Predicting a Startup's Acquisition Status

- **Deepak Nimbalkar**

Contents

Abstract

This project predicts a startup's acquisition status based on its financial statistics. In order to overcome the main challenge of biased data without under/oversampling the data, a novel ensemble model used. The resulting model combines a high precision model with a high accuracy model trained on a dataset transformed by the first model. Preliminary experiments suggest that this new model has the potential to yield higher precision predictions while preserving performance with respect to accuracy and weighted recall.

Introduction

The goal of this project is to predict a former start-up's acquisition status based on a company's financial statistics. The results of this project may be of particular interest to investors as well as job applicants to pre-IPO companies as it can be extended to look at the likelihood of the prospective company being acquired, closed or reaching and IPO. The resulting algorithm takes in a start-up's financial statistics such as total funding dollars, funding dates, number of funding rounds, and headquarter location as input's. The algorithm then predicts whether the start-up has been closed, acquired, is operating, or has reached and IPO.

Requirement and Tools

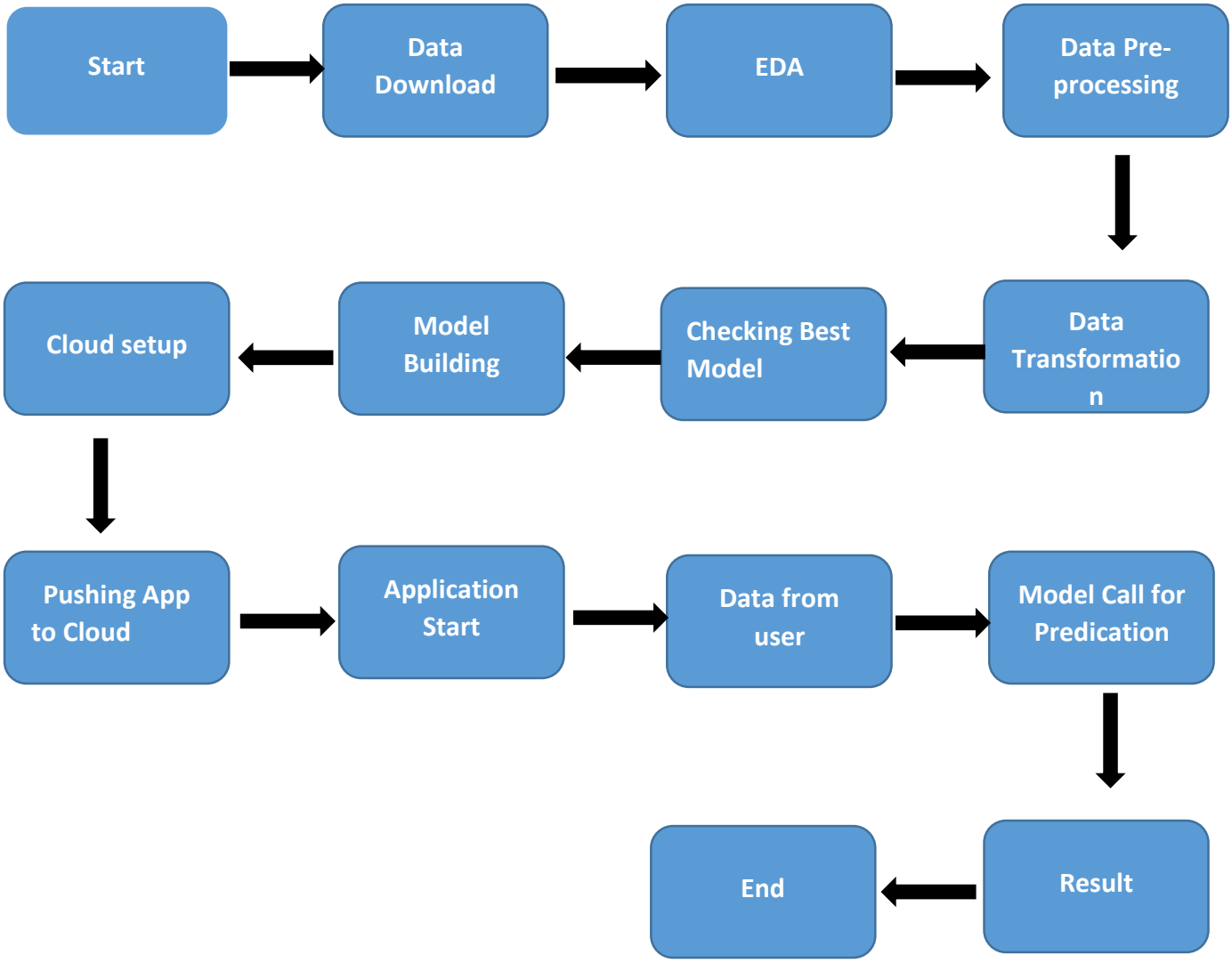
The dataset used for this project is a Kaggle dataset from Crunchbase called ‘Crunchbase 2013’ – Companies Investors. Tools –

1. Python
2. Numpy
3. Pandas
4. Sklearn
5. Pandas Profiling
6. Flask (HTML, css)
7. VS code
8. Jupyter Notebook
9. Heroku Cloud

Objective

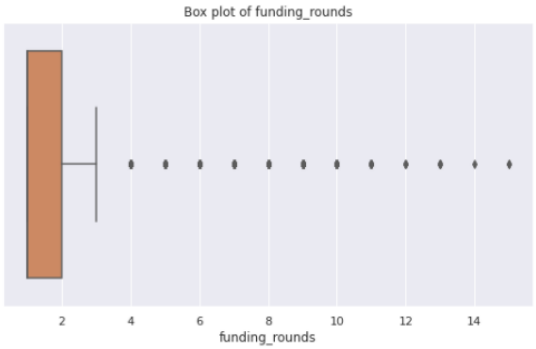
The goal of this project is to predict a former startup's acquisition status based on a company's financial statistics. The results of this project may be of particular interest to investors as well as job applicants to pre-IPO companies as it can be extended to look at the likelihood of the prospective company being acquired, closed or reaching and IPO. The results of this project may also give insight to which features have the most influence on the predictions. The algorithm then predicts whether the startup has been closed, acquired, is operating, or has reached and IPO. The main challenge for this problem is dealing with an imbalanced dataset where one class is overrepresented, but under/oversampling cannot be used as a technique to balance the data. In order to address this, an ensemblebased technique that combines the results of a high precision anomaly detection algorithm (QDA) with a random forest classifier.

Architecture



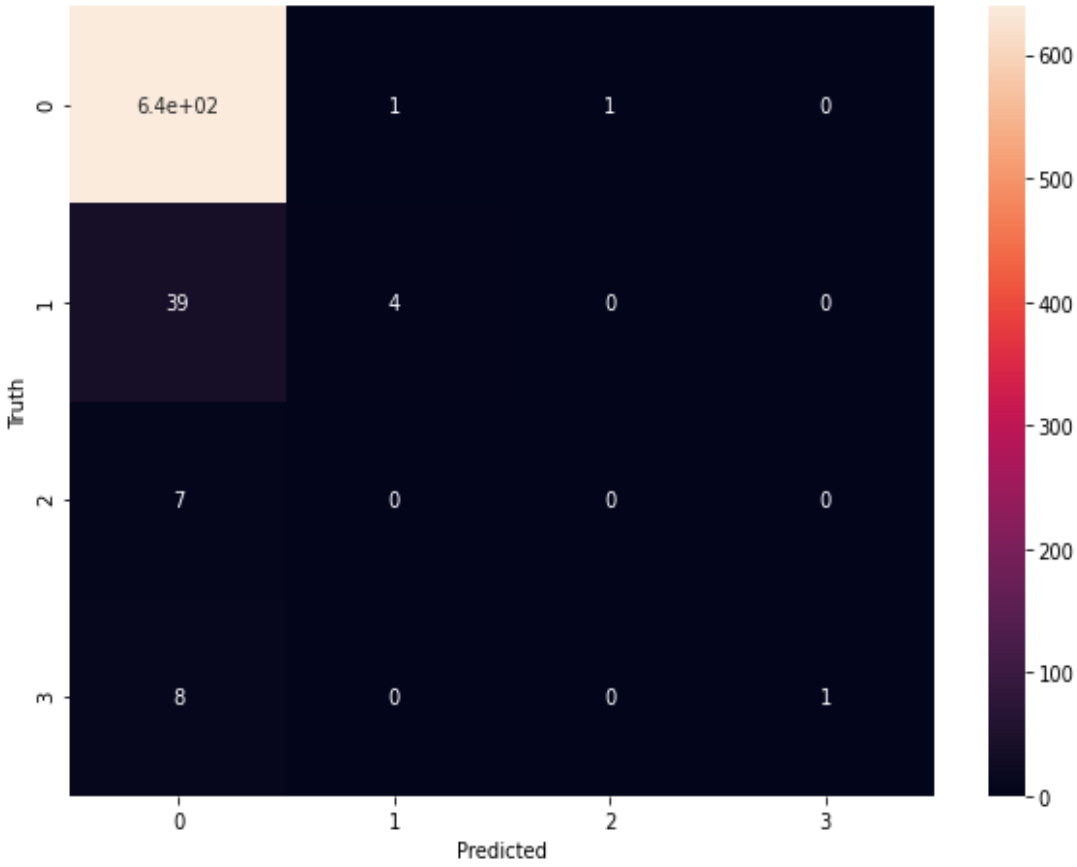
EDA

1. Data Cleaning –
   Delete region, city, state_code, as they provide too much of granularity. And deleting some other irrelevant features.
   Deleting duplicate values, deleting those row which has more than 98 % of null values.
2. Removing Noise
   Deleting instances with missing values for status, country code, category code and founded at.
   Deleting outlier for ‘founding total usd, and founding rounds.



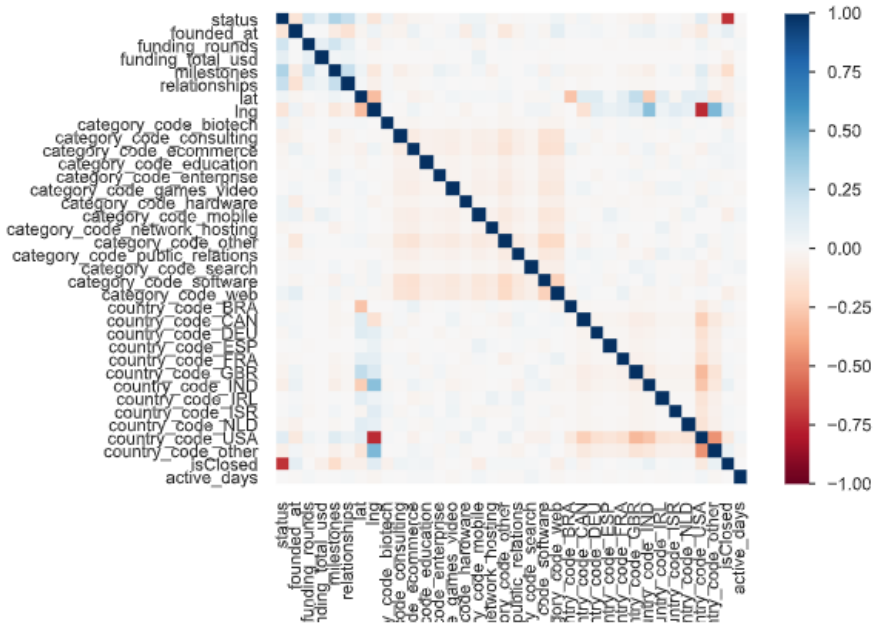
3. Data Transformation
   Changing original data only put the year except all date.
   Generalize the categorical data i.e category_code, status, category_code.
4. Creating new features
   isClosed and active_days
5. Applying one-hot encoding to categorical_code column.

Data Visualization



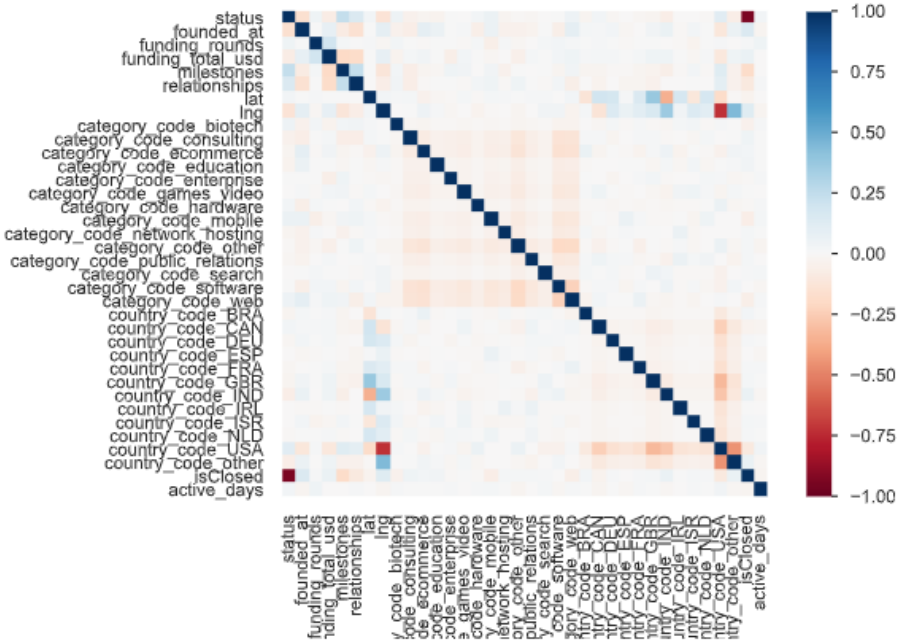
Pearson's r | Spearman's ρ | Kendall's τ | Phik (φk) | Cramér's V (φc)

Toggle correlation descriptions



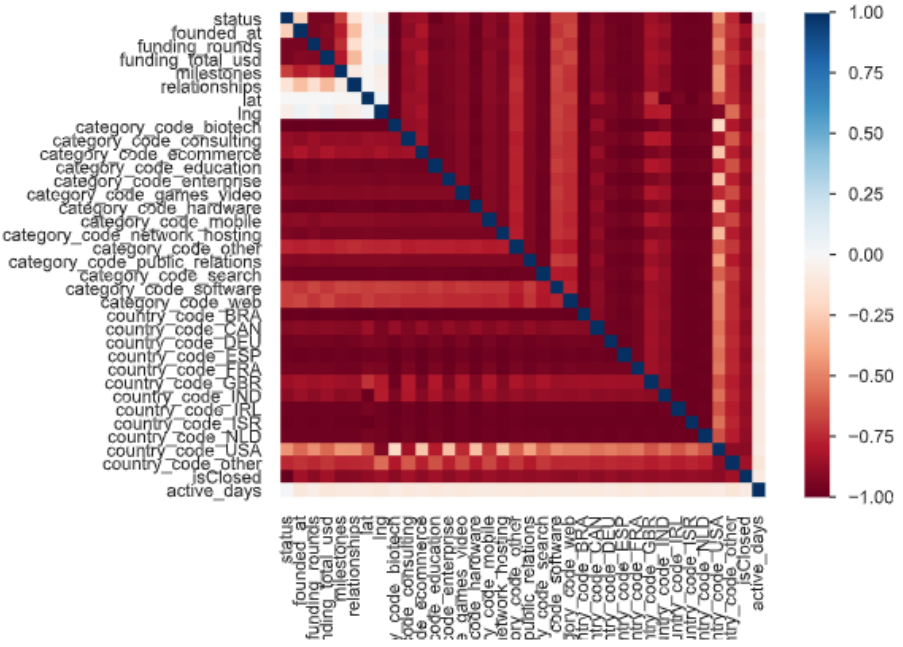
Pearson's r | Spearman's ρ | Kendall's τ | Phik (φk) | Cramér's V (φc)

Toggle correlation descriptions



Pearson's r | Spearman's ρ | Kendall's τ | Phik (φk) | Cramér's V (φc)

Toggle correlation descriptions



status

Categorical

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	4
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	27.4 KiB

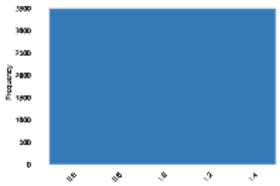
Toggle details

Overview | Categories | Words | Characters

Common Values

	Count	Frequency (%)
	3245	92.8%
	191	5.5%
	36	1.0%
	25	0.7%

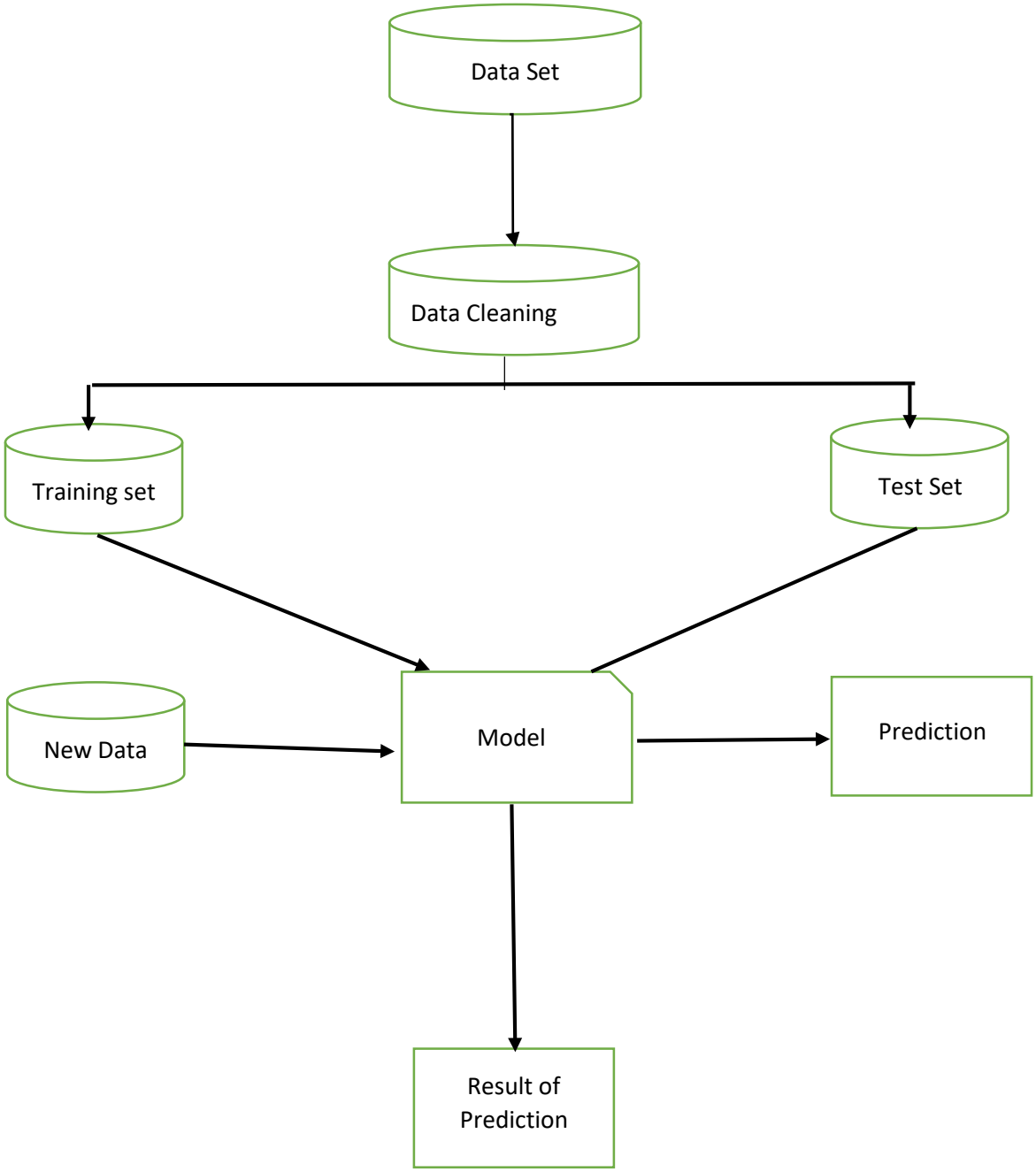
Length



Histogram of lengths of the category

Pie chart

Model Training and Evaluation



Deployement Process

1. Uploading all code in Github
2. Create account on Heroku cloud
3. Create a App on a Heroku cloud
4. Connecting github repo to App
5. Uploading all code in Heroku cloud
6. Deploying the all code on Heroku cloud(like install all required packages)
7. And checking the result

Conclusion

In the start-up's acquisition status prediction we will predict the status of start-up’s based on the Kaggle dataset from Crunchbase called ‘Crunchbase 2013’ – Companies Investors data used to train our algorithm, so we can identify the start-up acquisition status.