**Start-up Success Prediction**

**AIDI 1002: AI Algorithms**

**Statement of Work(V1)**

**FACULTY:**

**Marcos Bittencourt**

**STUDENT:**

**Shrita Gaonkar 100799307**

**EXECUTIVE SUMMARY**

Startups are businesses that are in the early period of their operations. North America, like SpaceX, Instagram, and Airbnb, is home to many of the most popular companies in the world that started this way. These companies have grown out of their start-up status, but start-up activity is still a breeding ground in the region. North American companies are among the world's highest-valued unicorns.

According to a research, in 2017 the Toronto-Waterloo startup ecosystem had an estimated 2,100 to 2,700 startups operating. These companies also help in increasing the economic growth of a country.

**RATIONALE STATEMENT**

Investors often suffer from huge losses for investing in the wrong company. In real life, it is daunting for investors to predict success for a start-up. My model helps to predict the success rate of the start-up companies based on various parameters like location, funding, type of business, etc. Using this model, investors will get a fair idea on which companies they should target for investment. A company will be considered a success based on the company’s mergers and acquisitions. A company will be considered a failure if it was forced to shut down.

Based on the rejections, a company will be motivated to come up with new products and new solutions.

**DATA REQUIREMENTS**

* The first data requirement will be to find out the key variables which define the quality of the data.
* For developing various types of models, the data should be large and diverse.
* The data should be in .csv or in excel format for better extraction.
* The data size must be feasible according to the available machine.
* The dataset is required to not have special characters in them so the data can be cleaned easily.
* The dataset should have unbiased entries.
* While collecting data, all the aspects must be considered.
* The data should be in alignment to solve the problem.

**ASSUMPTIONS**

* I am keeping in mind that there was no data manipulation before.
* I am assuming that we have enough data features available to build the machine learning models.
* I have enough data to conduct analysis.
* I am assuming that all columns are necessary, and all independent variables are correlated with the dependent variable and not changing any variable name.
* The dataset would be easily understandable to all the developers.

**CONSTRAINTS**

* One of the biggest constrain is that all the independent variables are not properly defined.
* The data is imbalanced because all the dependent variables have different total count.
* The variables that are most useful are yet to be defined.
* Few variables have outliers and we do not know that they are valuable for the model or not.
* All independent variables might not have correlated with the dependent variable.
* The dataset requires to have at least one dependent variable for conduct analysis.

**DATA SOURCES**

* The data was originally provided by Ramkishan Panthena which was further made available by Kaggle

I have downloaded the dataset from Kaggle

**DATA**

Data was acquired from Kaggle.

There are 48 columns/features. Some of the features are:

* age*first*funding\_year – quantitative
* age*last*funding\_year – quantitative
* relationships – quantitative
* funding\_rounds – quantitative
* funding*total*usd – quantitative
* milestones – quantitative
* age*first*milestone\_year – quantitative
* age*last*milestone\_year – quantitative
* state – categorical
* industry\_type – categorical
* has\_VC – categorical
* has\_angel – categorical
* has\_roundA – categorical
* has\_roundB – categorical
* has\_roundC – categorical
* has\_roundD – categorical
* avg\_participants – quantitative
* is\_top500 – categorical
* status(acquired/closed) – categorical (the target variable, if a startup is ‘acquired’ by some other organization, means the startup succeed)

**DATA LIMITATIONS**

* Short forms have been used to name the features. Data columns do not explain the feature quite well.
* There are columns which follow the method of label encoding. But it will not be ethical to not change the columns first and then apply label coding on my own.
* The dependent variable is categorical. A percentage is considered more valuable source in analytics
* Some column names are unnamed.
* Instead of using date, months can be extracted from the dates which can be further used as an important feature.

**TEST PROCESS**

|  |  |  |
| --- | --- | --- |
| **TASK** | **HOURS** | **DELIVERY DATE** |
| Collect information | 5 | 21-OCT-2020 |
| Analyze dataset | 20 | 26-OCT-2020 |
| Clean dataset | 10 | 15-NOV-2020 |
| Develop data processing pipeline | 15 | 15-NOV-2020 |
| Train all models | 30 | 10-NOV-2020 |
| Building all models | 25 | 17-NOV-2020 |
| Test models | 20 | 20-NOV-2020 |
| Evaluate models | 15 | 21-NOV-2020 |
| Feature selection | 15 | 22-NOV-2020 |
| Prototyping | 10-32 | 3-DEC-2020 |
| Refine model | 10 | 6-DEC-2020 |
| Develop report | 8-19 | 7-DEC-2020 |
| Develop score card | 10 | 10-DEC-2020 |
| Deployment | 20-25 | 12-DEC-2020 |

**Project V2(December 1, 2020)**

Removed null values

* After creating a new column there was a need to understand how it affected other columns
* It also led to a new problem of handling missing values in the new column
* Which were removed simply by indexing
* The other null values in age\_first\_milestone\_year, age\_last\_milestone\_year, age\_first\_funding\_year, age\_last\_funding\_year were filled my mean of that column.

Removed columns with same information

* There were some columns which had the same meaning Unnamed 6 which is just a combination of state\_code and zip\_code
* Removed state\_code because state/location of the company was already present in the other column

**Models Used**

* KNN
* Random Forest Classifier
* Decision Tree Classifier
* SVC