Yusuf Karadogan

Final Report:
Start-Up Prediction Success

Problem Statement

A startup or start-up is a company or project begun by an entrepreneur to seek, develop, and validate a scalable economic model. While entrepreneurship refers to all new businesses, including self-employment and businesses that never intend to become registered, startups refer to new businesses that intend to grow large beyond the solo founder. Startups face high uncertainty and have high rates of failure, but a minority of them do go on to be successful and influential. Some startups become unicorns: privately held startup companies valued at over US$1 billion.

This data set was obtained through Kaggle. The objective is to predict whether a startup which is currently operating turns into a success or a failure. The success of a company is defined as the event that gives the company's founders a large sum of money through the process of M&A (Merger and Acquisition) or an IPO (Initial Public Offering). A company would be considered as failed if it had to be shut down.

Startups play a major role in economic growth. They bring new ideas, spur innovation, create employment thereby moving the economy. There has been an exponential growth in startups over the past few years. Predicting the success of a startup allows investors to find companies that have the potential for rapid growth, thereby allowing them to be one step ahead of the competition.

After wrangling, exploring and feature engineering, I was able to compare over a dozen different machine learning models. When compared side by side the CatBoost Classification model was able to achieve an average precision of 0.82, accuracy of 0.77 and AUC of 0.85. This process can be repeated for many other startups and with more variables can be improved.

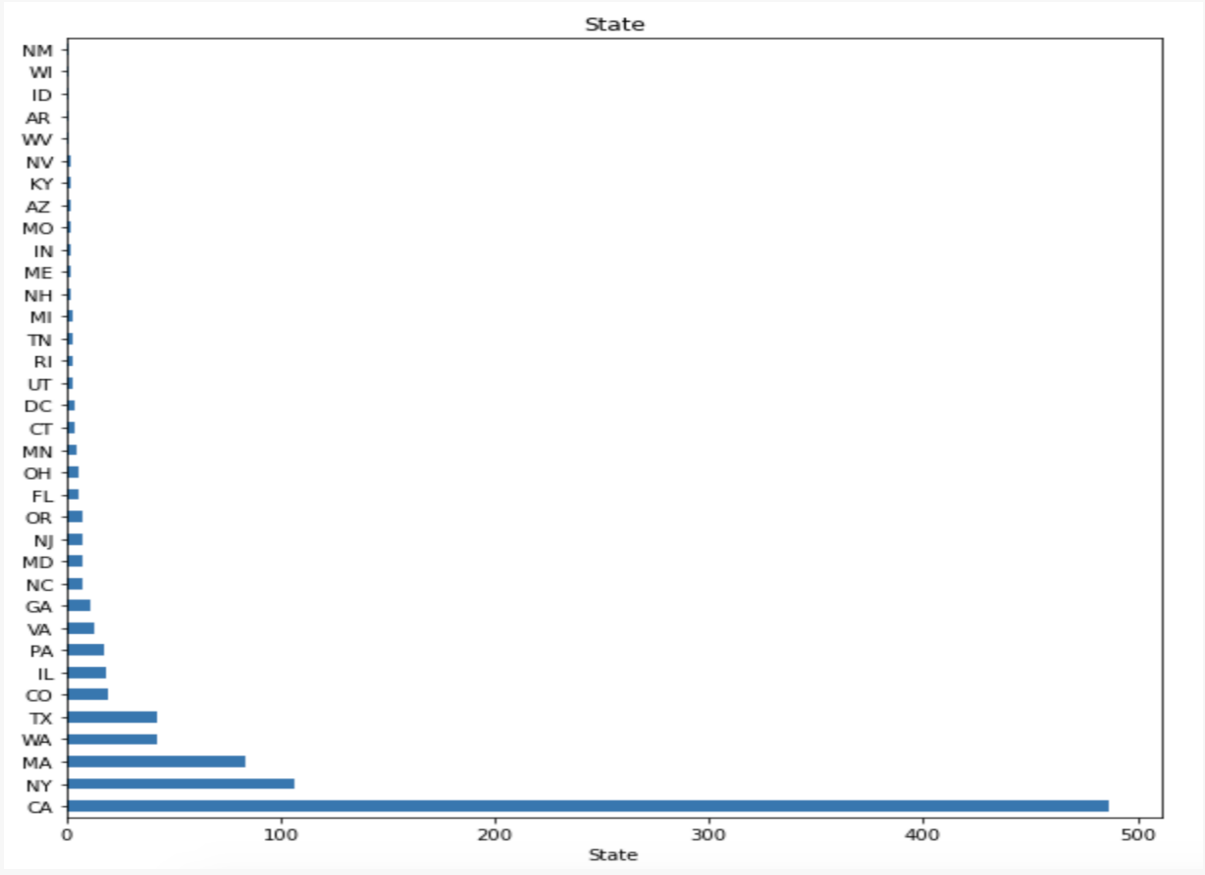
Data Wrangling

The initial data set contained 923 rows and 48 columns, while it was not a huge

set, it was still enough to work with and build upon. I started analysing the dataset as a whole, from there focused on the missing values for each column. Two of the most missing values were from the columns that had no impact in our model development; such as closed_at and unnamed_6. The closed_at column showed that 63.7% of the companies started were still operating. The other column without impact with the most missing values, unnamed_6, which was a combination of city, state, zip code, was eventually dropped.

Next I looked if there were any duplicate rows, fortunately there were only two. The information for Redwood Systems was entered twice, I went a head and dropped one of the duplicate rows. I then went ahead and dropped some of the unnecessary columns "Unnamed: 6", "Unnamed: 0", "id", "State_code.1" and "closed_at". From here after analysing the types of columns, I decided to convert the year columns ("founded_at", "first_funding_at", "last_funding_at") from object to datetime.

At this point without going too much into it, I was able to visually see that most startups were from California. The final shape of my dataset at this stage was 922 rows and 44 columns.



Exploratory Data Analysis

This is the stage where I dealt with the other two columns with the missing values; which were age_first_milestone_year and age_last_mile_stone year. The former represents some of the new companies who still were early stages and had not reached the first milestone they set for themselves. The latter had missing values because some companies still had not reached their last milestone. To fix these missing values, I filled both with the mean of each column. Next since we're trying to predict the outcome if a company is a success or not, I dropped the actual results and replaced them with dummie values, under the "status" column.

Next I wanted to visualize the start ups in different sectors and amount of funding and rounds for the top states. Fig. 1 displays the different sectors and Fig. 2 shows the funding amount and rounds.

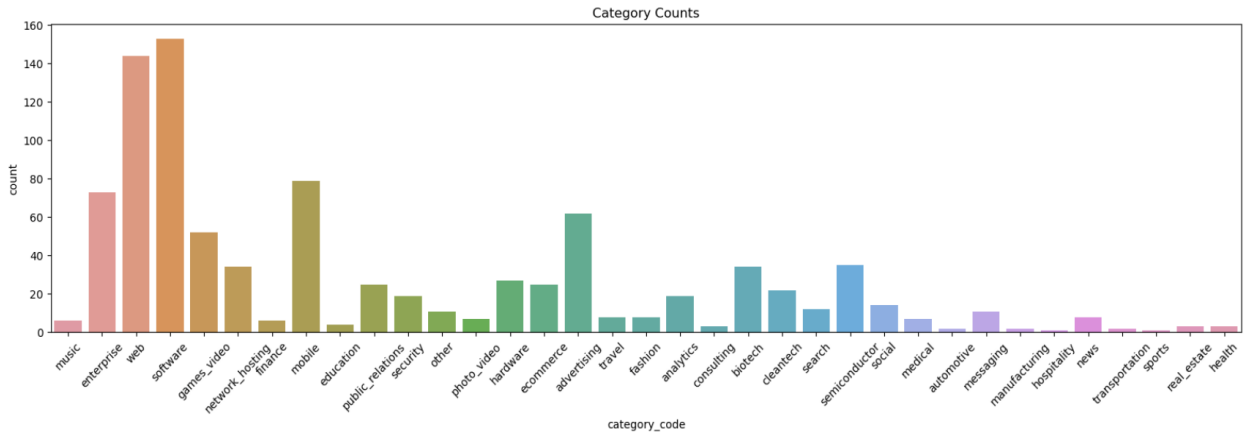


Figure 1: Bar Plot of the sectors

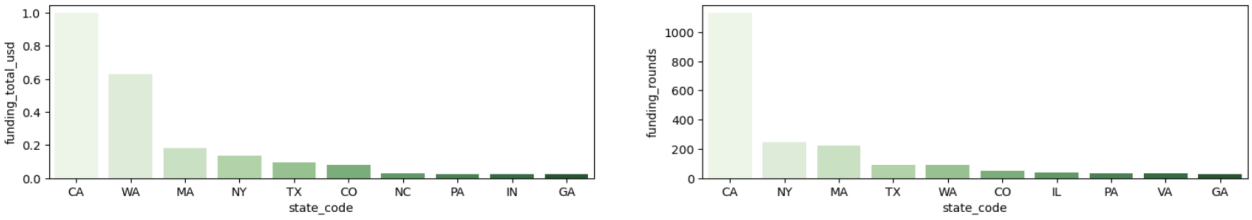


Figure 2: Bar Plot of top states in funding

I looked at the correlations between every column, which resulted in some strong correlations between the time value columns. These correlations between the time value columns ranged between 0.39 and 0.76, which showed significance. After some optimization and tuning, the Fig. 3 below shows some of these correlations.

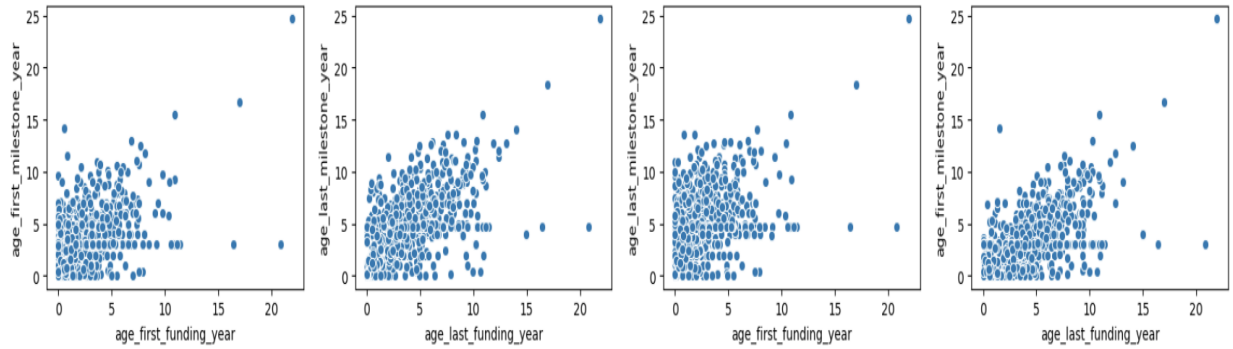


Figure 1: Scatter Plot of the time value correlations

Next I looked at some of the outliers. Fig. 4 shows there were a lot of outliers, so in order to fix this I took the log of the time value columns which improved the amount of outliers, seen in Fig. 5.

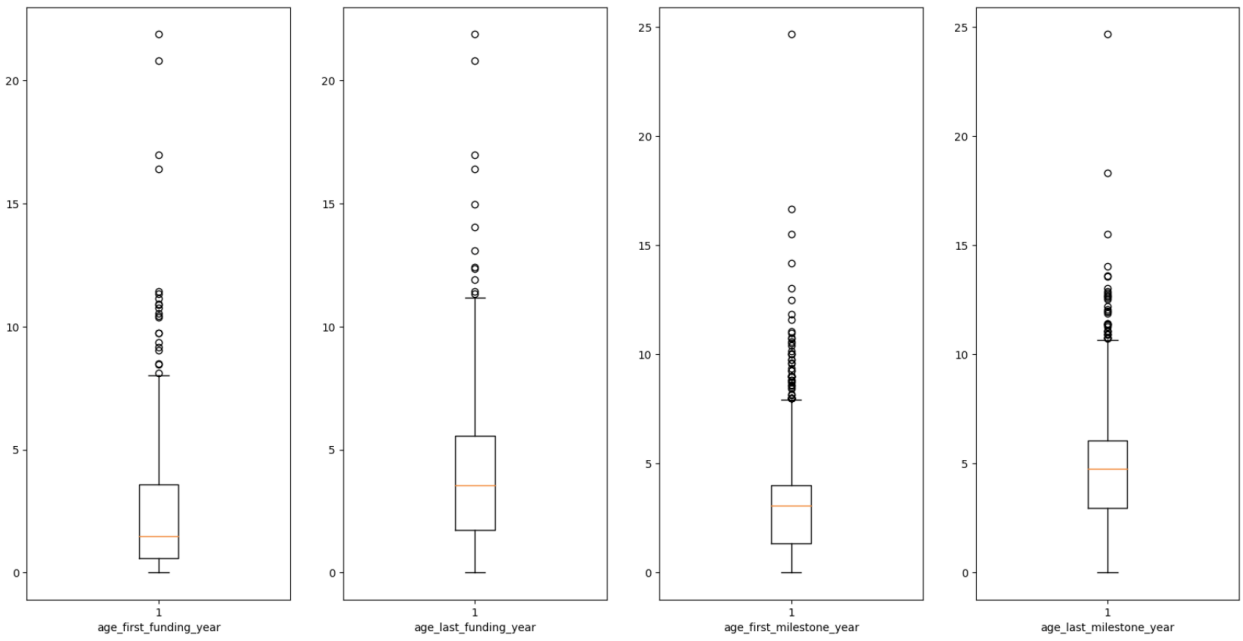


Figure 4: BoxPlot of the time value columns

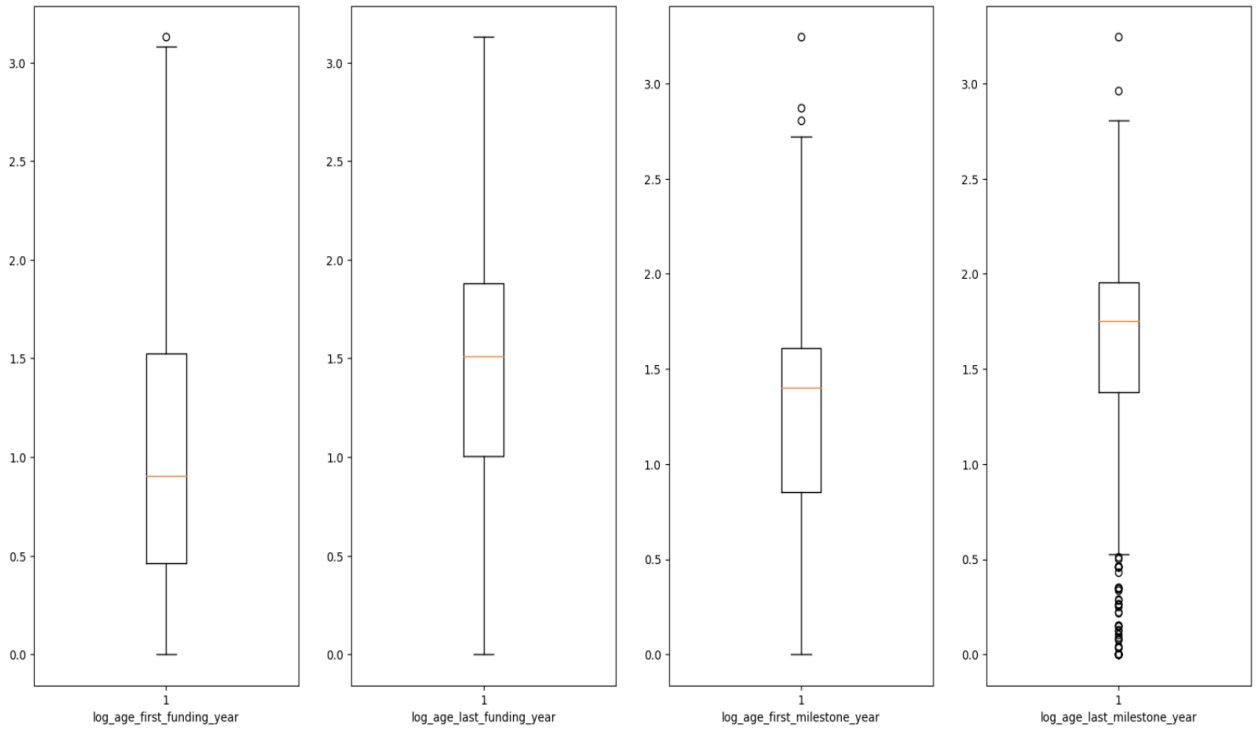


Figure 5: BoxPlot of the log of time value columns

Feature Engineering

Now the data is all cleaned and analysed to understand the relationships between different variables, I need to create and customize different features to prepare my dataset for modelling. Because this dataset was from kaggle, a lot of the feature engineering that was previously done was optimal for me to start modelling. The category/sector columb was further engineered to include the top 9 categories and the rest were grouped into “is_other”. The same thing was applied for the top 4 states and the rest were grouped into “is_otherstate”.

The main thing I did at this point was remove a few more features (“latitude”, “longitude”, '”zip_code'”, ”city”, “labels”, “object_id”, “is_top500”) that would not be needed for the modelling part, determined in the previous steps. After this I rounded all of the numerical values to 2 decimal points and was ready to start modelling.

Table 1 below is what I used for modelling.

	0	1	2	3	4
state_code	CA	CA	CA	CA	CA
name	Bandsintown	TriCipher	Plixi	Solidcore Systems	Inhale Digital
founded_at	2007-01-01 00:00:00	2000-01-01 00:00:00	2009-03-18 00:00:00	2002-01-01 00:00:00	2010-08-01 00:00:00
first_funding_at	2009-04-01 00:00:00	2005-02-14 00:00:00	2010-03-30 00:00:00	2005-02-17 00:00:00	2010-08-01 00:00:00
last_funding_at	2010-01-01 00:00:00	2009-12-28 00:00:00	2010-03-30 00:00:00	2007-04-25 00:00:00	2012-04-01 00:00:00
age_first_funding_year	2.25	5.13	1.03	3.13	0
age_last_funding_year	3	10	1.03	5.32	1.67
age_first_milestone_year	4.67	7.01	1.46	6	0.04
age_last_milestone_year	6.7	7.01	2.21	6	0.04
relationships	3	9	5	5	2
funding_rounds	3	4	1	3	2
funding_total_usd	375000	40100000	2600000	40000000	1300000
milestones	3	1	2	1	1
is_CA	1	1	1	1	1
is_NY	0	0	0	0	0
is_MA	0	0	0	0	0
is_TX	0	0	0	0	0
is_otherstate	0	0	0	0	0
category_code	music	enterprise	web	software	games_video
is_software	0	0	0	1	0
is_web	0	0	1	0	0
is_mobile	0	0	0	0	0
is_enterprise	0	1	0	0	0
is_advertising	0	0	0	0	0
is_gamesvideo	0	0	0	0	1
is_ecommerce	0	0	0	0	0
is_biotech	0	0	0	0	0
is_consulting	0	0	0	0	0
is_othercategory	1	0	0	0	0
has_VC	0	1	0	0	1
has_angel	1	0	0	0	1
has_roundA	0	0	1	0	0
has_roundB	0	1	0	1	0
has_roundC	0	1	0	1	0
has_roundD	0	1	0	1	0
avg_participants	1	4.75	4	3.33	1
status_closed	0	0	0	0	1

Table 1: Final feature engineered dataset

Model Selection

The objective is to predict whether a startup which is currently operating turns into a success or a failure. The success of a company is defined as the event that gives the company's founders a large sum of money through the process of M&A (Merger and Acquisition) or an IPO (Initial Public Offering). A company would be considered as failed if it had to be shut down. So for the modelling section my goal was to build the optimal model that would have the highest accuracy of predicting the success of a startup.

Instead of trying different individual models, I used PyCaret to run 14 different models for me and rank them based on their accuracy.

Table 2 below shows the different models ranked based on their accuracy score.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.7695	0.7991	0.5154	0.8118	0.6199	0.4672	0.4985	1.4600
gbc	Gradient Boosting Classifier	0.7629	0.7977	0.5507	0.7538	0.6267	0.4609	0.4787	0.0750
rf	Random Forest Classifier	0.7585	0.7842	0.4974	0.7782	0.5945	0.4386	0.4662	0.0810
lightgbm	Light Gradient Boosting Machine	0.7473	0.7770	0.5809	0.6977	0.6242	0.4387	0.4493	0.0220
et	Extra Trees Classifier	0.7363	0.7540	0.4364	0.7487	0.5454	0.3795	0.4095	0.0770
ada	Ada Boost Classifier	0.7361	0.7660	0.5886	0.6747	0.6187	0.4205	0.4306	0.0450
dt	Decision Tree Classifier	0.6741	0.6494	0.5522	0.5640	0.5532	0.2986	0.3020	0.0100
knn	K Neighbors Classifier	0.6409	0.6443	0.4092	0.5155	0.4473	0.1932	0.1979	0.0130
lr	Logistic Regression	0.6297	0.6531	0.0000	0.0000	0.0000	0.0000	0.0000	0.3710
svm	SVM - Linear Kernel	0.6297	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0090
ridge	Ridge Classifier	0.6297	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0120
qda	Quadratic Discriminant Analysis	0.5545	0.5627	0.5926	0.4294	0.4960	0.1175	0.1220	0.0170
lda	Linear Discriminant Analysis	0.5412	0.4367	0.2651	0.3610	0.3000	-0.0304	-0.0268	0.0250
nb	Naive Bayes	0.4186	0.4721	0.8581	0.3382	0.4848	0.0146	0.0316	0.0100

Table 2: All of the different models ranked based on accuracy

From the table above, I chose 3 different machine learning classification models: Catboost Classifier, Random Forest Classifier, and Gradient Boosting Classifier. The metric I focused on when building my models was accuracy. I wanted my model to predict the success of a startup and compared it to the known status of the company.

Before building the models, I would like to share the importance of different features that have the most impact on the success of a startup. Figure 6 shows the top features that have the most impact. In my initial analysis, I did not put too much emphasis on how important relationships were, however, we see that it is almost as important as the amount of funding that a startup receives. Anything above 7.5, I would consider as of high importance.

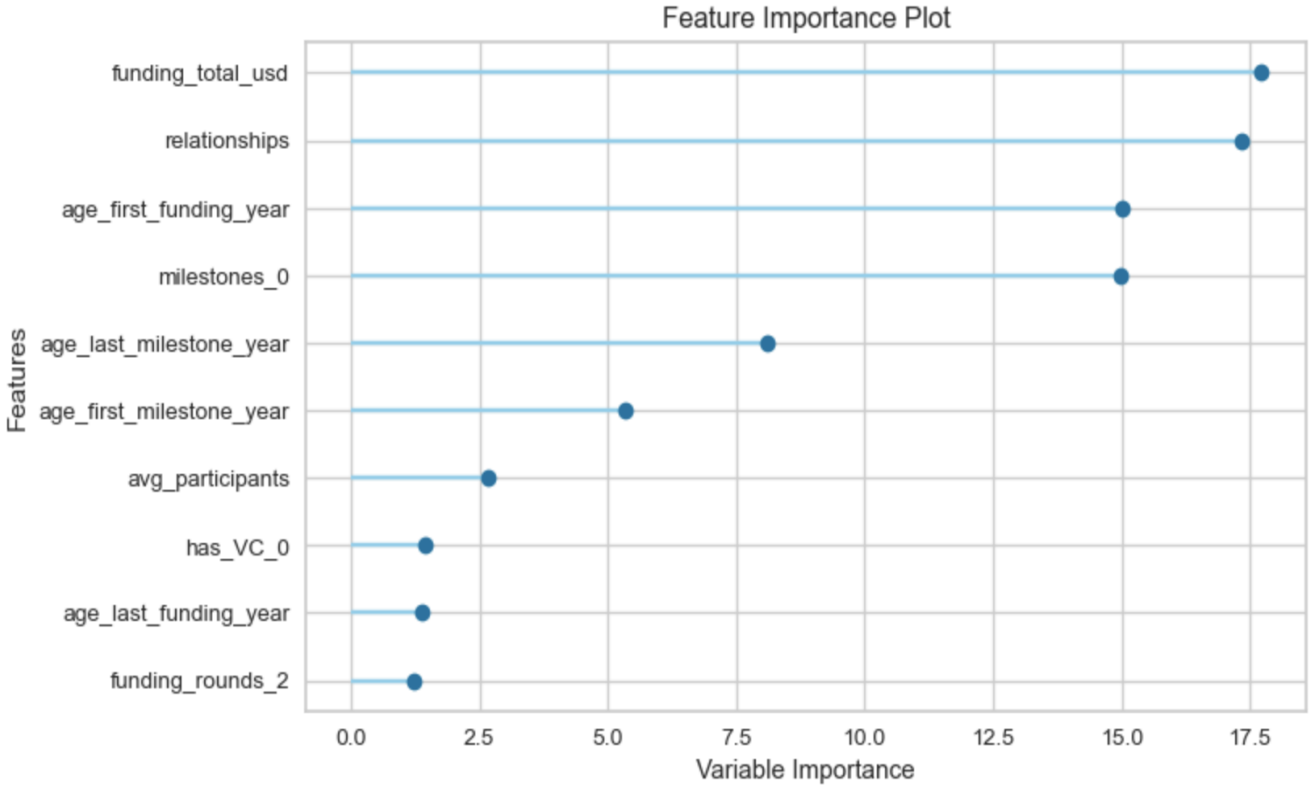


Figure 6: Feature importance

Going forward, I would like to focus on the most successful prediction model, which was CatBoost Classifier. Figure 7 shows that after tuning my model, I was able to increase my AUC score to 0.86 from 0.79.

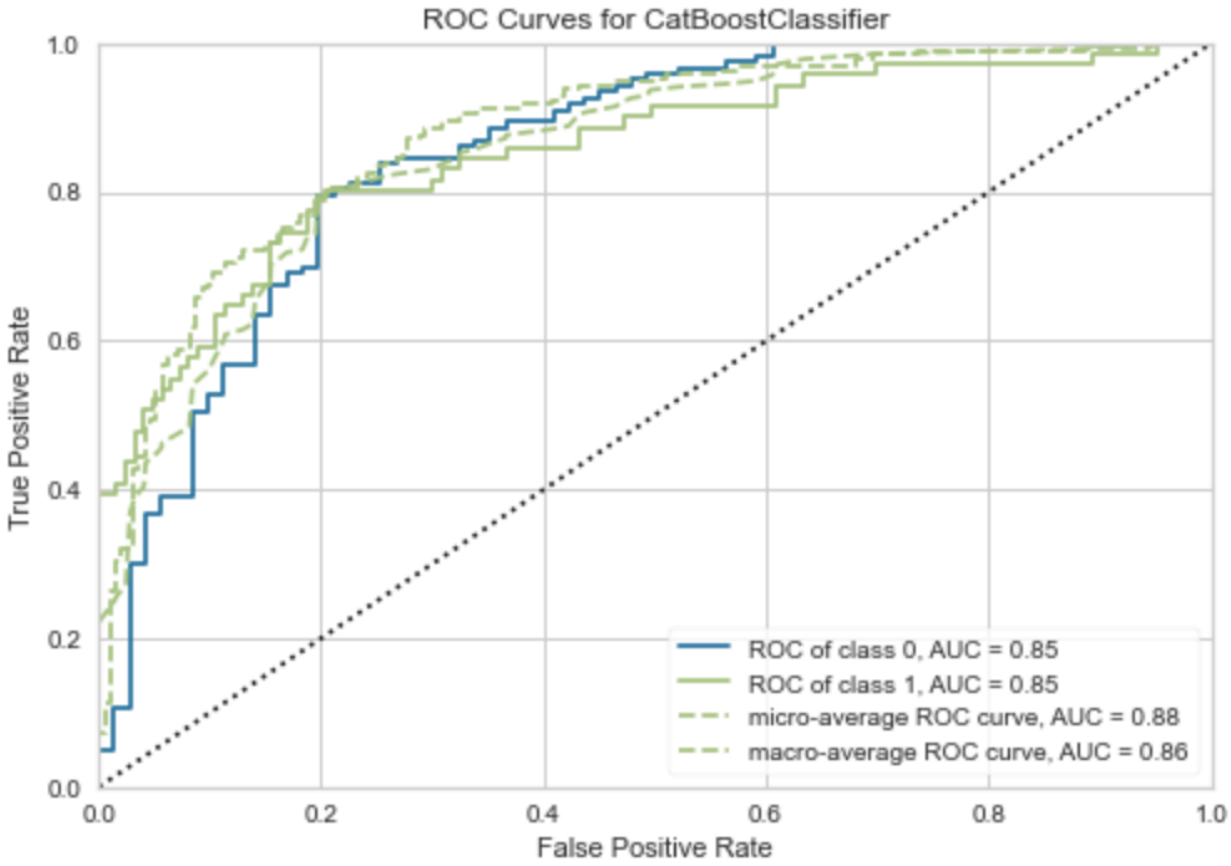


Figure 8: ROC curve for CatBoost Classsifier

Takeaways

We see that CatBoost Classifier is the best model for predicting startup success. Consistently outperforming the other models CatBoost Classifier gave the highest accuracy score each time which is the most important aspect of the prediction.

It is known that only a small number of startups end up succeeding. Most of the money that a VC invests does not bring any return, however when you average the losers with the homeruns with unicorns, VC and investors who know what they are doing and hedge their portfolio actually come up on top.

Although due to share dilution the returns in the later stage investing is lower, vs early angel and seed stage investing, I think investment firms and venture capital firms should wait to invest in companies to limit their risk and increase the likelihood of their returns.

Investors and VC's should also put a lot more emphasis on the amount of relationships that the founders have. As we saw in our model and feature importance, it is as important as the amount of money they've raised. In hindsight, this makes sense because a relationship with the right person can bring in the first client, investor, idea and employees. All these things are crucial for the success of a start-up.