A New Machine Learning Framework for the Prediction of Companies’ Business Success

Yiheng An, Xinyi Zhang, Beier Cao

Background

Due to the impact of COVID-19, the future of companies has become more elusive for many people. Given shocking value-destroying events nowadays, it has become increasingly crucial for investors to make wiser decisions. Fortunately, the innovation of algorithm and the release of computing power has provided new possibilities.

In this project, we introduced the exchange rate data and Google Trends data to build a robust predictive framework of companies’ business success. Many techniques in feature extraction, feature engineering, predictive modelling are involved.

Objectives

In this project, the first task is to set up a reasonable standard which defines the label of business success or not for each company.

Besides, we need to build competing supervised classifiers and evaluate their performance. Our evaluation would focus on the robustness of models in generalization and the value of the exchange rate features and web search features in this application.

Moreover, based on the performance of each algorithm, we hope to summarize possible user scenarios for powerful models and make our project more practical and useful.

Data

The datasets used in this project can be roughly divided into three categories, which are company-related data, exchange rate data and web search data.

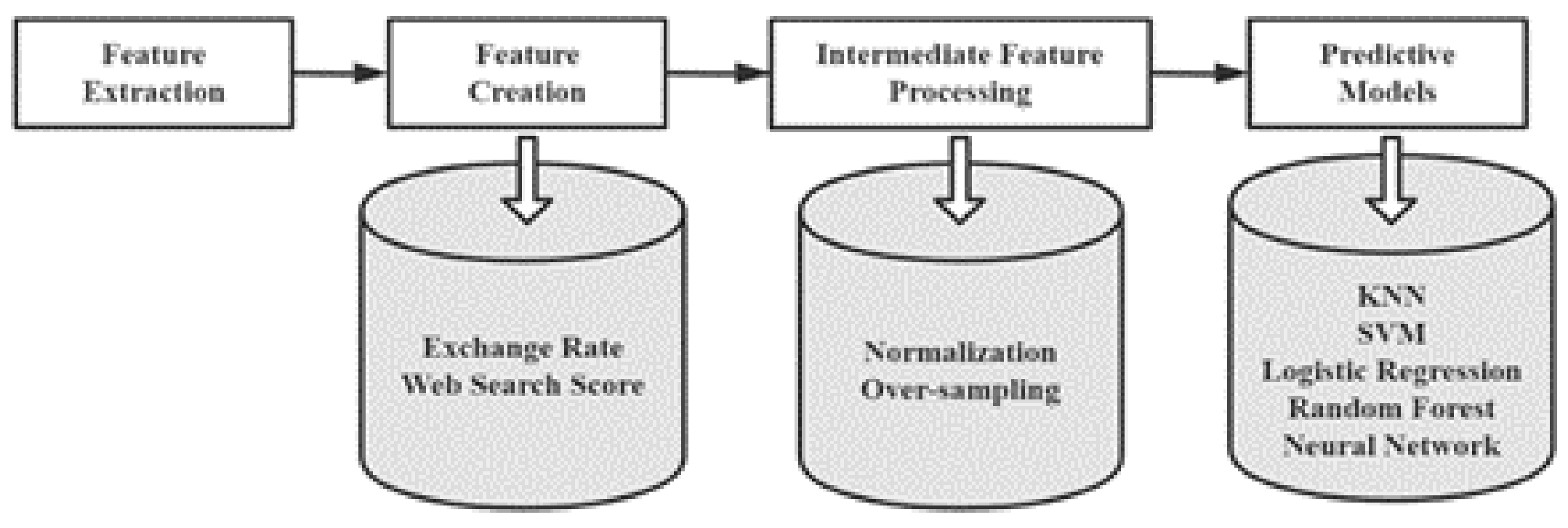
The **company-related data** is gained from Crunchbase Data Export, the exchange rate data is collected from OANDA database and the web search data is obtained from Google Trends. The raw datasets include records of more than 20,000 companies from 68 countries, January 2014 to February 2021.

As for the **exchange rate data**, we chose gold price to be the target (basic currency) and formatted the annually data for each currency based on the country codes for each company.

For the **web search data**, we formatted the monthly web search score of IPO and M&A for each country code. The time window of both exchange rate dataset and web search dataset matched the horizon of company-related datasets.

Framework Design

Our machine learning framework can be divided into four main components. They are feature extraction, feature creation, intermediate feature processing and predictive modelling.


Feature Extraction

Feature Creation

Intermediate Feature Processing

Predictive Models

Exchange Rate
Web Search Score

Normalization
Over-sampling

KNN
SVM
Logistic Regression
Random Forest
Neural Network

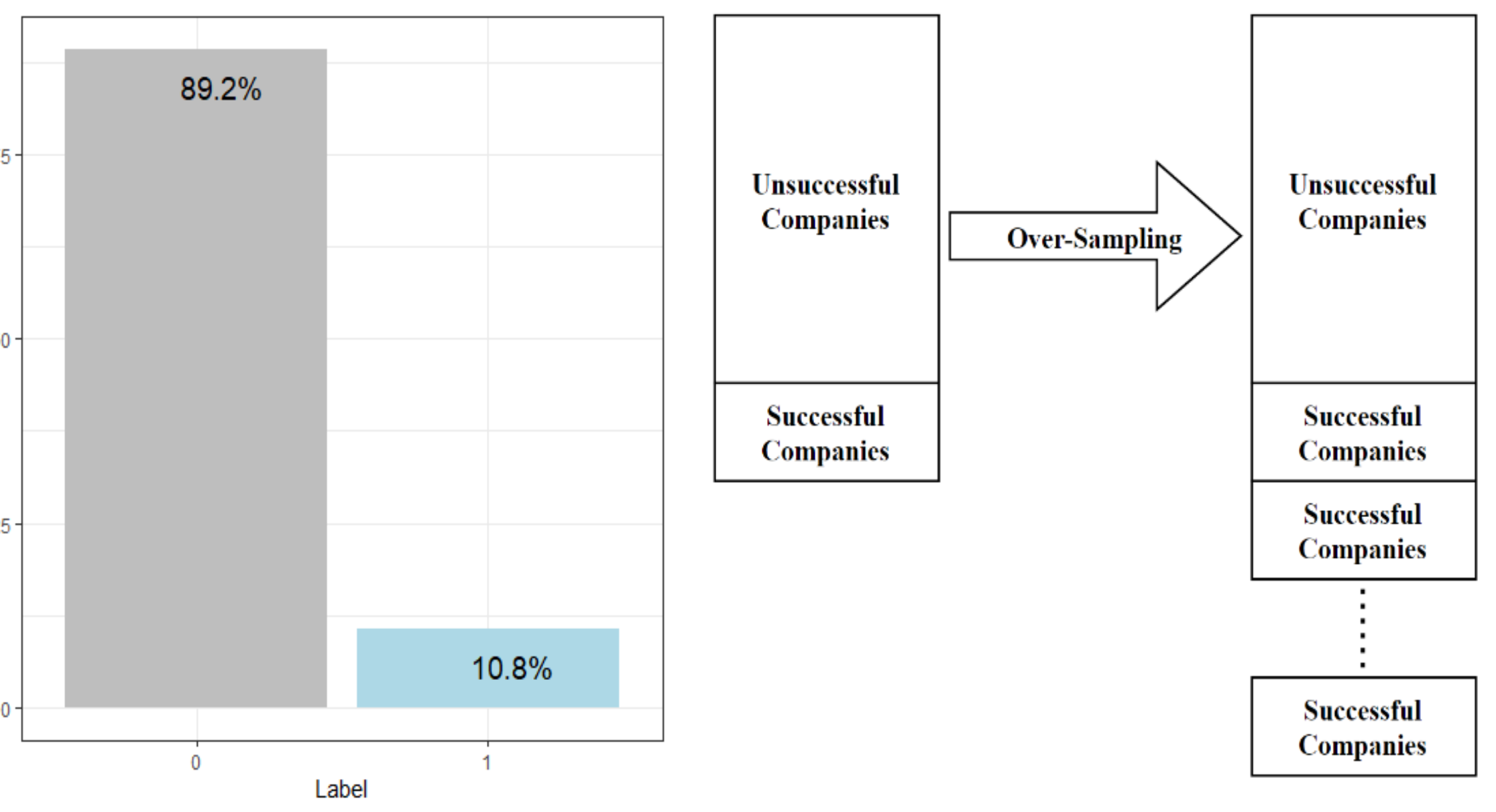
Normalization

Due to the requirement of many classification algorithms such as KNN, SVM, NNK and so on, we need to scale our features in a proper way. For numerical features, we use z-standardization method to normalize their ranges. The normalization algorithm follows the formula:

$$Z = \frac{x-\mu}{\sigma}$$

Over-sampling

Due to the issue of imbalanced label distribution, we need an effective method, over-sampling, to mitigate the problem. The figure below shows the issue and one way to implement over-sampling.



Unsuccessful Companies

Successful Companies

Over-Sampling

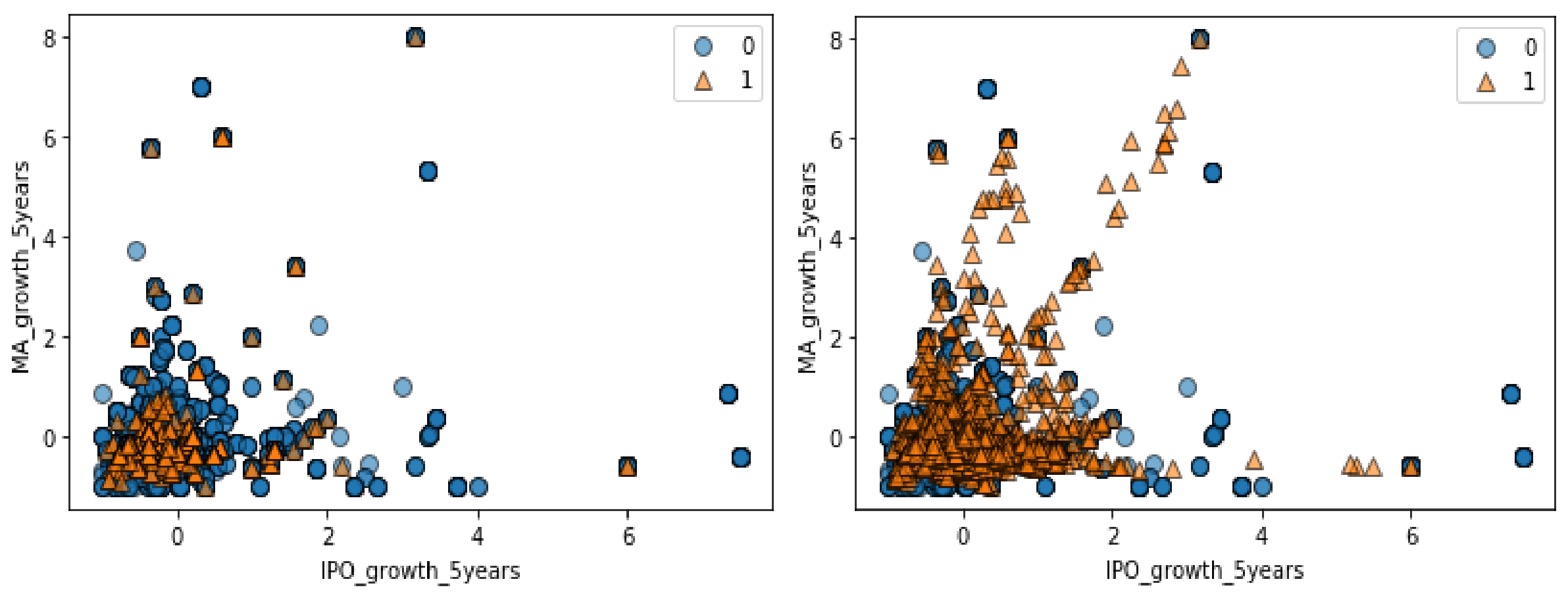
Unsuccessful Companies

Successful Companies

Successful Companies

Successful Companies

This KNN-based approach multiplies the successful companies in the training dataset to form a synthetic balanced dataset. The figure below shows the distribution of training labels before and after performing the over-sampling.



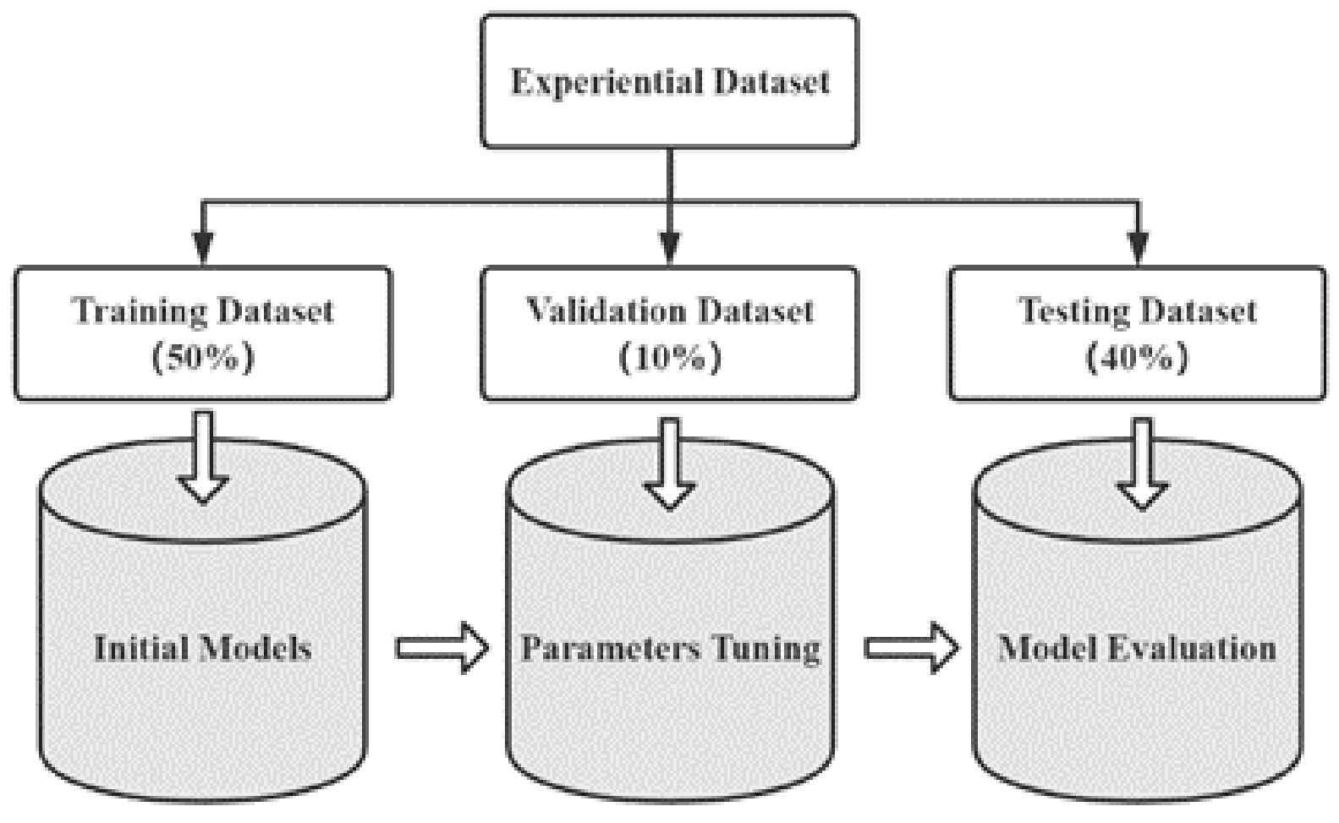
As a result, predictive algorithms will have more opportunities to learn the business success samples in the training process over this synthetic dataset.

Predictive Models

K-Nearest Neighbors	› n_neighbors=2
Logistic Regression	› C=8101
Support-Vector Machine	› kernel='linear', C=38
Random Forest	› n_estimators=111 › max_depth=29
Neural Networks	› 3 hidden layers › relu activation › threshold 0.8

Experimental Implementation

For each predictive model, we will firstly learn the model over the training dataset (50%), and then conduct the hyper-parameters tunning over the validation dataset (10%). Due to the cost of time and computation, here we only pick one or two parameters for each model and hope to obtain a set of moderate options after this iteration process.


Experiential Dataset

Training Dataset (50%)

Validation Dataset (10%)

Testing Dataset (40%)

Initial Models

Parameters Tuning

Model Evaluation

Results and Analysis

From Table II, we can see SVM has the best performance in terms of validation and testing accuracy, KNN has the best performance in terms of validation and testing F1-score, and LR is the weakest classifier among this model set.

TABLE II

ALL FEATURES: VALIDATION AND TESTING SET PERFORMANCE

Models	Validation Performance		Testing Performance	
	Accuracy	F1-score	Accuracy	F1-score
KNN	0.8103	0.8096	0.8260	0.8359
LR	0.6753	0.7284	0.6567	0.7235
SVM	0.8696	0.8183	0.8908	0.8489
RF	0.8088	0.8209	0.8144	0.8332
NNK	0.8412	0.8341	0.8547	0.8572

Based on hyper-parameters turning, most of models have a relatively moderate testing performance, and the performance difference between testing dataset and validation dataset is not large.

TABLE III

ROBUSTNESS TESTING: CROSS-VALIDATION ON 1.9K, 9.5K, 19K SAMPLES

Models	10% dataset (1.9K Samples)		50% dataset (9.5K Samples)		100% dataset (19K Samples)	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
KNN	0.8959	0.8655	0.8885	0.8560	0.8832	0.8489
LR	0.8933	0.8492	0.8951	0.8486	0.8913	0.8423
SVM	0.5134	0.5527	0.1717	0.1580	0.6314	0.6135
RF	0.8820	0.8603	0.8817	0.8592	0.8770	0.8536

Table III shows that over different datasets, the accuracy and F1-score values of KNN, LR and RF are consistent, which means such models are relatively robust and have a great potential of generalization. However, the performance of SVM shows highly visible changes over datasets with different sizes.

TABLE IV

FEATURES CONFIGURATIONS: TESTING PERFORMANCE COMPARISON

Models	Purely Company Features		Company Features + Exchange Rate Features		Company Features + Exchange Rate Features + Web Search Features	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
KNN	0.8028	0.8154	0.8113	0.8243	0.8260	0.8359
LR	0.5099	0.5955	0.6122	0.6873	0.6567	0.7235
SVM	0.1061	0.0254	0.8947	0.8496	0.8908	0.8489
RF	0.7370	0.7788	0.8421	0.8464	0.8144	0.8332
NNK	0.8738	0.8525	0.8730	0.8555	0.8547	0.8572

In Table IV, given the configuration of company-related features, exchange rate features significantly improved the testing performance of KNN, LR and RF. The effectiveness of web search features may be variable in different classifiers.

Potential Users Scenarios

Take LR and NNK as an example. For class A (successful companies), NNK has a higher precision of 31.57%. Given the great potential of generalization, when our model returns a positive prediction, we would be able to use this reliable probability in the calculation of the expected return of investment.

Besides, both models have an excellent precision for class B (unsuccessful companies). Under the assumption of good ability of generalization, if we receive a negative prediction from each model and they make the prediction independently, we will have strong confidence about its tiny possibility of IPO or M&A.

Acknowledgement

We would like to say thanks to those who had done previous research in the field of business success prediction. This project was inspired in part by their contributions.

Also, we want to thank all ECON 425 teaching staffs, including Prof. Xiaobai Liu and TAs for their help and assistant.