Project Proposal: Startup Success Prediction

Big Data & Machine Learning | Xinjie Lian & Zhihao Wang

Business Problem and Solution Overview (BLUF)

Startups, vital to global innovation and economies, heavily rely on early-stage funding. From the investors' end, to monitor startup performance and precisely predict startup future lucrativeness, it is important to analyze what features determine a Startup's success.

We will apply data processing techniques including KNN and winsorization to a dataset consisting of 48 features including location, founding time, and business category, of 923 Startup samples. Next, we will employ RFE to set up an effective feature set to make a startup be obtained at the end. Then, we will apply several classification models, such as logistic regression and Support Vector Machine, and tune the best one based on accuracy and CV_AUC Score, so that we can build a model that can accurately predict the success of a startup.

Dataset Description

The dataset we found from Kaggle contains 923 Startup samples in the US and 48 features. The target variable is “Label” in the file, which is a dummy variable suggesting whether this Startup is successful or not. If “Label” is 1, the company's founders receive a large sum of money through the process of M&A (Merger and Acquisition) or an IPO (Initial Public Offering). “Label” is 0 suggests it is considered as failed. As for the feature set, it consists of 2 main parts - basic company information and funding information. The prior part contains location and business scope, while the latter contains companies receiving funding dates and previous funding rounds. These are important factors mentioned in recent literature (Sevilla-Bernardo, Sanchez-Robles, and Herrador-Alcaide, 2022).

There are limitations in this dataset. Although a large number of important features are included, we cannot find much information about some variables in the dataset, such as “untitled 0” and “untitled 6” variables. We have to delete them to increase the model’s simplicity. Another limitation of this dataset is the lack of information on the current competitive environment and management teams of Startups.

Data Source: Startup Success Prediction (kaggle.com)

Workflow

1. Data preprocessing: Missing value & Duplicates check, Outlier drop, Balancing-SMOTE, Scaling, Label encoding, Descriptive analysis-sweetviz.analyze())
2. Predictors & Target separate-‘status’ column
3. Train & Test separate
4. Feature engineering-RFE & Importance graph (EDA)
5. Model fitting (RandomForest, KNN, LogisticRegression, DecisionTree, Gradient Boosting, AdaBoost, XGboost, lightGBM)
6. Comparison leaderboard-classifier metrics(Accuracy, cv_auc, recall, precision, F-1)
7. Tuning/Hyperparameters-Gridsearch
8. Model Evaluation-RAI Dashboard
9. Conclusion

Metrics and Evaluation

We will develop a leaderboard to aggregate all classifiers’ performance. The model with the most number of metrics ranking first will be picked. Metrics include Accuracy, CV_AUC Score, Precision, F-1. As our dataset will be balanced first, accuracy and CV-AUC score will be effective evaluation metrics. Moreover, in a startup success prediction, a false positive will be highly costly, as investors tend to be precautious in making investment decisions to avoid low ROI. As false negative events will depress the startup’s passion, F-1 should also be leveraged.

Reference:

Sevilla-Bernardo, J., Sanchez-Robles, B. and Herrador-Alcaide, T.C. (2022) ‘Success factors of startups in research literature within the entrepreneurial ecosystem’, *Administrative Sciences*, 12(3), p. 102. doi:10.3390/admsci12030102