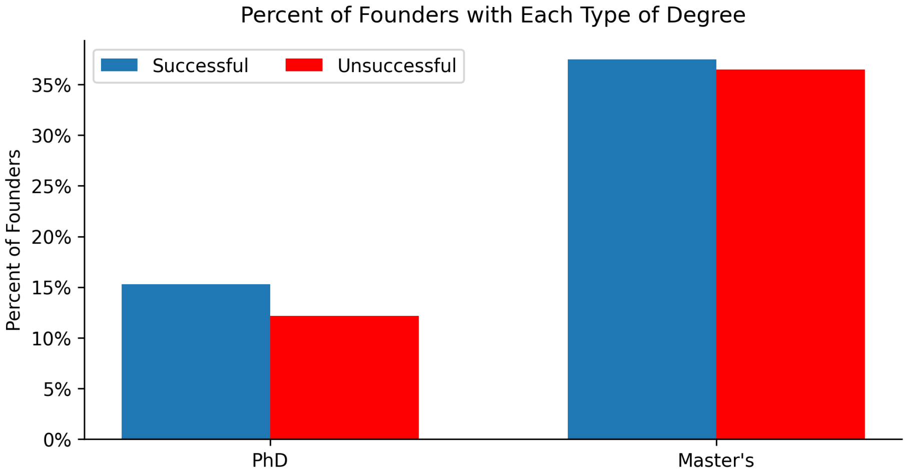


Figure 10: Bar chart of percentage of founders with each degree type

4 Modelling

To predict chances of startup success based on the extracted features, we used a neural network. The optimal structure for all of the feature sets employed was two hidden layers of size 40 and then 20. This was a good balance between accuracy and training performance. Between the layers we had a normalisation then a ReLU function. At the output, values were scaled to the range $[0, 1]$ using a sigmoid function. Success was designated > 0.5 and fail was designated < 0.5. Three different feature sets were used:

- **Temporal** had all of the temporal based values - age; number of jobs and of degrees; average length of jobs and of degrees; time to found from end of first education; form end of last education; from start of first job; from end of last job; unemployed time; number of jobs at time of founding.
- **No Subject** was the same as temporal but with degree type added.
- **All** had all of the features. Equivalent to no subject but with degree subjects added.

Each data set was split into training data (72%), validating data (18%) and testing data (10%).

4.1 Temporal Model

The temporal feature set had the highest true positive rate, albeit with a dataset which had a higher random true positive rate. The full set of data does not have an equal split of successful founders and unsuccessful founders, but has 50% more unsuccessful founders than successful. With the full dataset, it gets into a local minimum of predicting all startups will fail. When the number of successful and unsuccessful founders is equal, it does not have this issue and performs properly as can be seen in figure 11. The results of this model on the testing data can be seen below in table 1 and the confusion matrix in figure 12.

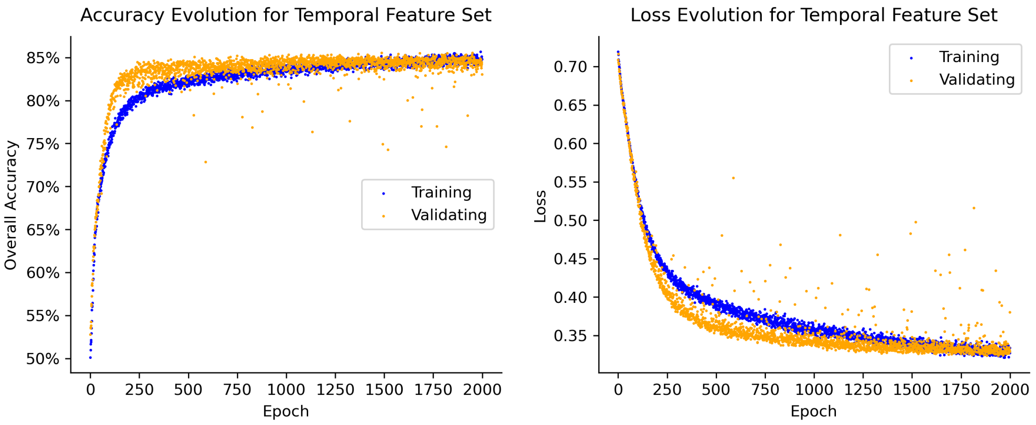


Figure 11: Accuracy and loss evolution over training period

Criteria	Model Result	Random Model Result
True Positive Rate	0.97	0.50
False Negative Rate	0.03	0.50
True Negative Rate	0.72	0.50
False Positive Rate	0.28	0.50
Precision	0.79	0.50
Accuracy	0.85	0.50

Table 1: Table of performance metrics for temporal model

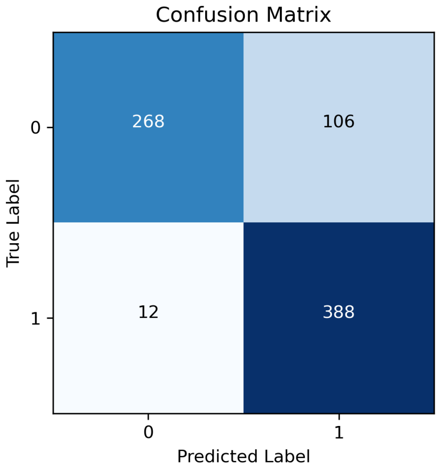


Figure 12: Confusion matrix for temporal model

The 97% true positive rate is highly impressive, however the true negative rate is not fantastic and this is somewhere where this model is lacking. The validation accuracy and loss sometimes significantly vary in a random manner during training, although a learning rate scheduler would quite likely go a long way towards fixing this.

4.2 No Subject Model

This expanded feature set performed the best. It was also possible to run the full dataset since it did not get caught by the local minimum. The performance during training can be seen in figure 13. The results of this model on the testing data can be seen below in table 2 and the confusion matrix in figure 14.

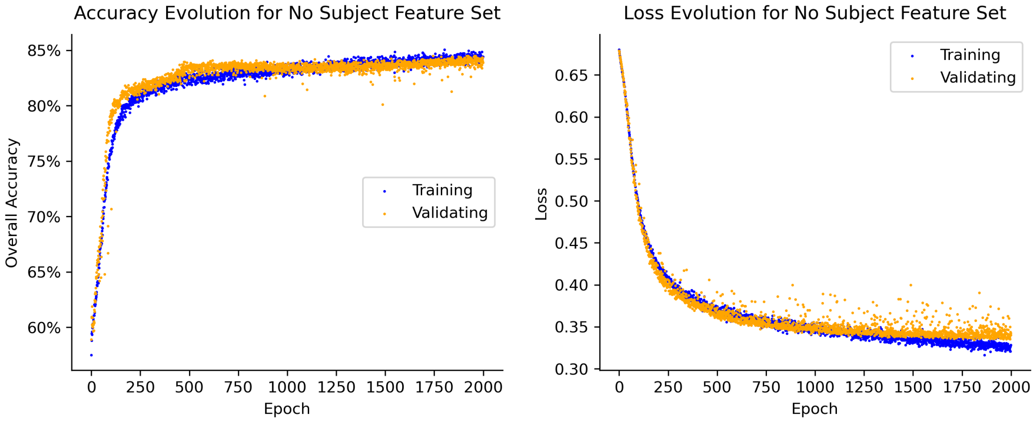


Figure 13: Accuracy and loss evolution over training period

The true positive rate may be lower on this model, but the true negative rate is much higher and the precision is more impressive when compared to the random model result, which is different since there is no longer a 50/50 split of successful/unsuccessful data. This model is also a lot more stable than the previous one, not getting stuck in local minima and with a lot less random variation during training. Degree type appears to be a worthwhile addition to the feature set.

Criteria	Model Result	Random Model Result
True Positive Rate	0.92	0.50
False Negative Rate	0.08	0.50
True Negative Rate	0.80	0.50
False Positive Rate	0.20	0.50
Precision	0.74	0.41
Accuracy	0.84	0.50

Table 2: Table of performance metrics for no subject model

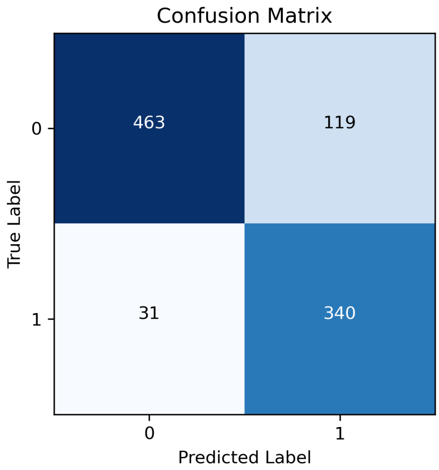


Figure 14: Confusion matrix for no subject model

4.3 All Feature Model

This expanded feature set did not perform as well as the previous one. The performance during training can be seen in figure 15. The results of this model on the testing data can be seen below in table 3 and the confusion matrix in figure 16.

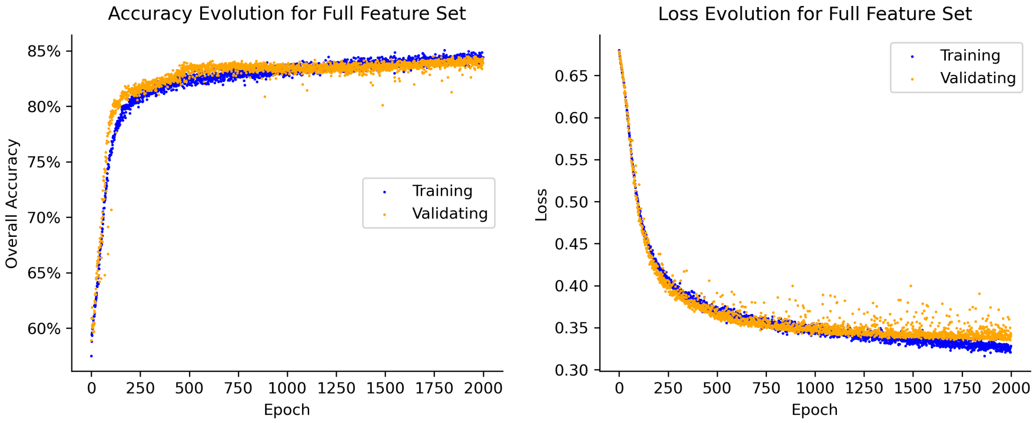


Figure 15: Accuracy and loss evolution over training period

Criteria	Model Result	Random Model Result
True Positive Rate	0.89	0.50
False Negative Rate	0.11	0.50
True Negative Rate	0.81	0.50
False Positive Rate	0.19	0.50
Precision	0.75	0.41
Accuracy	0.84	0.50

Table 3: Table of performance metrics for all feature model

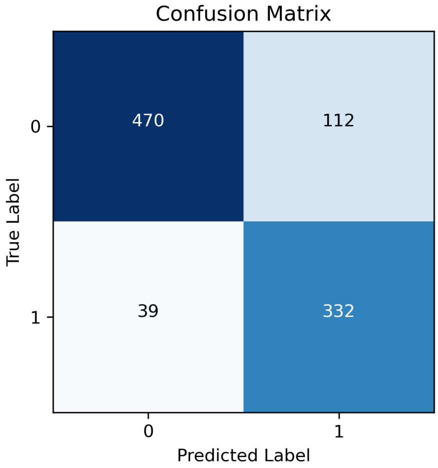


Figure 16: Confusion matrix for all feature model

The true positive rate is lower on this model but in return, the true negative rate slightly increases, but not by as much. Due to the split of the data being more heavily weighted towards unsuccessful founders, this results in a small increase to precision. As such, unless degree subject data can be better engineered or modelled, it may not be worth including in the model.

5 Conclusion

It appears that, even with somewhat basic features, founder timeline modelling has a great deal of potential and does not require a particularly advanced neural network to predict with high accuracy. The addition of degree types is of clear benefit to the model, however the integration of degree subject need improvement given its detrimental effect on the true positive rate.

Future work could focus on re-scraping the LinkedIn profiles for more accurate and up to date data, or classifying job titles and companies using embedding. Also using a k-nearest neighbours algorithm to fill holes in data sets may be of use. Some simpler ideas are to add learning rate scheduling and work out which of the features has the greatest significance. A more general idea is to weigh startups according to how recently they were founded in order to best reflect current markets, which may be most useful when applied to the Proxy model.