**INFO 442 Project Report**

**Members - Kayla Savage, Quan Ta, Steven Nie, Tapasya Sharma**

1. **Problem Statement**

As context, according to the Small Business Administration (SBA), many small businesses' growth and success is vital to the United States economy. Though this is the case, failure rates are high. Specifically, as of 2019, startup failure rates are around 90% and about 21% completely fail by the first year.([Investopedia](https://www.investopedia.com/articles/personal-finance/040915/how-many-startups-fail-and-why.asp)) There are various reasons why start-ups may fail, for example there could be a lack of resources, little research, wrong market projections, and possibly bad marketing. All of these aspects contribute to the financial success of business.By making this a priority we are contributing valuable conclusions about how businesses can function specifically at the start of establishment. Our project allowed us to show various ways prediction algorithms can impact analysis and business problem solving in the finance industry.

To successfully run a business you need capital and that is where listing on the capital market comes in. Over the years more and more companies are listed on the public stock exchange in the process known as IPO or Initial Public Offering. That is when a company is first listed on a public stock exchange whose share can be traded by anyone in the market. Traditional IPO is a stringent and time consuming process that takes years to complete. However, with the recent popularity of SPAC or Special Purpose Acquisition Company, when a public shell company can take a private company public by simply purchasing it allows it to quickly go public without a traditional process. This allows for more companies to IPO even when they don't have a great business yet. So studying a company through its IPO document is a great place to get insight into the company.

1. **Data Collection and Preprocessing**

Since a similar dataset doesn’t exist or is at least not publicly available we have to create our own dataset. While the financial fundamental metrics are derived and calculated from the number reported on the SEC filing, reading and extracting these numbers from the usually written in legal language that could be difficult to read through especially there are thousands of companies and each of their filings is hundreds of pages long it is better to seek data provider that already had these data and can be piped from their database easily with an API. For that, we use Intrinio which is one of the better financial data providers with a decent Python API. I wrote the program that collects the data in a way that fits the need of our project

The program is designed to get the snapshot financial data at the time of IPO or closest to the time of IPO it does not take into account new information reported after IPO which normally will be considered too but to analyze that with a machine learning algorithm will require a time-series analysis algorithm which is out of the scope of our project. There is other information the program will get from the API namely the country, number of employees and industry to get more relevant information about a company that could affect it. The goal of our project is to see if the company will be successful or not in turns of stock price information. It holds 50% of the IPO price the company is most likely still running. Otherwise if the price fall below 50% of the IPO price it is safe to say the company fundamental has gone quite wrong and most likely won't be able to recover from it and if it is delisted from the exchange that also mostly signal the company operation had gone terribly wrong so much so it is no longer considered worth it to list on the exchange.

The data we collected contains about 1000 companies that went public from the year 2012 to 2021 with roughly 400+ considered a success and 500 considered a failure so the dataset is not too unbalanced. Additionally, there are quite a lot of missing data, I manage to fill in a few but most metrics require extensive research on its filing so it is not possible to fill in them all so we just fill it in with mean value which is not a very good way to fill missing data for this kind of data but it will have to do.

1. **Exploratory Data Analytics**

Our Exploratory Data Analysis included understanding the features, data types, and relationships between features. Key findings:

1. The shape of the data is 906 rows and 13 columns
2. The columns are:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 906 entries, 0 to 905

Data columns (total 13 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 assetturnover 906 non-null float64

1 commontocap 906 non-null float64

2 currentratio 906 non-null float64

3 ebitdagrowth 906 non-null float64

4 investedcapitalturnover 906 non-null float64

5 leverageratio 906 non-null float64

6 nnep 906 non-null float64

7 profitmargin 906 non-null float64

8 roe\_simple 906 non-null float64

9 stdebttocap 906 non-null float64

10 enterprisevalue 906 non-null float64

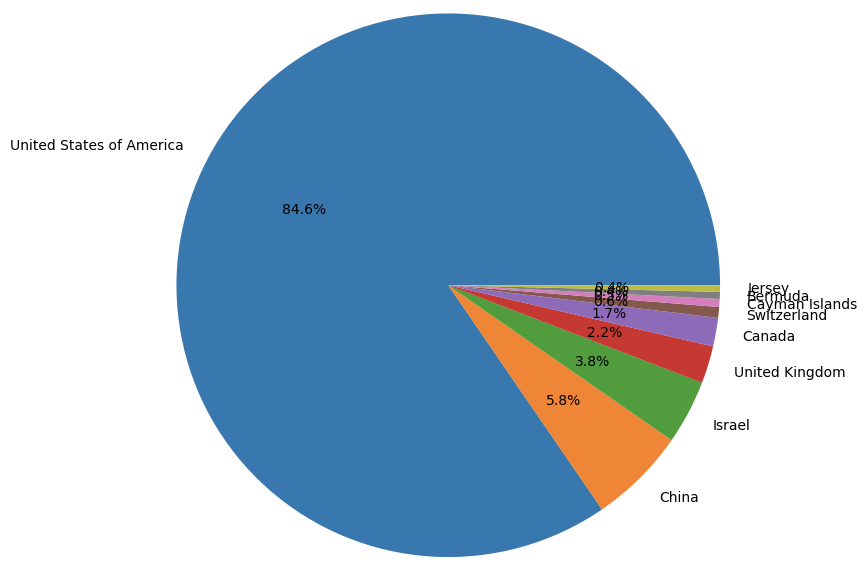
11 country\_label 906 non-null int64

12 employees 906 non-null int64

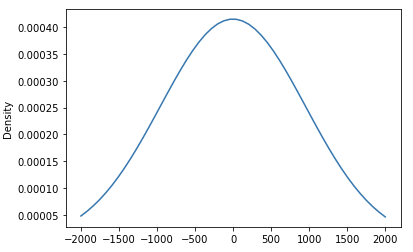
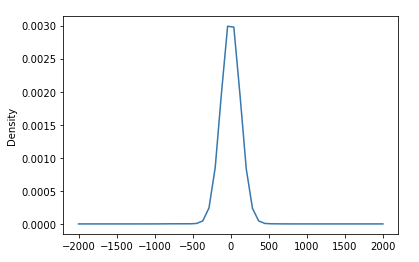
dtypes: float64(11), int64(2)

memory usage: 92.1 KB

1. Some graphical analysis



Distribution of countries

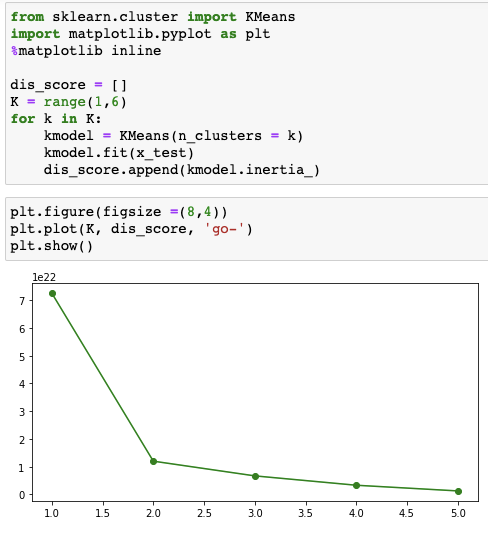
****

Distribution of EBITDA Growth for Successful Companies Distribution of EBITDA Growth for Failed Companies

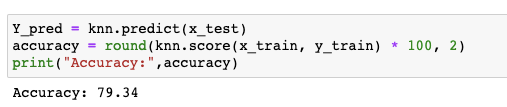
1. We have a sparse matrix and hence we set a 20% threshold that drops all rows and columns that do not fulfill that threshold
2. We also have numerous functions which delete duplicate symbols, impute missing data, remove null employees and clean the data further
3. This is also where we defined ‘success’ based on existing features
4. **Prediction Models**
5. **K Nearest Neighbors** 
   1. **Elbow Cluster Analysis**

To start, we conducted an elbow cluster analysis to find out the value of K. This allows us to quickly determine the k value.

This is the outcome, as you can see, the value of K is set at two.

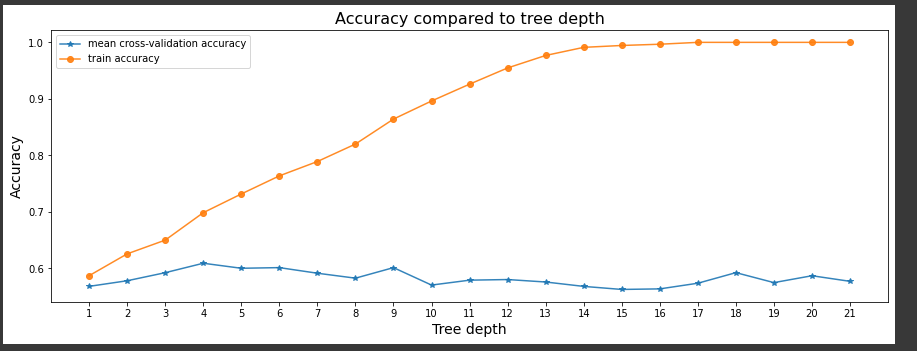


K Nearest Neighbor is a supervised machine learning algorithm. By finding the distances between each observation of data points, selecting the value of K closest to the query, it allows us to use each feature category and depict an outcome. We ran the k neighbor classifier on two different train/test data splits. One is 80/20 and the other is 70/30. We can see that there was a higher accuracy of 79% with a larger test data split.



1. **Decision Trees**

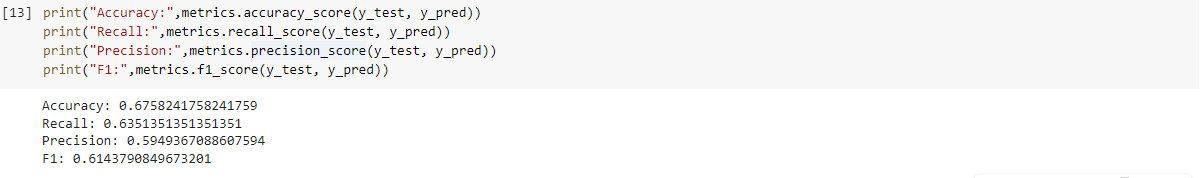
A decision tree is a machine learning model that resembles a tree-like structure used to make decisions from the root node until the leaf node until a target variable or a decision is reached. For our project, we utilized a decision tree model with a depth of 4. The depth 4 was decided through a cross-validation procedure that analyzed and compared results for Decision Trees up to depths from 1 to 20. Here are the results:



From the figure, we make the following observations:

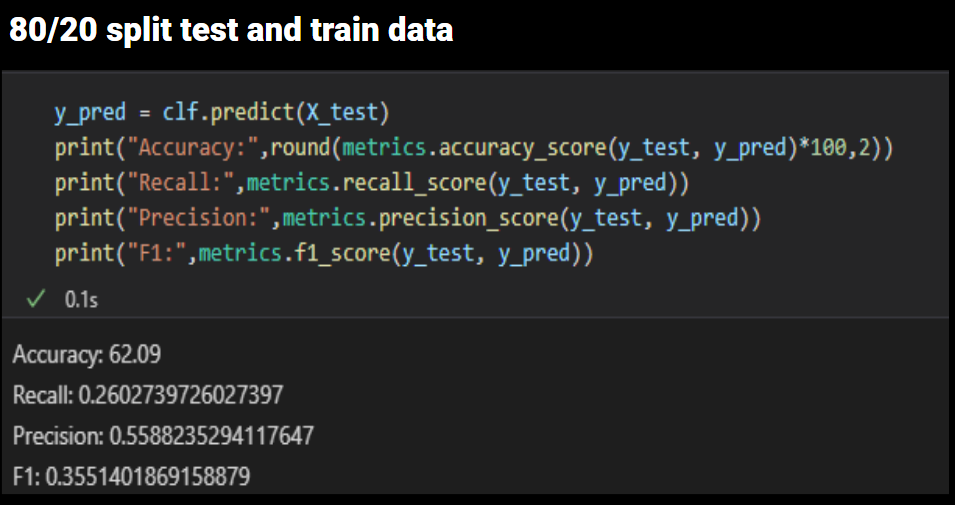
1. Depth 4 has the highest cross-validation accuracy (63.73%) with an acceptable training accuracy (69.88%).
2. From depth 13, the tree is on the verge of overfitting because the training accuracy is extremely high.
3. Tree with depth 9 is highly comparable to depth 9 with 63% cross-validation accuracy and 80% training accuracy
4. **Random Forest**

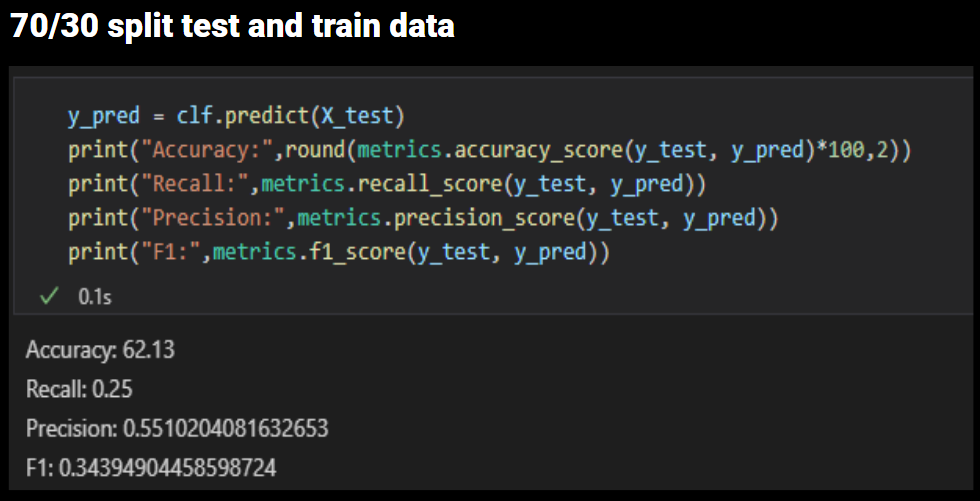
Random Forest is an ungraded version of thedecision tree generally with better performance than decision tree and here is the result:



1. **Support Vector Machine**

Support Vector Machine (SVM) is a linear model used for classification and regression problems. It is good for practical problems as well. For the project, we tried all the possible kernels and found the Radial Basis Function (rbf) one to be the best option for our dataset. To further improve the result, we ran the model two times with 80/20 and 70/30 train and test data split, and here is the result:





1. **Conclusion**
2. **Strengths of the project**
3. K-Nearest Neighbor showed to be the most accurate model which is an advantage as it is one of the most simple models to incorporate
4. Our project included the component of collecting and building a custom dataset from Intrinio which is replicable to real-life
5. Data Analysis of these data can provide great information for data-driven decision making
6. The project is a preliminary program that provides a guideline for a better future project
7. **Weakness of the project**
   1. Our dataset needs more variables for the model
   2. Our dataset consists of many missing values
   3. Our model has limited usability in the real world as our results are not as expected
   4. Our project does not find and quantify enough factors
   5. Our project may need time series analysis to gain more information and improve the models’ accuracy