Factors that Influence the Success of Startups

Arpan Chatterji, Jacob Bulzak, Rajsitee Dhavale

Spring 2022

Abstract

Data Science can be applied in venture capital to augment human insight with data-driven decision-making. This project aims to analyze trends in the startup ecosystem in the United States to find trends that would otherwise be invisible to the naked eye. Our dataset, obtained from Kaggle, consists of 923 observations and 49 variables. These startups belong to 35 different industries and were founded and acquired/closed between 2000 and 2013.

We approach the problem from the point of view of an investor performing due diligence on historic trends as part of his industry analysis. We aim to figure out which variables impacted the fate of the startups in the dataset. Through our analysis, we seek to develop a series of 0 to 1 outcomes to find probabilities on multiple aspects of a deal. These can then help the investor to better understand the current and future landscape and make decisions informed by data. Through our models, we try to pinpoint the most important factors that led to the success or failure of the startups. We define success as the startup getting acquired by a firm; we classify a startup as a failure if it shuts down. The variables we use are (i). location (state), (ii). type of the startup, (iii). the number of milestones achieved by the startup, (iv). whether the startup has venture capital backing, (v). whether the startup has angel investor backing, (vi). the total amount of funding (in USD) collected by the startup and (vii). whether it was acquired by a top 500 company.

We develop a Logistic Regression Model as a baseline model and then compare its results with those of the Random Forest model. We find that the factors that impact the success of the startup the most are the number of milestones crossed, the total amount of funding that the startup received, whether it received help from a top 500 company and, to an extent, its location (mostly in the state of California). Surprisingly, these factors seem to outweigh factors like the category of the startup and if they have VC and Angel Investor backing. The RMSE of the Random Forest model beats that of the Logistic Regression model as it is much lower. This project is constrained by the availability of adequate data since a lot of the relevant information is classified/proprietary.

Keywords: startups, random forests, logistic regression, venture capital, accuracy, data science.

Introduction

With the Neolithic revolution, humans transitioned from a nomadic lifestyle to one of agriculture and eventually, domesticity. This in turn facilitated the concentration of dense populations in a geographic region, paving the way for cities. Agglomeration economies are a core concept in Urban Economics, referring to the benefits that accrue to both firms and people when they are in close proximity to one another. The easy availability of labor and capital, coupled with the reduced costs of transportation and increased ease in the sharing of knowledge and resources, leads to the development of industrial clusters. Such clusters also foster an environment that is more friendly to doing business (through economic and taxation laws, benefits & other amenities), offers better social & civic opportunities and leads to higher wages. They also help attract

and retain both talent and allied industries to the region. Examples of such clusters are: Silicon Valley and the entertainment industry in California, and the Financial District in New York City.

Therefore, location plays an important role in our analysis of the startup ecosystem. Through our data, we see that most of the startups are concentrated in four states in the US, namely California, New York, Massachusetts, Texas and Washington. Agglomeration economies also help explain why startups in these states have a higher rate of success too! In summary:

- California has 488 startups. Of these, 332 were acquired and 156 closed.
- New York has 106 startups. Of these, 77 were acquired and 29 closed.
- has 83 startups. Of these, 64 were acquired and 19 closed.
- Texas has 42 startups. Of these, 23 were acquired and 19 closed.
- Washington has 42 startups. Of these, 24 were acquired and 18 closed.

The question we seek to answer revolves around finding some of the most important factors that influence the success of startups. It is very difficult to develop models to predict the success of startups, even if you have very detailed data. With our limited dataset, we seek to develop some rules of thumb that can be used to perform a preliminary screening of startups with potential. (This is similar to how insurance companies calculate insurance premiums for it is impossible to predict the life expectancy of a person. Therefore, they use certain parameters like the individual's basic information, personal habits, family medical history etc. as features to calculate the premium). Specifically, we approach this problem from the point of view of an investor trying to unearth previous trends in the startup ecosystem in the US before forming his analysis for current and future trends. The answers we hope to find might help this investor speed up his decision-making process for good investments are quick to lose out on.

We define success as the startup getting acquired by a firm; we classify a startup as a failure if it shuts down. A startup would only be acquired by a firm if it is making profits/shows promise or is a threat to the existing firm. Thus, we chose to define success as acquisition. The first variable we take into account is the location of the startup. Here, we only look at the state and not at the city or the exact location within the city to avoid repetition. We then look at the type of the startup as companies in certain fields like software have a high demand for their products and so find it easier to set up shop and sell. Further, these companies are easier to set up than brick-and-mortar companies as they have fewer hurdles to cross with respect to infrastructure and supply chain management. We then look at the number of milestones that a company has crossed to check it impacts the chances of the startup betting acquired. We also look at whether these companies have received any sort of aid from a top 500 company. Finally, we analyze the financials of the startup; the total amount of funding (in USD) that it has secured, whether or not it has Venture Capital backing and whether or not the startup received help from an Angel Investor. These financials serve as proxies of the confidence shown in the startups by industry experts.

Methodology

In this paper, we have used 3 classification models, namely logistic regression, decision trees and random forest model. In order to fit these models, we have used data from the startup database. The database consists of 923 observations (startups) and 49 variables. However, these startups have a very broad categorization, so we decided to create 5 categories within which these startups were placed, namely Science, Knowledge, Internet, Entertainment, and Others. The dataset has clustered different startups based on their industry and the state they are located in.

Category Buckets Explanation

1. is_science: Software, semiconductor, mobile, medical, health, hardware, cleantech, biotech, automotive, manufacturing

2. is_entertainment: Social, photo_video, games_video, music, public_relations, messaging, travel, sports
3. is_knowledge_service: Education, consulting, analytics, news, finance, advertising, hospitality, fashion, real_estate, transportation
4. is_internet: Web, search, network_hosting, enterprise, ecommerce, security
5. is_other: other

This data, obtained from Kaggle, has then been split into training and testing sets. The training set which uses 80% of the dataset has been used for building models, while the rest which comprises the testing set has been used to check the accuracy of our models. After splitting the data into training and testing sets, we fitted a logistic regression model using all of the features without any interactions into the training data. The logistic regression model is the base model used, which yields a 72.29% accuracy. Then, the model has been used to make predictions on the testing set. In order to test for out-of-sample accuracy, we created a confusion matrix with a threshold of 0.5.

One of the major reasons we used decision tree models is because they are easy to interpret and can automatically detect non-linear relationships and interactions. However, the decision trees lack prediction accuracy. We can get a largely different decision tree with different data. If we split out data into multiple training sets, the structure of the tree will differ significantly for each training set. In order to avoid this limitation, we can aggregate across several decision trees. Aggregation techniques like random forest helped us improve our prediction accuracy significantly. In our random forest model, "status" was used as a variable. With every prediction method, we could test for the percentage of correct predictions and decide the best method by choosing the method with the highest percentage value.

Results

Logistic Regression: Baseline Model

```
##
## Call:  glm(formula = status ~ has_VC + has_angel + is_top500 + funding_total_usd +
##     is_internet + is_entertainment + is_knowledge_service + is_science +
##     is_other + is_CA + is_MA + is_NY + is_TX + is_WA + milestones,
##     family = "binomial", data = startup_train)
##
## Coefficients:
##          (Intercept)               has_VC            has_angel             is_top500
##           -3.648e+00           -4.156e-01           -5.164e-01            1.229e+00
##    funding_total_usd          is_internet      is_entertainment  is_knowledge_service
##            1.255e-09            2.173e+00            1.997e+00            2.161e+00
##           is_science             is_other                 is_CA                 is_MA
##            2.272e+00                   NA            4.567e-01            8.595e-01
##                is_NY                is_TX                 is_WA           milestones
##            6.916e-01           -2.568e-01           -4.295e-02            5.956e-01
##
## Degrees of Freedom: 691 Total (i.e. Null);  677 Residual
## Null Deviance:       893.3
## Residual Deviance: 742.3     AIC: 772.3
```

The variables `has_angel`, `is-top 500`, `funding_total_usd`, `is_TX`, `is_WA` and `milestones` are not significant because they do not have a linear relationship with the dependent variable i.e. `status`.

```
##    predict_reg
##       0   1
##   0  46  40
##   1  17 128
```

```
## [1] "Accuracy = 0.753246753246753"
```

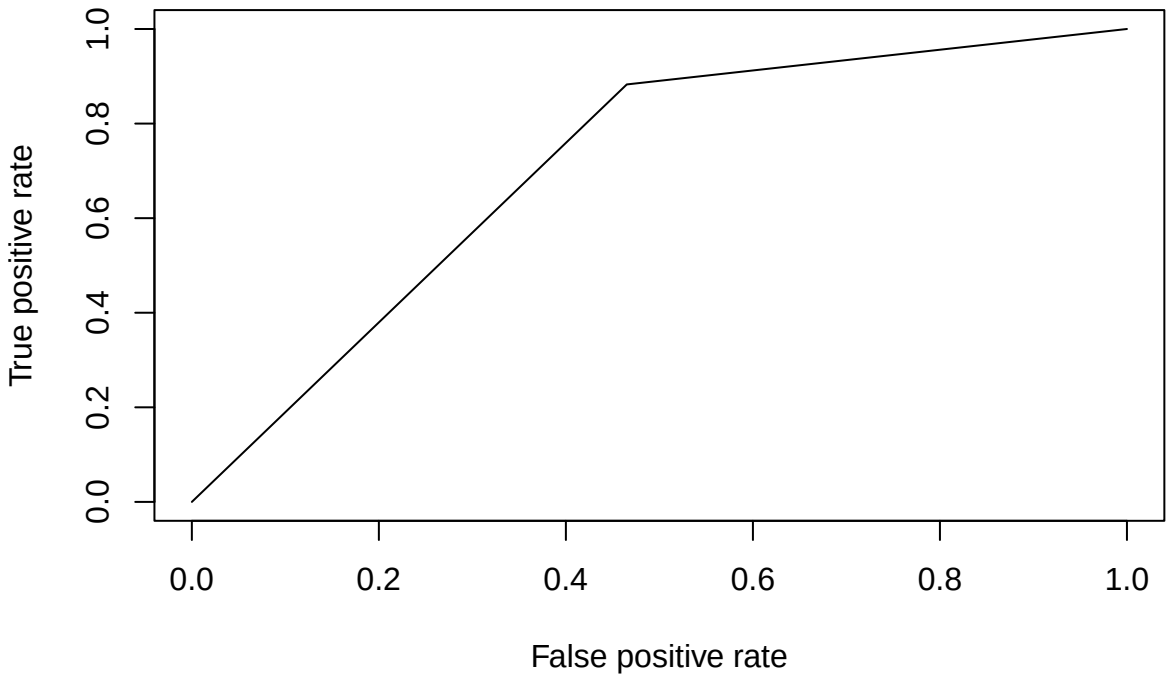
Given the decent accuracy, the model is a good predictor of true positives. However, several variables are not significant.

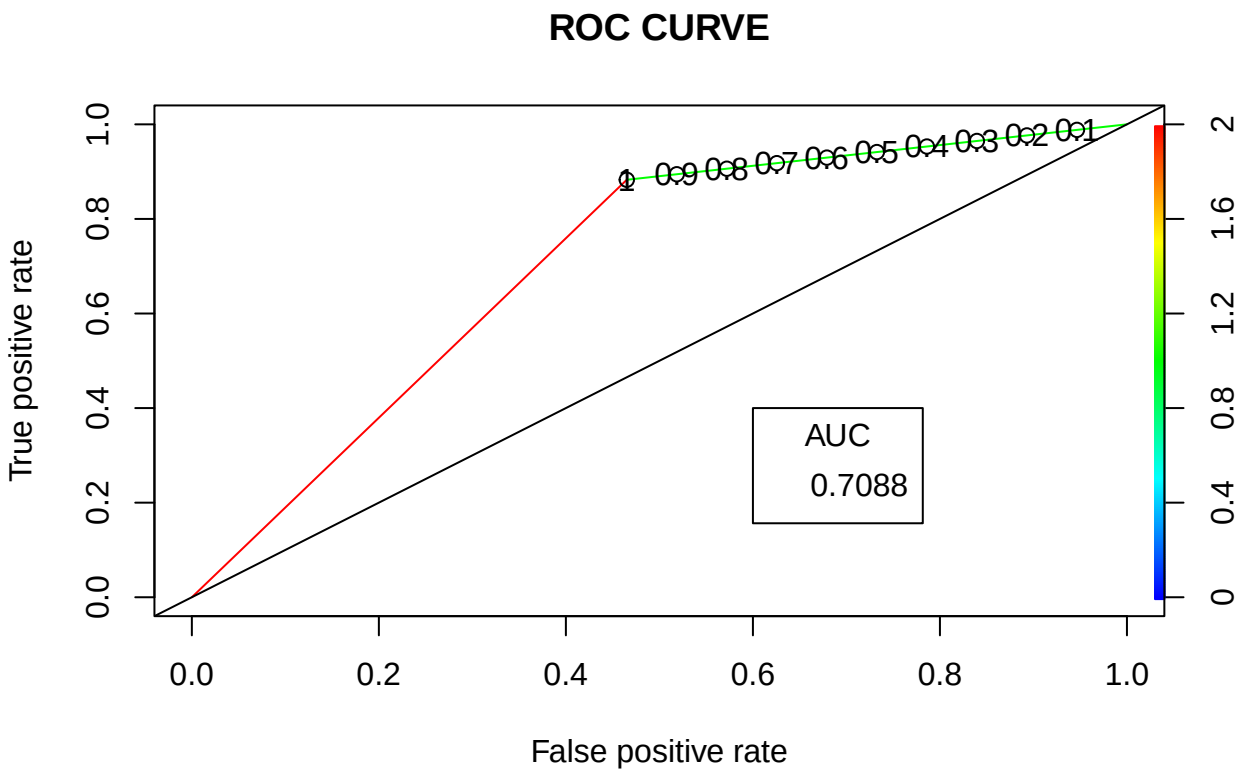
```
## [1] 1.025874
```

The RMSE for the logistic regression is rather poor, as seen directly above.

```
## [1] 0.7088212
```

The model is an above-average predictor of accuracy given the area under the curve (auc) shown above.





Random Forest

We arrived at the conclusion that a random forest model was a consistent winner when it came to performance relative to the other models tested. After all, random forests are the benchmark standard for supervised learning techniques and have the added benefit of being remarkably easy to implement.

```
conf <- startup.forest$confusion
# conf

##          acquired closed class.error
##acquired      439     43  0.08921162
##closed        132    124  0.51562500

## Corresponds to roughly 76.28% accuracy
```

The confusion matrix presented above corresponds to an accuracy of approximately 76% which is the highest achieved so far across the models we have tested.

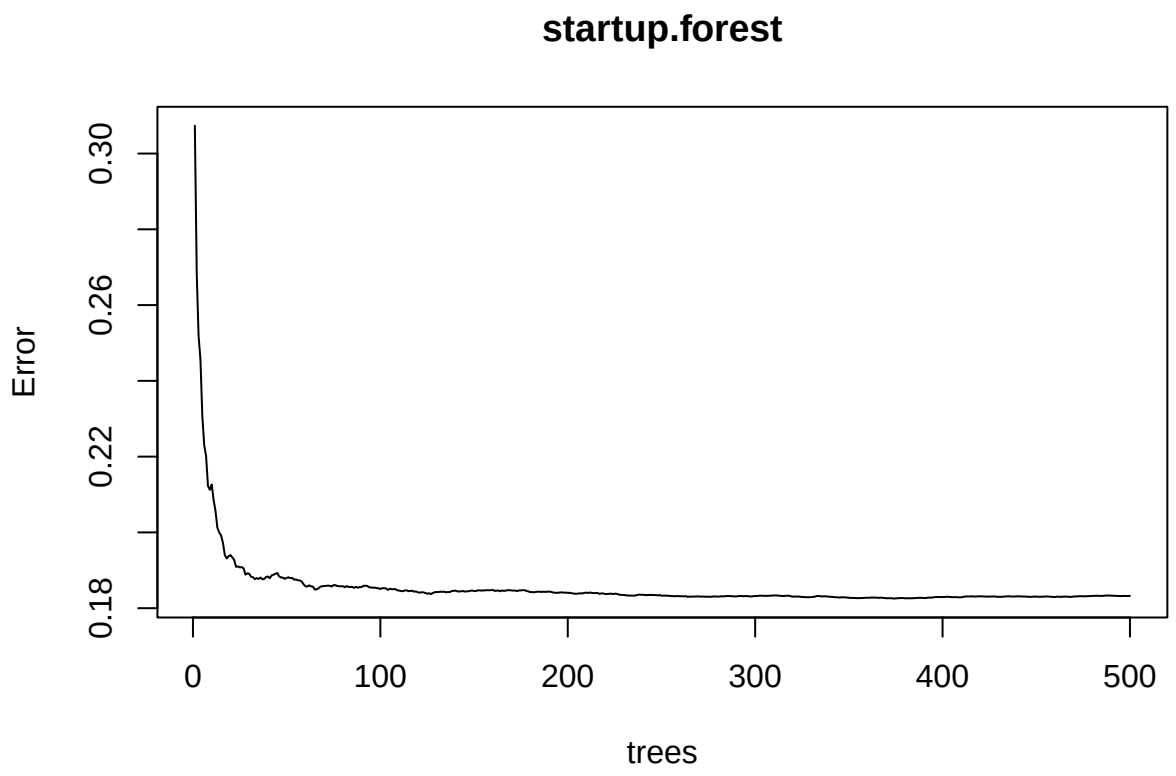
```
modelr::rmse(startup.tree, startup_test)
```

```
## [1] 0.4693375
```

```
modelr::rmse(startup.forest, startup_test)
```

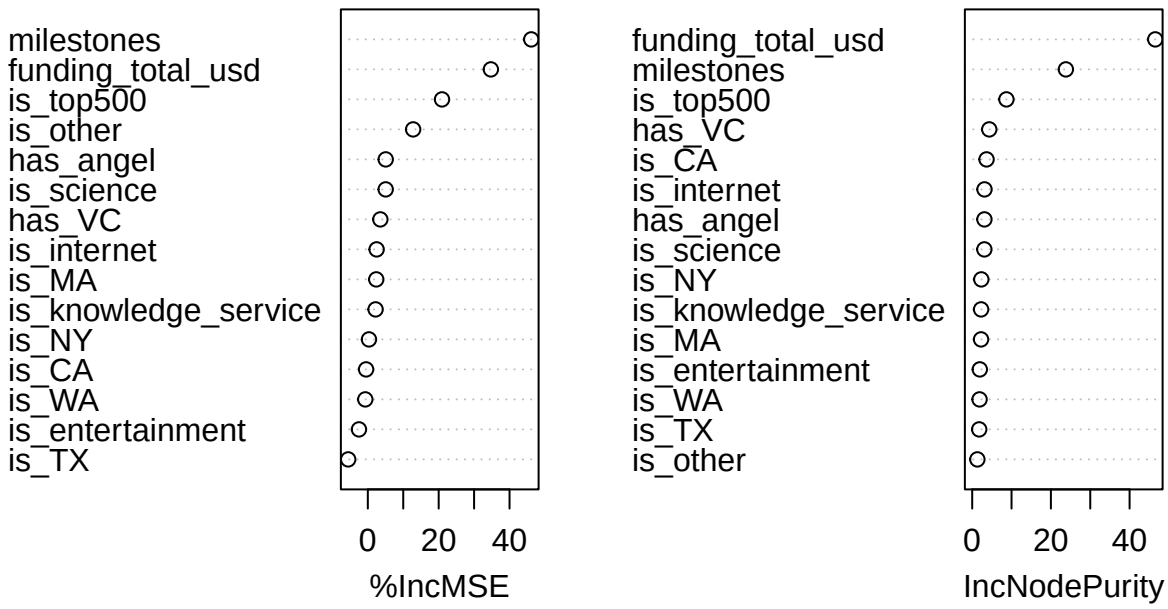
```
## [1] 0.4167074
```

Furthermore, if we observe the RMSEs of the random forest model far outperforms the aforementioned RMSE of the logistic regression, and further confirms our decision to use a random forest model.

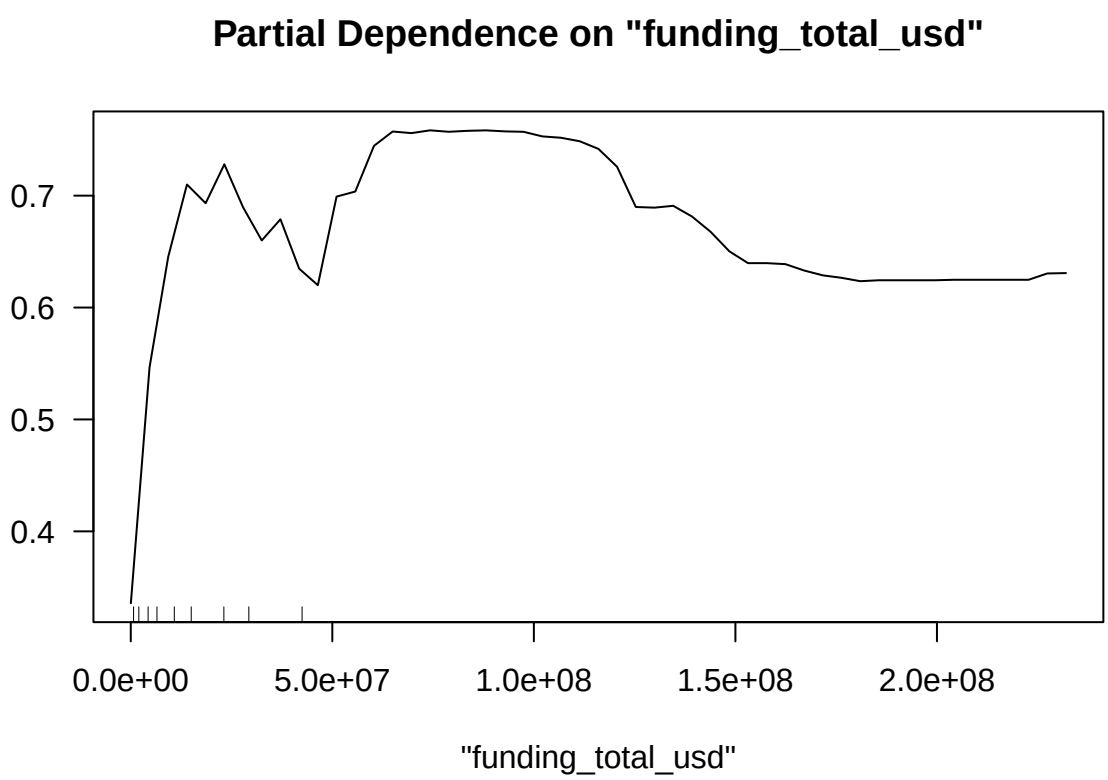


The plot above shows out-of-bag MSE as a function of the number of trees used. In each case we see that the number the error initially decreases rapidly at low numbers of trees used but then rapidly decreases and plateaus. Overall, once we exceed over 100 trees, there is very little appreciable change in the error.

In order to ascertain which variables had the strongest effects on whether a startup would be acquired, we introduced the variable importance plot made from our forest model.



In the figure above we notice two plots. The left plot tracks the mean decrease in accuracy that the model suffers when a given variable is removed, ceteris paribus. The right hand plot shows the mean increase in node purity by variable. The variable importance plot thus allows us to choose variables that have the strongest effect of the probability a startup is acquired. We thus single out `milestones`, `funding_total_usd`, and `is_top500` given their high contributions to model accuracy and high node purities which are indicative of their explanatory power.



We find that the benefit from funding increases rapidly until around 20 million, then falls slightly to a trough at around 50 million before again increasing, and finally gradually decreasing to a plateau.

In the figure above we see the relationship between `milestones` and probability of acquisition. A "milestone" can be defined in a variety of ways e.g. developing minimum viable product, getting your first customer etc. While these milestones vary somewhat across different startups, it is reasonable to assume that the number of milestones reached provides a good proxy for the momentum of a startup, and will be important to investors. In the figure above we observe that acquisition probability increases and then plateuas. The behavior seen here calls to mind diminishing marginal returns.

Conclusion

Returning to the variable importance plot, we see that belonging to one of the category "buckets" e.g. `is_science`, `is_entertainment` etc. seems to not be as significant as other factors. Indeed these variables fall rather low on the scales of %IncMSE accuracy and IncNodePurity. A possible reason for this is that it is not so much the "type" or "category" of startup that matters for eventual acquisition, but rather financial factors such as the ability to secure funding and meet specific milestones. Thus, we can conclude that in general, it may be the case that startup investors are "sector-agnostic" i.e. they place less weight on a startup's "bucket" relative to more concrete indicators of financial performance such as funding.

In the beginning for this project, we hypothesized that agglomeration economies would play a critical role in the success of startups. Ergo, we would expect that an early-stage company's geographic location would be very significant in determining its eventual success. Indeed we saw that variables such as `is_NY` and `is_CA` ranked rather highly in terms of %IncMSE and IncNodePurity, lending some credence to our hypothesis. However, location is not the whole story. Many other variables outweigh any location indicators. A possible explanation for this is that as the world economy becomes more interconnected through the internet, being in close proximity to VCs and other startups is slowly losing relevance. It would be interesting to see whether the recent COVID-19 pandemic lessened the significance of location even further given the massive transition

to remote work. Granted, the pandemic is outside the scope of our data, however this topic may be worth exploring in a future study.

Another interesting finding that was briefly touched upon in the *Results* section is how the partial dependence of acquisition on `milestones` exhibits diminishing marginal returns. Furthermore, roughly 20 and 60 million seems to be the optimal level of funding for a startup, beyond which the probability of acquisition decreases. This finding may give some insight into how VCs evaluate firms. It seems that in the investor's estimation, there does not exist a linear relationship between the aforementioned factors and acquisition. Instead, it would appear that VCs have a certain benchmark level for funding and milestones reached they want to see. Past this, the significance of further milestones or funding becomes less relevant and VCs might turn their focus to other factors.

Finally, we note that while this dataset includes many variables, there are several other factors that influence the chances of success of a startup. Many of these are not mentioned in our dataset and so impose a constraint on the model and it's accuracy.

Limitations

This project is constrained by the lack of availability of adequate data in the dataset. There are only 923 observations in this dataset. Further, there are certain crucial variables that are not available to us. We would have liked to view this problem from the perspectives of the government, the investors and the entrepreneur.

Information about the founders (the total number of founders, their age, their age, education, the ranking of the university from which they graduated, their previous work experience) would greatly help develop the framework from the perspective of the entrepreneur. Details on the tax reforms per state during 2000 to 2013, the number of employees hired by the startup, the incentives provided to the employees, the revenue growth experienced by the startup (yearly and overall) and the social relevance of the company would help approach the problem from the point of view of the government too. Details regarding the data orientation and the pivots of the firm would also aid in assessing the leadership and chances of success of the company.

References

1. Diego Camelo Martinez (2019). Startup Success Prediction in the Dutch Startup Ecosystem. Delft University of Technology.
2. Cerme Unal, Ioana Ceasu (2019). A Machine Learning Approach Towards Startup Success Prediction.
3. Daniela Santos da Silva (2016). Portuguese Startups: A Success Prediction Model.
4. Gilles Duranton and William R. Kerr (2015). The Logic of Agglomeration. The Wharton School of the University of Pennsylvania.
5. Urbanization and the Development of Cities. Lumen Learning.
6. Ahmad Takatkah. How VCs and LPs Data Science to Make Informed Investment Decisions. Medium.
7. A Machine Learning Approach to Venture Capital. McKinsey & Company.

Appendix

`milestones`: number of milestones met

`funding_total_usd`: Total funding received by a firm in USD

`is_CA`: firm based out of California

`is_TX`: firm based out of Texas

`is_MA`: firm based out of Massachusetts

`is_NY`: firm based out of New York

`is_WA`: firm based out of Washington

`Is_top500`: help received from a top 500 company

`Is_science`: Belongs to the science bucket

`Is_knowledge_source`: Belongs to the knowledge source bucket

`Is_entertainment`: Belongs to the entertainment bucket

`Is_internet`: Belongs to the internet bucket

`Is_other`: Belongs to the other bucket

`has_VC`: firm has received help from a VC

`Has_angel`: firm has an angel investor