Final Project Report

Introduction

India is seeing a tremendous growth in start-up companies. The opportunity to perform analytics on the data associated with these start-ups is huge. The outcomes of this analysis are very useful and valuable in decision making for investors.

1. A lot of investors in India are beginning to fund start-ups and growing companies. The following statistics obtained from a fin tech website indicate the Venture Capital (VC) deals in India over the last 5 years

Date	No. of Deals	Aggregate Deal Value ($M)	Average Deal Size ($M)
2017	938	10,700	17.9
2018	964	9,900	13.6
2019	1127	14,700	18
2020	1088	11,200	14.6
2021	1376	36,100	32.9

2. Studies have also shown that 90% of the start-ups fail in India within their first 5 years. This indicates a high level of risk involved when it comes to investing money.

https://trak.in/tags/business/2017/05/22/indian-startup-failure-ibm-study/

Whether it be Venture Capitalists or Angel Investors, it can be of extreme importance around the globe for investors to know the probability of success or failure based on some criteria such as growth, acquisition, etc. using historical data and based on the attributes of the company such as Business Domain, Sales, Gross Margin, Daily Active Users, Funding rounds, Location of the Start-up company, etc.

Problem Statement and Objective

Predicting whether a start-up company will succeed or not, success defined by Acquisition/IPO and failure defined by Closed. Our project aims to provide investors with a model that helps them to invest money with confidence.

Data set Description

Data Set Link - https://www.kaggle.com/datasets/justinas/startup-investments?select=funding_rounds.csv

The Data set contains several files. For the purposes of our analysis, we will be using the following files -

- **objects.csv** - It contains contains descriptive information about each company. This includes -
  - Category - The domain in which the company operates. Example - Software, Security, Biotechnology, E-commerce, etc. (Categorical)
  - Country - The country where the start-up originated from. (Categorical)
  - Relationships - The number of connections that exist between all entities that engage in commerce. (Numerical)

  - Milestones - The number of significant moments of accomplishment that mark a company's development and growth. (Numerical)
  - Funding Rounds - The number of funding rounds. (Numerical)
  - Twitter Presence - Whether the company has a twitter account and engages with their customers. (Binary)
  - Total Amount of Funding Received (Numerical)
- **people.csv** - Information about a person and the organization they are affiliated with. The primary key (Person ID) will be used in other files and will act as a foreign key.
- **relationships.csv** - A file that joins people and objects and contains information regarding the person's title in the company. Our focus on this analysis will mainly be on the number of Founders. (Numerical)
- **degrees.csv** - A foreign key is placed which indicates the Degree obtained by each person. In our analysis, we will be focusing on -
  - Number of MBA Graduates (Numerical)
  - Number of Masters Graduates (Numerical)
  - Number of PHD Graduates (Numerical)

  It is important to note that these values are calculated for the Top Level Management positions only (Example - VP, SVP, CEO, etc.). Therefore, the numerical quantities may be on the lower end when compared to that of the whole company.
- **acquisitions.csv** - Information about the acquired company, the acquiring company, and the price at which the company was acquired
- **ipos.csv** - Information about the IPO Date, Amount of money raised, and Valuation.

Approach and Methodology

Data Pre-processing and Feature Selection

We noticed that the data required some pre-processing to be brought to the correct format to help assist in performing EDA as well as building models.

The following steps were performed during the course of data pre-processing:

1. The data set contained data that went all the way back to the 1930s. For our analysis, we considered only records post the year 1980.
2. We filtered the records with status - 'Acquired', 'IPO' or 'Closed'. Success is defined by 'Acquired' or 'IPO' while Failure is defined as 'Closed'.
3. We filtered the records where the entity type was a 'Company'.
4. Some of the categories (example of categories: e-commerce, software, biotech, etc.) had missing values. We could have grouped these data points into a separate category called 'other', however this group already existed. Due to this, we eliminated missing values as the fact that placing them in the 'other' category may lead to creating/imputing incorrect data. Example: A missing value for an originally Software company might be placed into the other category.
5. We had missing values for the Countries as well. We removed these data records so as to not provide inaccurate information to the model and to not hinder EDA.
6. Instead of including the twitter username for each company, we modeled it as a Binary Variable which indicated Social media presence/absence.
7. We convert the category and country into factor type in R.

8. The response variable is converted into a binary variable. Success is indicated by 1, while Failure is indicated by 0.
9. We read in the relationships and degrees data and only filter based on the companies that exist in our original company data frame. We then merge them based on the person ID which acts as the foreign key.
10. We perform some pre-processing to find the following four features -

a. No. of Founders
b. No. of MBAs
c. No. of MSs
d. No. of PHDs

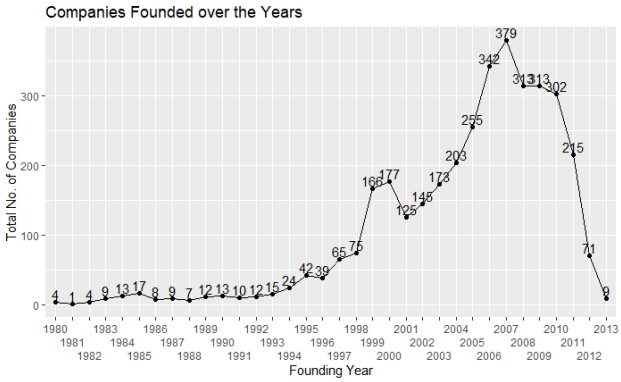
The reason for including these features is to observe the impact of education of the top management of a company and whether it contributes to success.

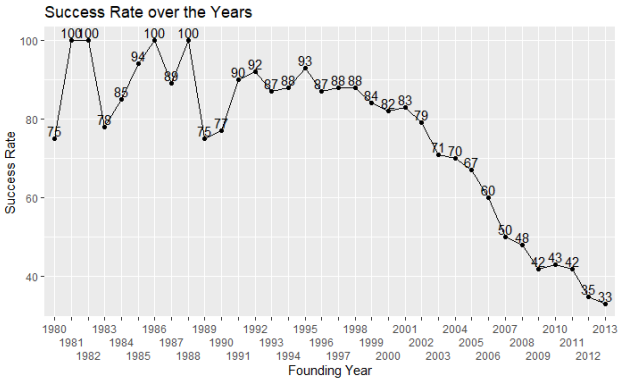
11. We perform a left join based on our original company data frame and the relationship-degree data frames. The ID here will refer to the company ID placed as a foreign key in the relationship-degree data frame.
12. We then model the countries as to whether they belong to a top 5 country or not, top 5 indicated by the no. of start-ups emerging from that country and whether they were top contributors towards the start-up population or not.. The total no. of dummy variables increases if we take the country categorical variable as is, therefore we engineered this feature to be a flag.
13. We consider only records with funding rounds greater than 0 because when we consider companies with 0 funding rounds that turn out to be successful, they are self sustainable and the investors don't get an opportunity to invest in such companies. Keeping companies with 0 funding rounds would thus degrade our prediction models.
14. Some of the Funding Total values were missing. There were approximately 8% of missing values. Due to the percentage value not being below 5%, we could not consider imputation and removed the records.

Exploratory Data Analysis

Growth of Startup Culture

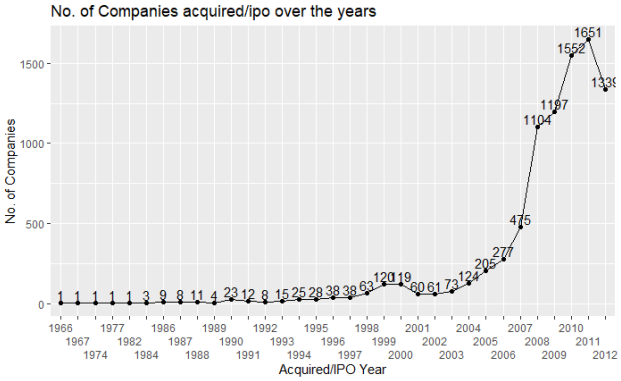
- We observe a big spike in the number of start-ups being found over the years, especially after computers became popular in the late 90's (Intel's Pentium 2 was launched) following which the internet reached the common household.
- Due to this increased competition and the hype around the start-up culture, we see a sharp drop in the start-up success rate.

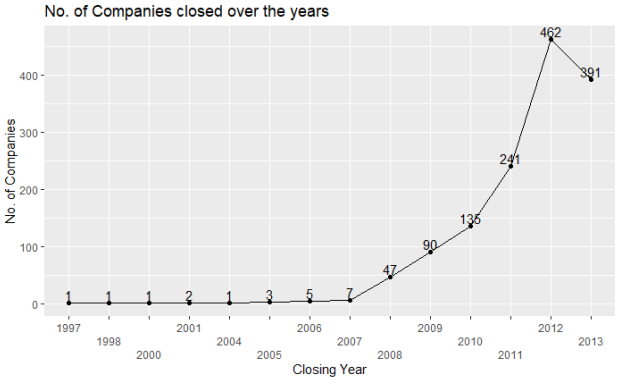




Acquisitions (& IPO) / Closures based on Year

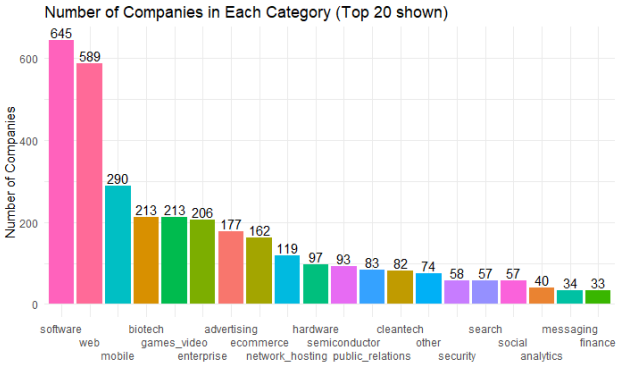
- We observe that acceptance and success stories of start-ups are increasing over the years.
- Opposing our hypothesis that the dot com bubble and 2008 economic crash should have had highest closures, we observe that closures are increasing every year, as new companies' numbers are also increasing. This can be attributed to the after effects of the recession.

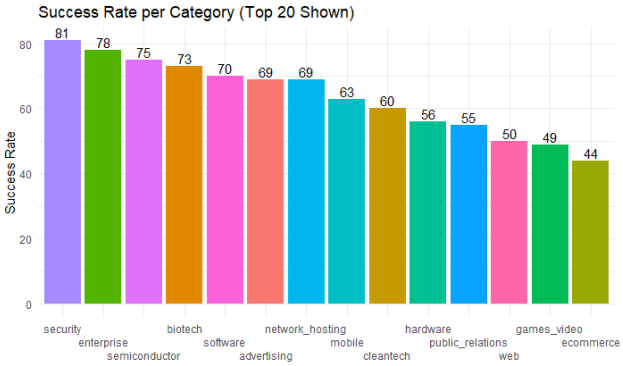




Analysis of Categories

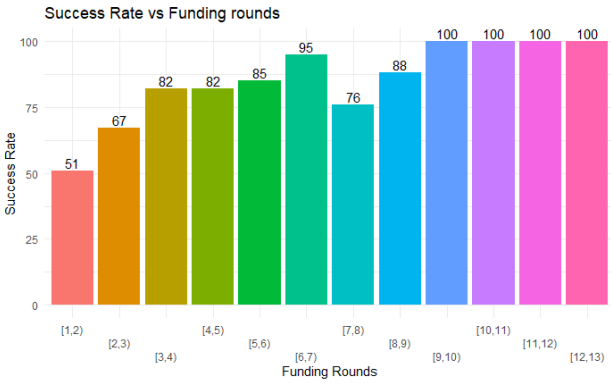
- We observe that the top two categories are occupied by Security and Bio-Tech having a success rate of 87% and 85% respectively.
- 50% of the successful start-ups came from the sum of the following four segments - Software, Web Technology, Mobile and Video Games.

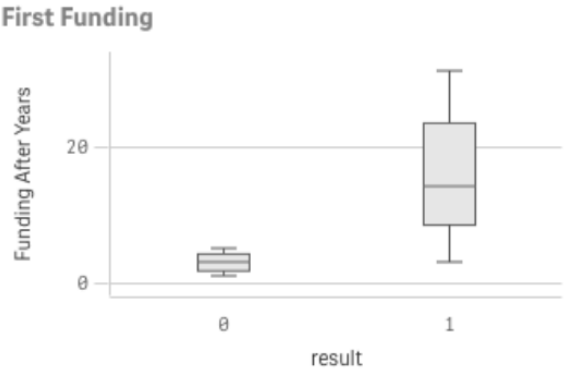




Analysis of Funding Rounds

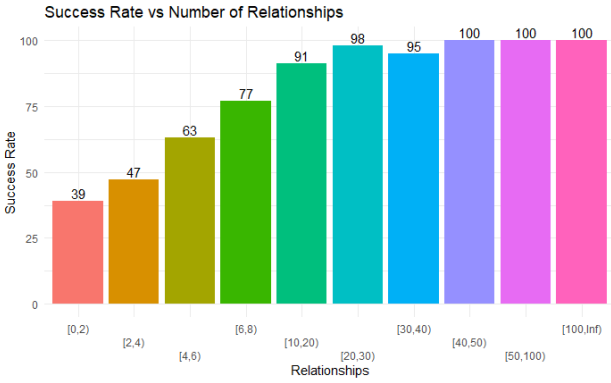
- We can clearly observe that more than half the companies that could not secure more than one round of funding saw closure.
- Companies with 9 or more rounds of funding lead to acquisition/IPO.
- From the box plot, we can see that start-ups that saw closure were funded within the first 5 years of their founding date and more than 75% of successful start-ups were first funded after 10 years of their founding, i.e. after they had established themselves.

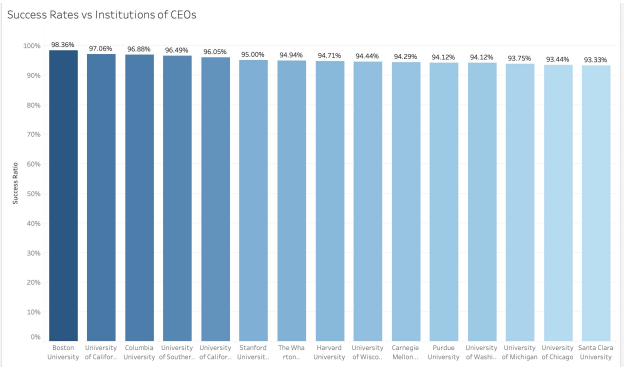




Analysis of Relationships and CEOs

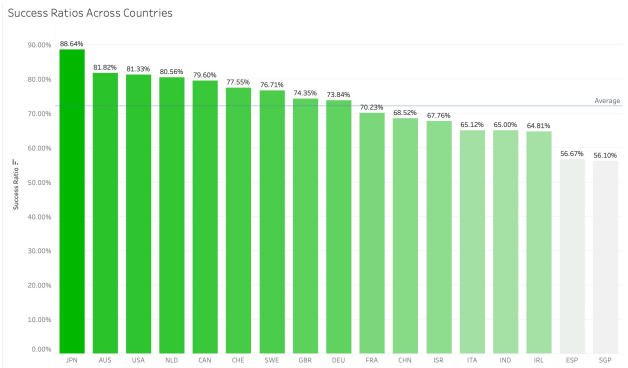
- We see a correlation between number of people associated with a company and its success.
- Alma Mater of CXOs and Founders also affects these relationships and success.
- Educational background of founders and leaders is also key in ensuring a start-up succeeds.

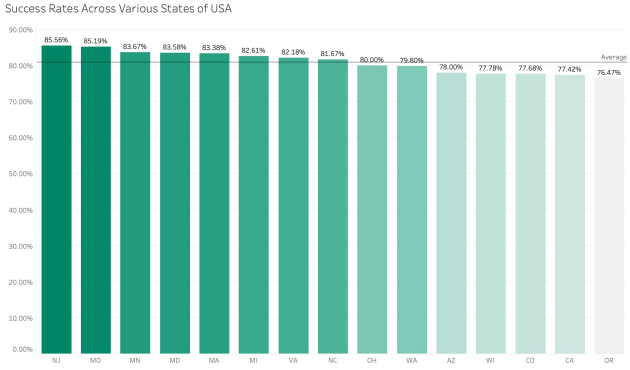




Analysis of Geography

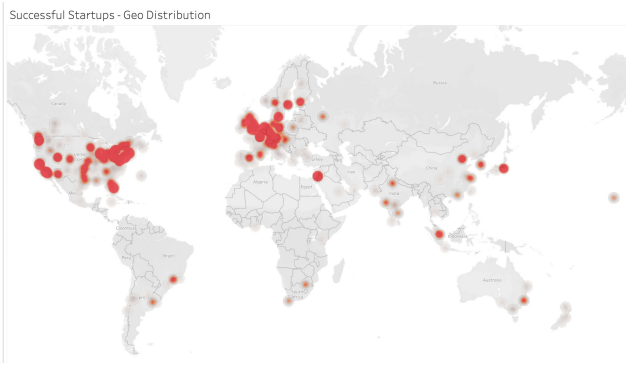
- Geographical factors have an important role to play in the success of a startup.
- Policies favoring the success of start-ups vary in each country.
- We expect a concentration of successful start-ups in USA, Europe and Israel.
- Japanese start-ups also have high success rate, which was a surprising find.
- In the United States, NJ seems to have the highest success rate.





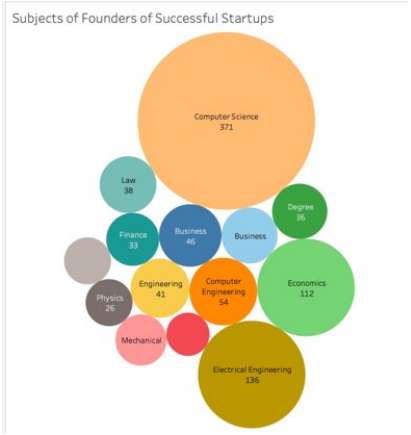
Successful Start-ups visualized on the World Map

- We observe a high concentration in the United States as well as Europe.
- In the United States, we can further observe a cluster of points on the east coast.



Analysis of Degrees

- We observe that the degrees are dominated by the Computer Science.
- Nearly 25% of Founders of successful start-ups have a STEM Degree.
- This correlates with the fact that we observe a large number of software and web start-ups.



Models

Logistic Regression

- Logistic regression is a probability regression model and is known to be very effective in Binary Classification problems.
- For our problem statement, the output of the model is the probability of success for a start-up company.
- We further went on to perform analysis by setting different threshold values to define the outcome of classification, i.e. whether say 70% probability was required to classify a company as successful or 30% was sufficient.
- We observed that the peak for accuracy is observed at a 0.5 threshold.

Confusion Matrix

		Actual Values	
		0	1
Predicted Values	0	150	65
	1	72	248

Measures	Value
Accuracy	74.39%
Sensitivity	79.23%
Specificity	67.57%

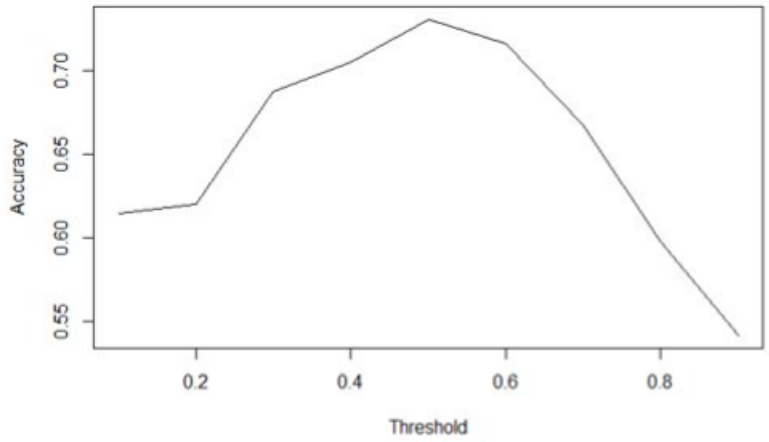


Figure 1: Threshold vs Accuracy

Decision Trees

- The decision tree is a statistical model that is useful in both regression and classification problems.
- At each node, DT minimizes the Gini coefficient and thus splits the data into purer class subsets with the class leaf nodes at the bottom of the decision tree.
- Best Hyper Parameters -
  - minsplit = 200 - The minimum of number of data points in a node that can be split
  - minbucket = 50 - The minimum number of data points that can be present in a bucket.
- From the results of the Decision tree, we can observe that Funding Total, No. of Milestones and No. of Masters' Degrees held by the top management in the company are important splitting variables.

Confusion Matrix

		Actual Values	
		0	1
Predicted Values	0	108	29
	1	102	296

Measures	Value
Accuracy	75.51%
Sensitivity	91.08%
Specificity	51.43%

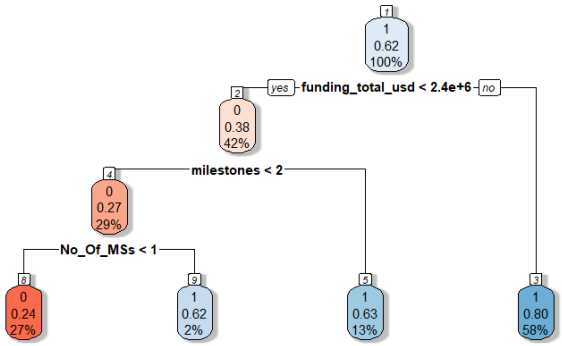


Figure 2: Default Model Hyper parameters

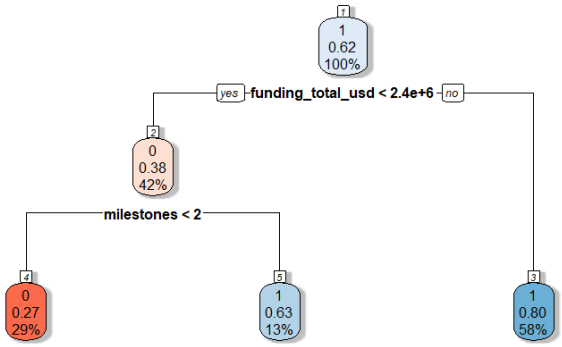


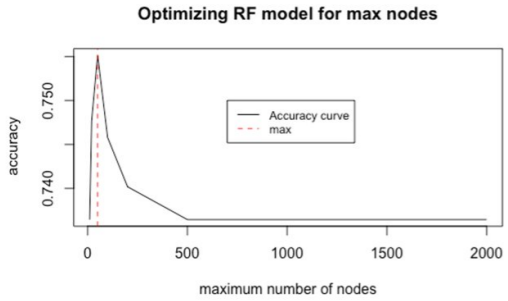
Figure 3: With Hyper parameter tuning

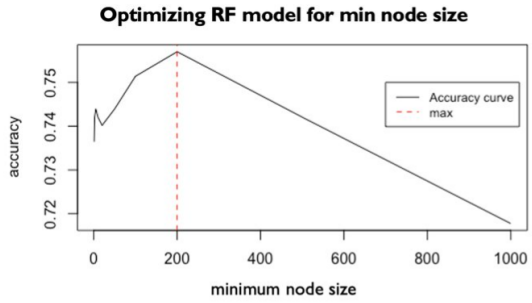
Random Forest

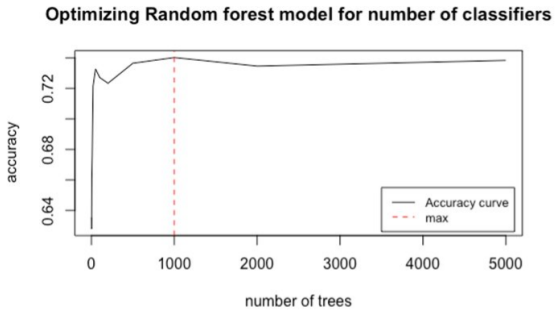
- The random forest model is a boosting algorithm, based on an aggregation of multiple decision trees.
- It helps mask the shortcomings of Decision trees, wherein it tends to over fit the data.
- Best Hyper parameters:
  - No. of Estimators - 1000
  - No. of splitting variables per node - 3
  - Maximum Nodes - 50
  - Node Size - 200

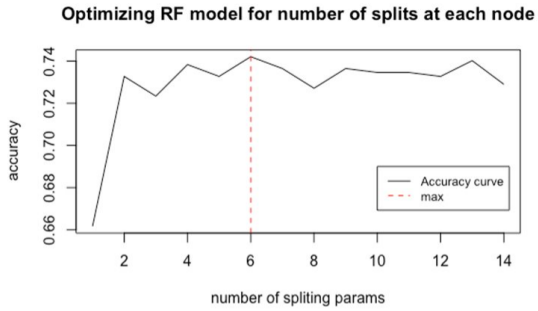
		Actual Values	
		0	1
Predicted Values	0	115	29
	1	95	296

Measures	Value
Accuracy	76.82%
Sensitivity	91.08%
Specificity	54.76%







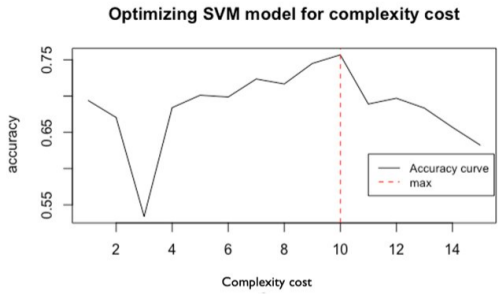


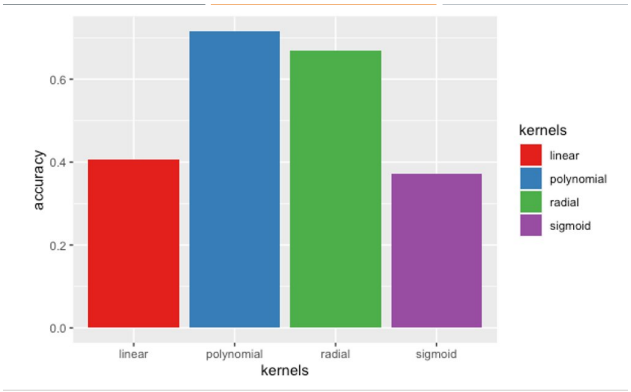
SVM

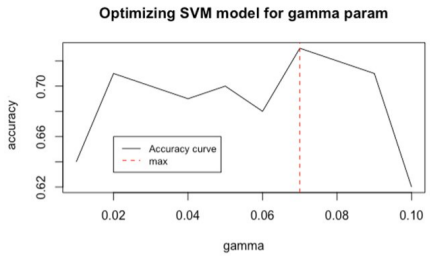
- SVM is one of the most popular statistical learning approach used in classification problems.
- It is highly robust due to the fact that it can easily be extended to multiple dimensions.
- For our problem since the data is not linearly separable, modified kernel methods provided a good way to define a classification boundary.
- Best Hyper parameters:
  - Kernel - Polynomial
  - Degree - 3
  - Gamma - 0.09
  - Cost complexity - 10

		Actual Values	
		0	1
Predicted Values	0	121	49
	1	89	276

Measures	Value
Accuracy	74.21%
Sensitivity	84.92%
Specificity	57.62%







KNN

- By varying the parameters of k, we obtained the highest accuracy for k=9.
- Intuitively this means that each point in the test set is classified based on its 9 closest neighbors, and for this value of k, we obtained the highest accuracy.
- Best Hyper parameters
  - k - 9
  - Distance - Euclidean

		Actual Values	
		0	1
Predicted Values	0	128	55
	1	82	270

Measures	Value
Accuracy	74.39%
Sensitivity	83.08%
Specificity	60.95%

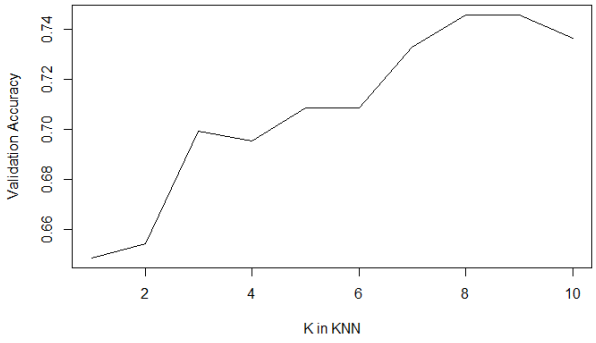


Figure 4: K values vs Accuracy

Compilation of Final Results from All Models

- We observe that Random-Forest Model outperforms all the other models in terms of accuracy. This is followed by the Decision Tree. This seems to indicate that aggregation of trees, i.e. Bagging seems to perform well.
- Although SVM produces the least accuracy, it is not too far off when we compare it to the rest of the models.

Model	Accuracy
Random Forest	76.82%
Decision Trees	75.51%
Logistic Regression	74.39%
K-Nearest Neighbors	74.39%
SVM	74.21%

Addressing Misclassification Cost

We additionally addressed a case where we apply a penalty when our model predicts a start-up with actual value as closed as successful. The Cost is calculated by considering the fact that predicting a closed start-up as successful is 5 times worse than predicting a successful start-up as closed. The reason is simply because we would rather take a missed opportunity over losing a lot of money.

Model	Misclassification Cost
Random Forest	240
Decision Trees	247
SVM	334
K-Nearest Neighbors	357
Logistic Regression	397

Analysis of Results

- Our results suggest that this problem is not a linearly separable problem.
  - Various factors like domain and number of founders are significant in determining start-up success.
- Total funds received by a start-up is a good indication of its success.
- Education background of people in key roles is important
  - Nearly 83% of failed start-ups had no key person with a business or related education
  - This number was nearly 40% for successful start-ups
- Geographical influence is also observed
  - 38.6% - Success rate in start-ups outside of top 5 geographical regions
  - 34.67% - Failure rate in top 5 geographical regions
- Higher number of founders reduce the risk on an individual.
  - Risk of mismanagement in areas like marketing, accounting and development gets distributed
  - Co-founders help in avoiding tunnel vision
- Companies in the internet and web domain have a smaller prototype to product cycle. This leads to -
  - Easy pivot opportunities
  - Requirement of low capital investment
- High proportions of external investment from angel, and/or equity, indicate good market demand.
  - Financial and operational risks get audited by investors and are flagged early.
- Having a business education helps in keeping the business afloat.
- Countries with a nurturing start-up environment identify good ideas early
- More capital is available in the market for companies and individuals
- Conducive government policies and economics enable and increase the risk appetite of corporations.

Future of the Project

The scope of this project given today's trends is very high. There are several improvements that can be made for the future while trying to predict the success of a start-up. This can be based on -

- Better Data Collection Methods - In our data set, we encountered missing values. It's important to reduce missing data through imputation or regression methods, however better yet, by creating a data pipeline that will scrape the data from a reputable data source.
- Feature Engineering and Selection - For the purposes of our analysis, we used features we considered important based on our data set and EDA. However, new features can be used such as Political, Economic and Environmental Factors.
- Models - we can also try Bayesian optimization techniques and Bagging techniques. Neural networks can also be used to perform prediction.