Using logistic regression to predict success and to recognize success attributes in startups

ABSTRACT

Predicting the success of a startup is useful to investors and founders alike. In this paper, we present a machine learning model that classifies companies based on information about the industry of the startup, the founders, and the investors which have invested until a certain point. The definition for a successful company we use is having more than USD 500M valuation. On the other hand, an unsuccessful company is as startup that has raised more than USD 1M more than 5 years ago but has never surpassed USD 500M valuation. Our model achieves around 77% accuracy on unseen data. Results indicate that there is a strong correlation between many of the attributes in the data and success of a company, both in the positive and negative direction.

1 INTRODUCTION

Understanding what makes a startup successful could aid investors to make better decisions. The insights the logistic regression model provides could further strengthen investors' observations about success or even present them with a new perspective. Another way machine learning could be useful in this context is in finding new promising startups. The model can analyze a huge number of companies, which investors are not be able to go through themselves.

The number of startups which have raised more than USD 1M is not large. Our data has around 2300 entries. Intuitively, this means that complex models like neural network will not do well on this dataset. That means we are limited to simpler techniques like k-nearest neighbors, random forest, and logistic regression. Those models, because of their simplicity, also have the advantage of being easier to interpret. Logistic regression is especially good for our use case as it produces a coefficient for each attribute of the data.

The dataset we work with includes only countries from the United States. It consists of 13 attributes - *'name'*, *'founded_year'*, *'country_code'*, *'city'*, *'category_list'*, *'category_groups_list'*, *'universities_of_founders'*, *'degrees_of_founders'*, *'subject_degrees_of_founders'*, *'gender_of_founders'*, *'city_of_founders'*, *'prev_companies_of_founders'*, *'prev_title_of_founders'*, *'investor_name'*. We only use some of them in our model. Additionally, we create a new feature *'number_of_founders'* that is deduced from the other attributes. Our model uses the one-hot encoding of this data. This is done because most of our features are categorical values or lists of categorical values. Our results indicate that this method of processing the data produces satisfactory predictions.

2 METHODOLOGY

In this section, we describe our method for processing the data and training our model.

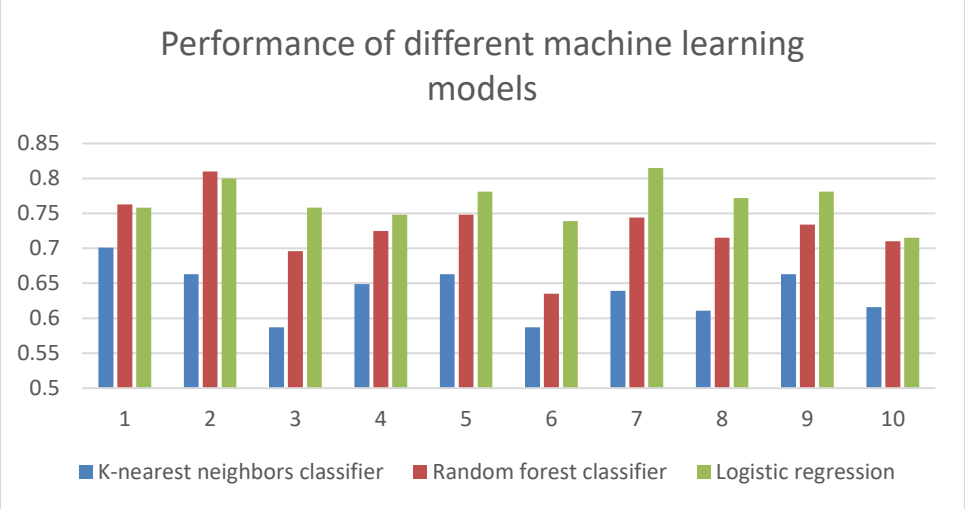
Firstly, we load the data from the different sheets of our tables of data (one for successful and one for unsuccessful startups). We have chosen to keep only the data that has an entry in each of those sheets, in order to have the most information about each company. After that, we need to change the names of some of the universities, as some universities have different names in the successful and unsuccessful dataset.

The next step is creating the new feature *'number_of_founders'*. Because of inconsistencies in the data, we cannot be sure that this value is accurate. That is why we try to deduce it from two sources – *'universities_of_founders'* and *'prev_companies_of_founders'*. The feature *'universities_of_founders'* should theoretically have the same length as the number of founders. The attribute *'prev_companies_of_founders'* includes the startup of the founders. Therefore, it should theoretically be included the same number of times as their number. That is however not the case because sometimes there are missing values. That is why we take the maximum of those two values to get a better approximation of the true count.

After this, we remove the features *'name'*, *'founded_year'*, and *'country_code'*. As all the startups are from the United States, *'country_code'* is always 'US'. Although *'name'* could have some relevance, we deem it does not carry important information. *'founded_year'* is removed because it wouldn't make sense to insert it as a categorical value. It could be added as an adjusted value in a future version of the model.

The remaining features are then one-hot encoded. The number of instances of the same value in a feature is ignored, i.e. the final value is either 0 or 1. After that, the *'Other'* value in *'category_groups_list'* is removed, as it is only present in the unsuccessful companies dataset, which is most likely an error of the data.

To test our models, we split the data into a training and a test batch. We have tested a k-nearest neighbors classifier, a random forest classifier, and a logistic regression model. We found that, on average, the logistic regression model outperforms the other two, with the random forest classifier being second. Additionally, we have produced a sorted list of the most important features for the success of the company from the logistical model.



3 RESULTS

Our logistic regression model predicts whether a startup will be successful with around 77% accuracy on average. The sorted list of the most important values it creates has some interesting results:

Top 5 most important values according to the model (both successful and unsuccessful)

Top 5 most successful	**Top 5 most unsuccessful**
prev_title_of_founders: Board Member	*city_of_founders*: null
investor_name: Accel	*degree_of_founders*: unknown
category_list: Enterprise Software	*gender_of_founders*: female
universities_of_founders: Stanford University	*prev_companies_of_founders*: Inc. (with a space in front)
category_list: SaaS	*number_of_founders*: 1

4 CONCLUSION

Machine learning can be used to predict the future success of companies. It can also help investors find and choose successful companies. The logistic regression we have developed finds some fascinating patterns in our dataset.

According to the model, having more than 5 company founders is better than having less. Additionally, having just 1 or 2 founders is among one of the worst indicators for a startup.

Investors are also a very important indicator. The best ones are Accel, SV Angel, GV, Sequoia Capital, and Benchmark. On the other hand, some investors are a bad sign, e.g. 500 Startups, Kima Ventures, Betaworks, Boost VC, and QueensBridge Venture Partners.

The best subjects for degrees of a company's founders are the following: Computer Science, History, Psychology, Management, and Computer Science and Engineering. Ironically, Entrepreneurship seems to be among the most unsuccessful subjects.

The best universities for a startup are Stanford University, Technion, Massachusetts Institute of Technology, Tel Aviv University, Yale University. The worst ones are University of Washington, Vanderbilt University, Caltech, Purdue University, University of Oxford.

One surprising finding the model makes is that having a woman on the founders team is a negative sign for the success of the business. This might be for many reasons, one of which is discrimination.

Two other attributes that correlate negatively with success according to the logistic regression are previous titles " CEO" and " CTO". Both start with a space and actually indicate a title that includes a comma, e.g. "Founder, CEO" (that is because our data is comma separated). This might indicate that using commas in titles is a bad idea.

5 FURTHER IDEAS

There are further ideas we have not developed. To start, any improvements on the quality or quantity of data will lead to an improvement in the model. To add to that, new features like financial data or number of Internet searches would be useful. The final improvement we could do is to change the logistic regression model to an AutoGluon or LightGBM model.