Startup Success Prediction

Abstract

The aim of this project is to develop a classification model to predict the success of the startup. The dataset used is "Start Up Investments" from Crunchbase which is a platform for finding business information about private and public companies.

Design

This project is mainly intended to help investors and startups owners. The dataset includes the status of the startups (Operating, Acquired, Closed). By classifying and understanding the important features that determines the success of a startup. The survivorship of a startup could be determined which helps the investors to make the right investments choices and startups owners to determine the flaws in their businesses.

Data

The dataset used is "Start Up Investments" from Crunchbase which is a platform for finding business information about private and public companies. The dataset is acquired through a secondary source which is Kaggle. The data set contains 49438 rows, each row of the dataset represents a startup company with 39 features such as: name, category list, market, funding total in USD, country, funding rounds number, round A-H series funding and the target feature for this analysis: status. Upon initial view, the dataset seems imbalanced as the data contains %6 acquired startups, %5 closed and above 80% operating startups.

Algorithms

Feature Engineering

1. Extracting year from dates variables
2. Converting categorical features to Encoded variables
3. Extract the difference between the founding year and first year of funding

Models

The main classifier used is Random Forest. It was chosen as its performance was the highest compared to models such as Logistic Regression and Naïve Bayes. Moreover, Random Forest is optimal for multiclass classification. As for handling imbalanced classes two models were used: Random Over Sampler and Synthetic Minority Over-sampling Technique.

Model Evaluation and Selection

The original data contains 49438 rows. After balancing the data contains 41820 rows which was split into 80/20 training and testing set. Stratified 10-fold cross validation was used on the training data set.

The metrics used to evaluate the performance are accuracy, F1, Precision and Recall

Random Forest with SMOTE evaluation

Training Accuracy: 91%

Testing Accuracy 92%

Testing F1:92%

Testing Recall: 92%

Testing Precision: 92%

Random Forest with ROS evaluation

Training Accuracy: 97%

Testing Accuracy: 97%

Testing F1:98%

Testing Recall: 98%

Testing Precision: 98%

Tools

The tools used for this project include:
1-Python3: as the main programming language used
2-Jupyter: as the IDE for python
3-Numpy: to manipulate the dataset
4-Pandas: to clean and preprocess the dataset
5-Seaborn: to visualize the dataset
6- sklearn: for feature selection and modeling
7- imblearn: to deal with the imbalanced dataset

Communication

Presentation can be found in github repository.
https://github.com/lamahr0/Startup-Success-Prediction/blob/main/Startup%20Success%20Prediction%20PPT.pptx