**TECHNOCOLABS DATA SCIENCE INTERNSHIP**

**Project Report**

**Prediction Startup‘s Acquisition Status**

**Under the Guidance of**

**Bidhya Pokharel**

By Mega Satish

**ABSTRACT**

Start-up is a company or project undertaken by an [entrepreneur](https://en.wikipedia.org/wiki/Entrepreneurship) to seek, develop, and validate a scalable [business model](https://en.wikipedia.org/wiki/Business_model). While entrepreneurship refers to all new businesses, including self-employment and businesses that never intend to become registered, startups refer to new businesses that intend to grow large beyond the solo founder. At the beginning, startups face high uncertaintyand have high rates of failure, but a minority of them do go on to be successful and influential.

Many nations have experienced an exponential increase in start-up agreements during the last decade. As a result, it seems to be a considerable difficulty comprehending what makes these high-risk excursions productive and, as such, attractive to financial experts and company visionaries. A start-success up's is defined here as the event that provides a large sum of money to the original founders, investors, and initial employees.

This project aims to determine the status of a start-up based on its financial records. We have the Crunchbase Database on which exploratory data analysis is performed. We have deployed the app on the Heroku Cloud Application Platform. Here, we intend to base an evaluation on every basic criterion that is taken into account when predicting the status. Random Forest Model is a classification algorithm consisting of many decision trees. It is used by a variety of software and computers to determine the best feasible action or path in a given situation.

By providing inputs such as year of foundation, relationships, funding rounds, total amount of funding, number of founding rounds, the predictor may be used to determine the status of a particular status. The goal of this project is to learn and get hands-on experience in Data Analytics and Machine Learning.

**INTRODUCTION**

Startup acquisition is the process by which a large corporation purchases a small company/startup and gains control of it by acquiring most or all of the company's shares or assets.

A firm may wish to acquire/purchase a startup for a variety of reasons. If an entrepreneur is ready to sell the firm and pursue a new venture, the company will want a plan to guide them through the full purchase process. There is a certain procedure that will highlight critical features of acquisition and how an entrepreneur may reduce the likelihood of failure.

For most firms or startups, being purchased by another firm not only validates that the company is on a development path in the particular market, but it also bridges the financial gap that the company has been attempting to fill for quite some time. Acquisitions and mergers are both thrilling and hard for entrepreneurs running dynamic businesses.

Mergers and Acquisitions are quite active in the UK business (M&A). Almost 1,400 mergers and acquisitions were completed in the first half of 2019 across all major industries, including telecommunications, insurance, manufacturing, information technology services, and wholesale trade.

**AIM & OBJECTIVES**

The primary goal of this research is to create a prediction model to determine whether a business is viable: operating, closed, acquired or IPO The most significant stage in entering startup acquisitions is to understand its financial information, such as total financing amounts, funding dates, number of funding rounds, and headquarters locations.

**RANDOM FOREST CLASSIFIER**

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is:

*A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.*

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don’t constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction. So the prerequisites for random forest to perform well are:

1. There needs to be some actual signal in our features so that models built using those features do better than random guessing.
2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

**\**

**METHODOLOGY**

The steps to create and deploy a model are:

1. Dataset collection, exploration and analysis.
2. Build Model
3. Train Data.
4. Test Data.
5. Evaluate Model.
6. Deploy Model.

**Data Collection and Selection (CrunchBase Dataset from Kaggle)**

The dataset used for this project is a kaggle dataset sourced from Crunchbase called: “Crunchbase 2013- Companies, Investors, etc.” There are nearly 196553 rows and 15 columns, each row of the dataset contains a startup’s information. The dataset labels show that the dataset is extremely biased.

* Data Cleaning:

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.

The dataset had many features and attributes, all were not relevant for the prediction of the startup’s status such as “Company Name”, “unnamed: 0” and many other columns. So we dropped all the unnecessary data columns.

Afterwards, we removed the inconsistencies present in the datatype of some columns to have a uniform dataset.

Then, handling missing values and removing the duplicates values was primordial in order to continue with the data exploration and analysis. Removing the outliers with the help of IQR method was important as outliers are excessively deviating the value from the scale of the feature.

* Data transformation

It is the process of converting data from one format or structure into another.

Here we did One Hot Encoding on Category and Country\_code columns as there were so many countries and the data can be provided to machine learning algorithms to improve predictions. This created new columns for each Country\_code.

Two more columns were added: “isClosed”: tells us whether the startup is still running or closed and “Active days” which shows the number since the startup has been running.

**MODEL BUILDING**

### Random Forest Model

The random forest is a model made up of many decision trees. Rather than just simply averaging the prediction of trees (which we could call a “forest”), this model uses two key concepts that gives it the name random:

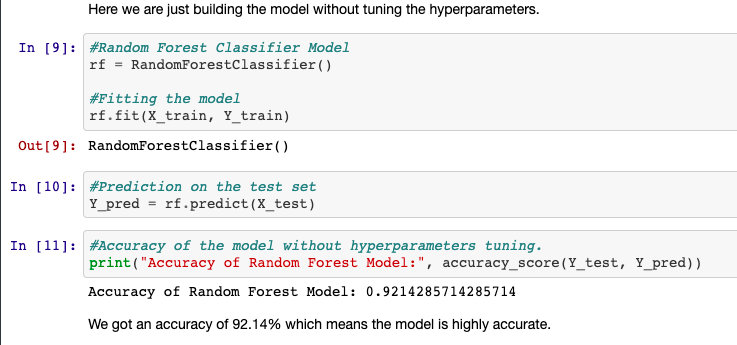
1. Random sampling of training data points when building trees
2. Random subsets of features considered when splitting nodes

We have built one model without hyperparameters tuning and one with hyperparameters tuning.

The final data consists of 3496 rows and 7 columns.

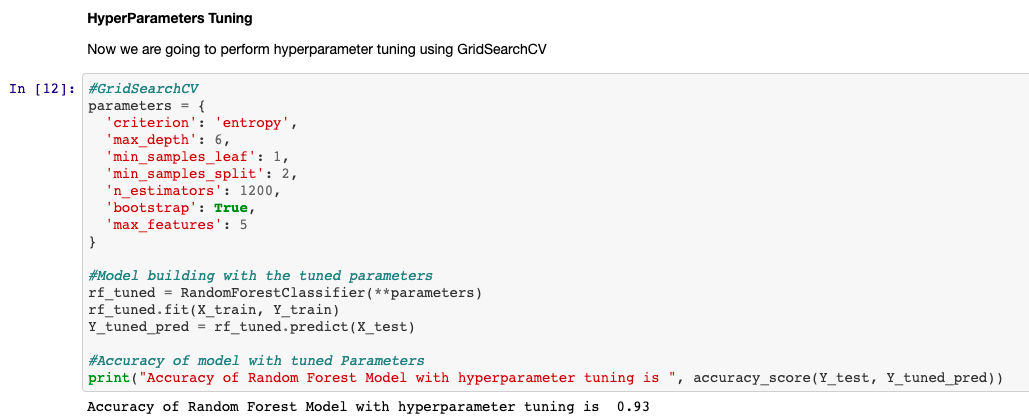


**Without Hyperparameters Tuning**



We got an accuracy of 92.14% which is pretty good.

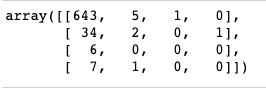
**With Hyperparameters Tuning**

****

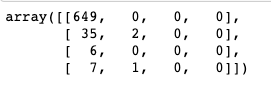
For the Random Forest Model with hyperparameters tuning, we got an accuracy of 93%.

**MODEL EVALUATION**

Confusion Matrix for Normal Random Forest Model

****

Confusion Matrix for Random Forest Model with Hyperparameters Tuning

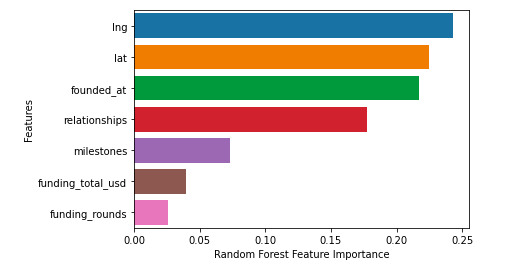


**Feature Importance**

Machine learning (ML) algorithms such as neural networks and Random Forests (RF), as well as any other machine learning algorithm, are frequently regarded as producing black box models because they do not provide a direct explanation for their predictions.

However, because they can model complex relationships in data, these methods frequently outperform simple linear models or decision trees in predictive performance.

Feature importance is used to clarify how features affect model performance and to determine which features are the most effective for our model.



**MODEL DEPLOYMENT**

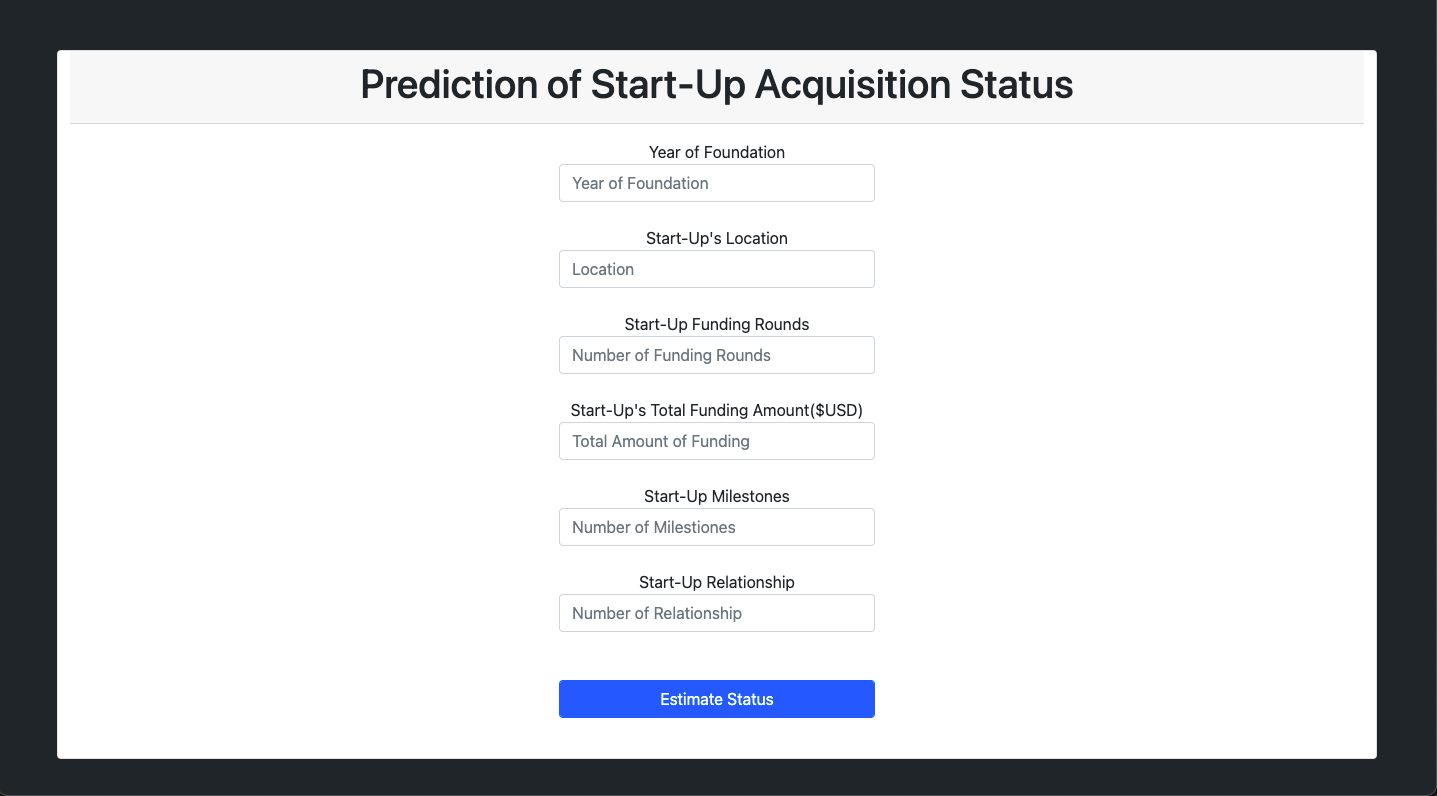
The project is hosted on Heroku which is a cloud Platform as a container-based Service (PaaS). Heroku is used by developers to launch, manage, and grow contemporary programs. Heroku is an open-source software platform for machine learning and data science that makes it simple to develop and publish attractive, bespoke web apps. The benefit of web apps is that they are platform agnostic and may be operated by anybody with an Internet connection. Their code is run on a back-end server, which processes incoming requests and answers using a common protocol that all browsers can understand.

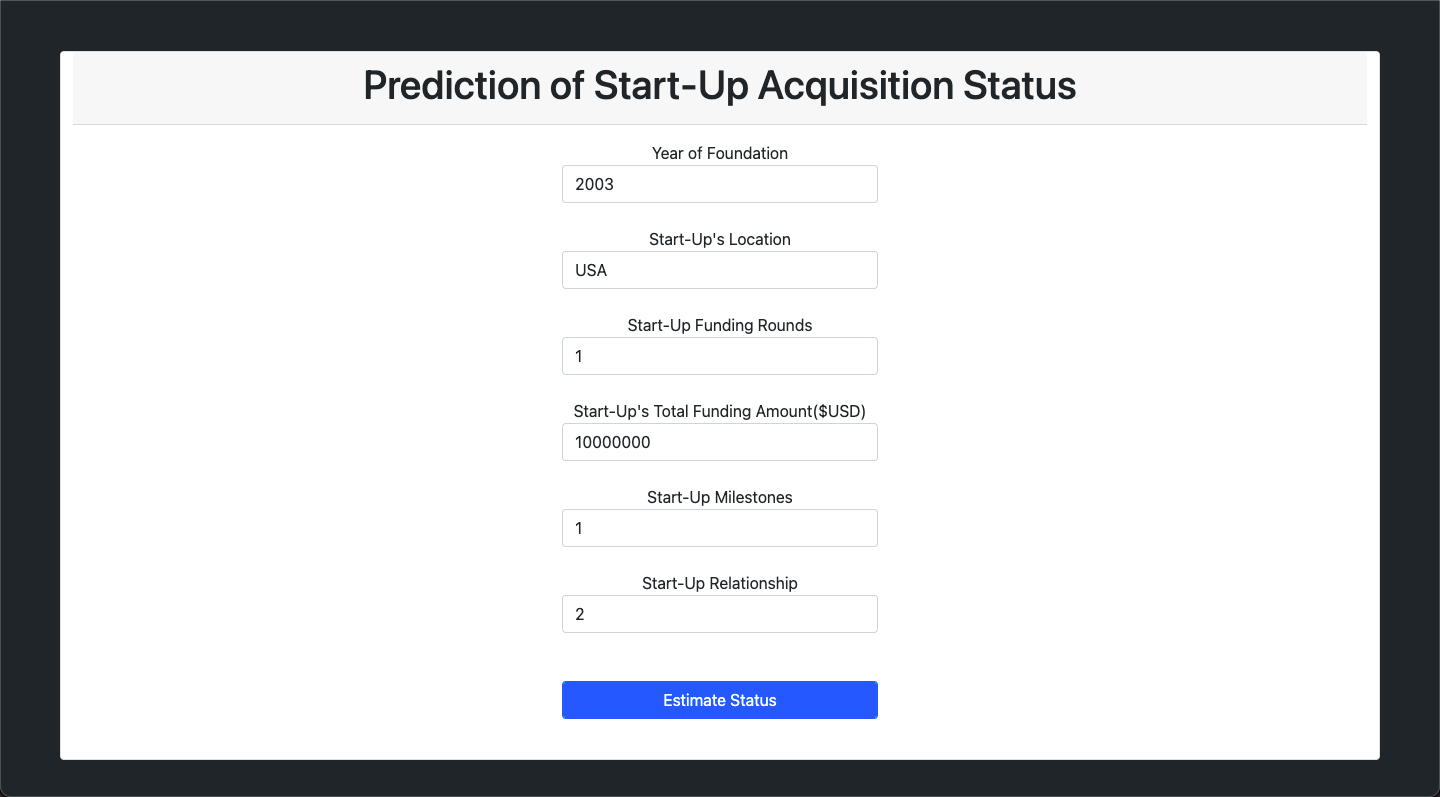
Necessary Files:

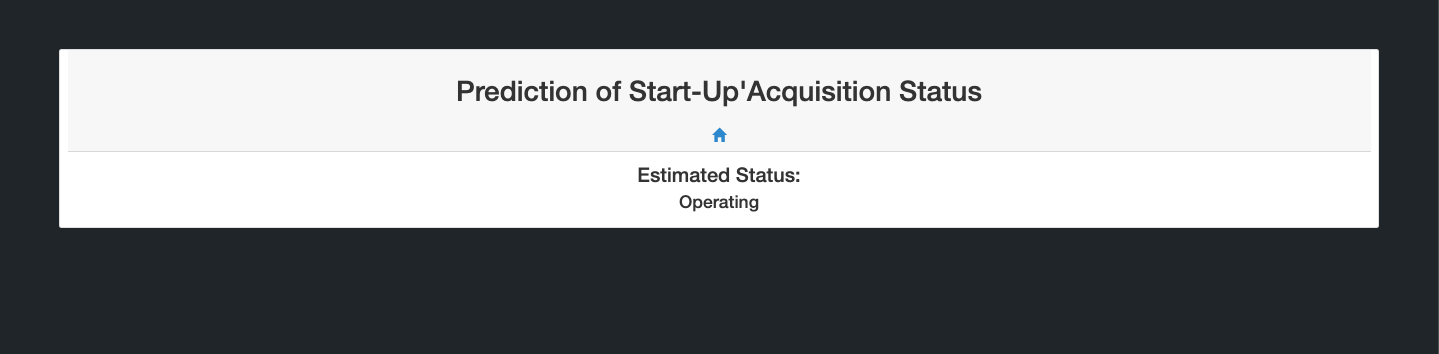
1. [app.py](https://stock-trading-with-rl.herokuapp.com)
2. requirement.txt - Contains a list of all the dependencies that your code requires in order to function properly.
3. Procfile - In an app, a Procfile is a list of process types.

**RESULTS**

**Web App:** [**https://startup-acquisition-status.herokuapp.com/**](https://startup-acquisition-status.herokuapp.com/)

****

****

****

**CONCLUSION**

The primary goal of this study was to develop a methodology for categorising organisations or start-ups as operating, IPO, closed, or acquired. The goal was met by developing multilabel classifiers to classify startups. The model can efficiently categorise not only the total number of successful firms in the dataset (TPR, recall), but also all of the successfully-classified enterprises that are successful (Precision).