St. Joseph's University

Bangalore, Karnataka

PROJECT REPORT

ADVANCE STATISTICAL METHODS

Submitted by:

JITHIN JOJI -222BDA09

DAIN JINS – 222BDA19

SHONE BENNY – 222BDA37

JOEMON PAULSON – 222BDA64

Submitted to:

JAYATI KAUSHIK

Assistant Professor

Department of Advanced Computing

St. Joseph's University

Predicting Success Rate of Startups using Machine Learning

Our aim is to create a machine learning model that can predict the success of startups by analyzing various factors such as funding, location, industry sector and previous experience. Looking at the problem we can see that it can be answered using any classification algorithm. But preprocessing the dataset for the classification model is the most important part, because if any mistake happens in this step the whole model fails. So, we actually took a lot of time to clean and preprocess the data.

The dataset we got contained company information's such as company name, url, market, country, state, region, city, founded date, first funding date, last funding date. It also had data on different investment types such as seed, venture equity crowdfunding, undisclosed funding, convertible note, debt financing, angel, grant, private equity, post ipo equity, post ipo debt, secondary market, product crowdfunding, round A-H series funding. The dependent variable here is status which were also available and segmented by acquired, operating and closed. So, it's a multi class classification.

For hypothesis purpose we defined null and alternative hypothesis for t test and chi square test. The alternative hypothesis (H1) would be that there is a significant association, while the null hypothesis (H0) would be that there is no significant relationship between the independent factors and diabetes.

The model that we made will predict whether the company will acquire, operate, or close by analyzing various important features that we found out in this project. This model can help investors and entrepreneurs make informed decisions about which startups to invest in and can also provide insights into the factors that contribute to the success or failure of a company.

Domain: Business and entrepreneurship

Problem Statement: Startup investment can be very risky due to the high failure rate of startups. People like angel investors and venture capitalists have a very high risk while they are investing in startups. To assist startup investors with their decisions, in this project we aim to find the important features that

lead to startup success and forecast a company's success with supervised machine learning methods.

Introduction

Startups are young businesses. Entrepreneurs start a company to make a promising product or service. These startups have high costs and low revenue, so they need multiple funding sources. Due to demand, this newly founded company has gained traction. A startup wants to grow quickly by filling a market need. These companies lack a corporate structure and, more importantly, sufficient funding to grow. Most of these businesses are founded by risk-takers. Many startups fail. That's why the first month matters. Startups can use seed capital for research and strategy. A startup's initial funding may come from family, friends, and other contacts with a high risk of failure.

Machine learning uses algorithms to create models that reveal patterns from data, allowing businesses to gain insights and make predictions to enhance operations, better understand customers, and solve other issues. There are numerous algorithms to choose from. These help to predict the outcomes of a start-up will be profitable or not. Learning the mapping $Y = f(X)$ to develop predictions of Y for new X is the most prevalent sort of machine learning. Our goal is to make the closest result possible, which is called predictive modelling or predictive analytics. Algorithms those are trained on historical data and any other relevant data sets may evaluate data and go through various possible scenarios at a size and pace that humans can't match, allowing them to give suggestions on the best course of action. Machine Learning algorithms assist start- up businesses in reading consumer behaviour. As a result, it enhances accuracy, lowers errors, and speeds up the pace of work, benefiting both employees and customers. Meanwhile, forward-thinking companies are figuring out how to use machine learning to spark new start-up prospects that will help them stand out in the marketplace. In a word, this Machine Learning prediction helps a start-up to grow or succeed and encourages to take risks based on failure percentage.

Literature Review

In recent times, startups have been entering increased attention in numerous regions of the world. Startups are mainly high-risk ventures and analyzing their eventuality is pivotal for making informed investment opinions. Through analysis, investors can estimate the viability, growth eventuality, and profitability of a startup, helping them decide whether to invest their finances or capital in the enterprise. Success for startups can be twofold: launching an IPO (Initial Public Offering) or getting merged/acquired by another company. The plenty of published workshops on the content of successful startups emphasized the need for fresh study.

Success rates of well- established businesses are the primary focus of the current body of exploration. Being models are harmful for calling the success of startups due to the large gaps between the success rates of established companies and new gambles. Startups have several crucial features that distinguish them from further established businesses. They are generally erected around a new and innovative business model that differentiates them from their challengers. This could involve offering a new product or service, using a new distribution channel, or borrowing a new pricing strategy. Also, they are frequently concentrated on a particular request niche or member, rather than trying to appeal to a broad followership. This allows them to target a specific client need and make a constant client base. Third, startups are generally characterized by their small size and lack of finance. This means that they're frequently light and suitable to adjust quickly to changing request conditions. In addition, startups frequently play a critical part in breaking established industries and creating new requests, which can lead to significant profitable benefits in the long term. Conducting a thorough analysis of a startup is critical to relating openings and challenges, and developing strategies to achieve success. Industry analysis, business model analysis, financial analysis, request entry strategy, and competitive analysis are five implicit startup analysis systems that can give precious insights into a startup's performance and help to identify openings for enhancement. By conducting these analyses, startups can place themselves for success and achieve their goals.

Aim of the work

The aim of the project is to predict whether the Startups will be a success or failure using supervised learning and using some hypothesis-testing techniques to validate the project.

Method and Materials

- We have used Pandas and NumPy library for basic operations with dataset.
- We have used Matplotlib and Seaborn libraries for plotting graphs for Exploratory Data Analysis.
- We installed Scipy and stats models to do Anova test, t-test and Chi-square test.
- From the sklearn library we used train_test_split which is used to split the data into training data and testing data, where training data is used to train the model and testing data is compared with the predicted output and it is checked whether the model is working correctly or not.
- From sklearn library we have used DecisionTreeClassifier and RandomForestClassifier for the model training.
- From the sklearn library we import the accuracy score, `classification_report` , confusion matrix, etc. to know how accurate the model is predicting and to validate the model

Dataset

The data had around 54k rows and 39 columns. The dataset had company information such as name of the company, url, market, country, state, region, city, founded date, first funding date, last funding date. It also had data on different investment types such as seed, venture equity crowdfunding, undisclosed funding, convertible note, debt financing, angel, grant, private equity, post ipo equity, post ipo debt, secondary market, product crowdfunding, round A-H series funding. Detailed descriptions of the different funding types is available here. Status of the companies were also available and segmented by acquired, operating and closed.

Study Design

Dataset

Startup data is a dataset of Kaggle on startup success prediction.
Startups Investments (Crunchbase)
https://www.kaggle.com/datasets/arindam235/startup-investments-crunchbase
There are 39 columns. There are 22 float and 16 objects.

Data cleaning

Data pre-processing is one of the key steps in the whole data mining process. That have been cleaned and merged into the final dataset. One of the biggest problems we had with the dataset was that it had a lot of zeros and a lot of columns to choose from. To clean the data, we removed extra spaces from different columns and also removed things like ",", "–" where ever necessary. We made sure that columns that had numbers were being read as numbers and also converted all of the date columns into date data types.

Rows with null values were removed. Columns with a high percentage of null values like state, city, region, and found date were removed.

Exploratory Data Analysis:

Here we used Heatmaps, Bar plot, Count plot with the help of Matplotlib and Seaborn libraries.

Algorithms used:

We have used 2 different types of algorithms in this dataset to find different results generated from this dataset. They are given bellow with a short description

Decision Tree

Decision Tree may be a Directed learning procedure that can be utilized for both classification and Relapse issues, but for the most part it is preferred for understanding Classification issues. It may be a tree-structured classifier, where inner hubs speak to the highlights of a dataset, branches speak to the choice rules and each leaf hub speaks to the result. In call Trees, for predicting a category label for a record we tend to begin from the basis of the tree. we tend to compare

the values of the basis attribute with the record's attribute. On the idea of comparison, we tend to follow the branch reminiscent of that price and jump to following node.

Types of Decision Trees:

1. Categorical Variable Decision Tree
2. Continuous Variable Decision Tree

Random Forest

Random Forest is a well-known machine learning calculation that has a place to directed learning Strategy. It can be utilized for Classification and Regression issues as well in Machine Learning. It is based on the concept of gathering learning, which could be a handle of combining numerous classifiers to unravel a complex issue and to progress the execution of the demonstrate.

Properties of random forest are given bellow:

1. It takes less preparing time as compared to other calculations.

2. It predicts yield with tall precision, indeed for the expansive dataset it runs effectively.

3. It can to keep up exactness when an expansive extent of information is lost.

Exploratory Data Analysis (EDA)

- First, we did a count plot to find out the frequency of the target variable *status*. We also realized that the *status* column had around 80% of the companies as operating status and the rest as closed and acquired companies.
- Then we segmented the dataset using the column *founded_year* and plotted a count plot for that. From that plot we understood that most of the startups were founded after the year 1990.
- Again, we used count plot to find the trend of emerging startups after the year 1990 to get a detailed plot. We found that the trend of emerging startups had a gradual increase up to the year 2012, after which there is a drastic fall.
- The data contained 115 countries. The dataset was joined with a different dataset from github that contained the country name and continent. We made a new column for the continent name. We used a barplot to show which continent has the most startups. America has the highest no of startups.
- The data contained 753 different market values. As there were a lot of different types of markets, we wanted to reduce this number. To reduce the number of markets, we grouped markets into different industry groups segment on the industry grouping list produced by crunchbase. The new column Industry Group had 43 industry groups. We made a barplot to show the no of startups under different industrial sectors. Software industry has got the highest no of startups. We also made a barplot of the closed startups using the Industry group column. The failure trend seems to follow the frequency of startups as the most failures are in the industries with the most start up. This makes sense as there might be immense competition in these categories.
- After changing all the categorical variables into numeric by using different Encoding techniques, we made a heatmap to visualize the correlation matrix between the variables. Using the heatmap we found out that columns *equity_crowdfunding, undisclosed, convertible_note, grant , post_ipo_equity, post_ipo_debt, secondary_market, product_crowdfunding, round_G, round_H* can be dropped, as there is low correlation.

Results

After doing EDA, we did some hypothesis testing like one sample t-test, two-sample t-test, chi-squared test, ANOVA Test, etc.

i. One sample t-test to determine if the mean total investment is significantly different from a specific value. Here, we took 5000000. We accepted the null hypothesis that mean total investment is not significantly different from 5000000.
ii. Two sample t-test to determine if the mean total investment is significantly different between two groups (successful startups vs. unsuccessful startups). We concluded that the mean total investment is not significantly different between successful and unsuccessful startups.
iii. Chi-square test of independence to determine if there is a relationship between startup success and industry group. We rejected the null hypothesis and concluded that there is a significant relationship between startup success and industry group.
iv. ANOVA: to compare the means of a numerical variable between more than two groups. When we checked for total investment and industry group, we got a p-value of 0.019362, which is less than the standard significance level of 0.05. Therefore, we can reject the null hypothesis and conclude that there is a significant difference in the means of the total investment variable between at least two Industry Group. We also checked between total investment and continent name, there also we rejected the null hypothesis because the p-value was less than 0.05.

ML Model

We used different types of models on the data to understand which model would be the best. We were trying to predict for closed, acquired and operating companies, that means we were doing multi-class classification. We thought the best way to classify them was by using Decision Tree Classifier.

When we decided to use only default values on the decision tree model, the model was overfitting. We used cost_complex_pruning_path from sklearn to find the most efficient alpha value that would help us prune the tree. The alpha value that would give the highest accuracy rate was used in the model. To tune the model, we also tested with grid search. Tuning with

cost_complex_pruning_path and grid search gave the same accuracy results but their decision tree looked very different from each other.

The next model we used was Random Forest. Random forest consists of a large number of individual decision trees that operate on ensemble. Ensemble method means that multiple models are generated and combined to solve the problem. For Random Forest, each individual tree in a random forest spits out a class prediction and the class with the most votes is the model's prediction. Here also we used random grid search for hyper parameter tuning. Using these parameters, the accuracy rate came up to be 0.86.

Codes

The link to the github repository is given below:

https://github.com/joemon-paulson/Startups-Prediction-using-supervised-learning-Methods

Conclusion

As the dataset contained 80% of the data that had operating companies, the multiclass model was good at predicting for operating companies. According to the Decision Tree model venture, Industry group, and continent name were the most important features. Random Forest model showed that Industry Group is the most important feature. We received the best result when we used SVM and Random Forest for Multi Class Classification. At first both the models showed signs of over fitting as the accuracy of the training set was high. After using grid search to tune the hyper parameters, random forest showed a slight improvement. This dataset is a real-life dataset and we have come to an assumption that if anyone want best result from it, he should take the Random Forest Classifier in consideration as random forest has the highest accuracy rate for these types of datasets. If anyone wish to start a startup with his unique idea, following the way generated from the decision tree will be the best for him. We hope this finding of ours will help others and add some value in machine learning industry.

Future works

Startups are the backbone of the growing economy and society. A startup can create a lot of new opportunities for unemployed people. Startup usually brings new services, products which help in growth of the economy. Startups attract a lot of venture capitalists, investors, and other new resources. To have a successful startup there are many things that are taken into consideration. One must have a deep knowledge of the market that they are going to handle. Market analysis helps a lot. It helps in identifying potential customers and the needs of the market.

Predicting the success or failure of startups is an interesting and complex problem. Supervised learning is a good approach to solve this problem. There is high potential for this project. Improving the accuracy of the model is a significant part of it. For future scope, we would like more data for closed and acquired companies, and test model with one-hot encoding. In this project we are using random forest and decision tree but if we were able to use Gradient Boost, Logistic Regression, MLP Neural networks then we would get a better result.

Using Crunchbase API, we can also make a real time dashboard and deploy a model so that it can assist investors and founders. In the current models, the state of global economics and crises have not been considered. Moderate funding during a recession or an extraordinary global event such as the pandemic could point to a more successful startup than higher funding in normal times. Geographic location and the implication of a nation's economic scenario have not been considered.

Facilitating more data and introducing new algorithms will be very useful. A larger volume of data is always better for analyzing. Supervised learning model usually works on quantitative data but what if we can bring in the qualitative data such as the customer review then it will be more useful in predicting the success of the startup. When we predict the outcome as success or failure using the model, if the model is able to find the cause of the failure or success then it can be very helpful in improving the overall performance of the model.

Different evaluation metrics can be used to assess the performance of predictive models, and the choice of evaluation metric can have an impact on the accuracy of the predictions. Optimizing a model for precision may achieve high precision but low recall, while optimizing for recall may achieve high recall but low precision. If we are able use better metrics, then this type of problems can be reduced to a certain level. Whenever a prediction model for startup is created, we

use data such as financial statements and customer review. If we are able to use other alternative data source, then it could be more efficient. In general, we can say that if we improve the prediction accuracy, consider other external factors, use different evaluation matrices, get real-time data, different data sources then can help in improving the project in reaching new heights.

Key Learnings

1. Feature Selection: Feature selection plays an important role in the success of a predictive model. Identifying the most relevant features that have the highest impact on the prediction is crucial. In the context of startups, features such as funding, team size, location, market size, and previous experience can be important factors to consider.
2. Model Selection: Decision trees and random forests are powerful algorithms for classification tasks such as predicting the success of startups. However, selecting the best model for a particular dataset and problem is not always straightforward. It is important to compare the performance of different algorithms and select the one that provides the best results.
3. Data Preprocessing: Preprocessing the data before feeding it into the model is essential. This includes handling missing data, handling categorical data, and normalization/scaling. Preprocessing ensures that the model is trained on a clean and consistent dataset, which improves its accuracy.
4. Interpretability: Decision trees and random forests are interpretable models, which means that the decisions they make can be easily explained. This can be useful in the context of startups, where investors and stakeholders may want to understand the factors that contribute to the success or failure of a company.

References

Predicting Startup Success

Harjo Baskoro1 , Harjanto Prabowo2 , Meyliana3 , Ford Lumban Gaol4

https://www.kaggle.com/datasets/arindam235/startup-investments-crunchbase

https://www.analyticsvidhya.com/blog/2021/11/startups-profit-prediction-using-multiple-linear-regression/

Predicting Success Rate of Startups using Machine Learning Algorithms

Yashvi Mehta