Machine Learning Prediction of Companies' Business Success

Chenchen Pan (cpan2@stanford.edu), Yuan Gao (gaoy@stanford.edu), Yuzi Luo (yuziluo@stanford.edu)

INTRODUCTION

Thousands of companies are emerging around the world each year. Among them, some are successful, been acquired or IPO, while others may vanished. What makes this different and lead to the different endings for the companies? In this project, we want to build a binary classification model to predict the success of companies. Previous work using similar dataset only compared the model between Logistic Regression and Random Forest. We explored K-Nearest Neighbours (KNN) classifier, and use F1 score as the metric to compare the models. And found KNN performs better on this task.

DATASET & METHODS

Dataset

The dataset we use is extracted from Crunchbase Data Export containing 60K+ companies' information updated to December 2015.

Logistic Regression

Logistic regression is s widely-used algorithm to model a binary dependent variable with many independent variables.

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}},$$

Random Forest

Random Forest is an ensemble learning method for classification with constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes.

K Nearest Neighbours

We classify an object by a majority vote of its K nearest neighbours.

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + \ldots + (x_n - x'_n)^2}$$

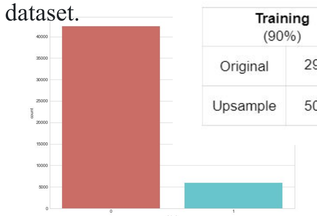
$$P(y = j|X = x) = \frac{1}{K} \sum_{i \in A} I(y^{(i)} = j)$$

EXPERIMENTS

Data Preprocessing

- Extracted and merged the companies' information from several original files.
- Labelled all the data with 1 or 0 based on the companies' status. 1 = Acquired or IPO; 0 = Otherwise.
- Edited, filtered and selected meaningful features.
  - category_list Audio|Mobile|Music
  - funding_total_usd 440000
  - country_code AUS
  - funding_rounds 3
  - Num_of_investor 3
  - funding_duration 425
  - first_funding_at_UTC 15461
  - last_funding_at_UTC 15886
  - label 0
- Used up-sample method to balance the training set.
- Normalized numerical features.
- Encoded text features using bag-of-words model.

This table below shows the number of training, evaluation and test data for original and up-sampled dataset.



	Training (90%)	Validation (5%)	Test (5%)
Original	29428	1635	1635
Upsample	50040		

Model Selection

we present three metrics:

- Accuracy: The proportion we have predicted right.
- F1 Score: $F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$

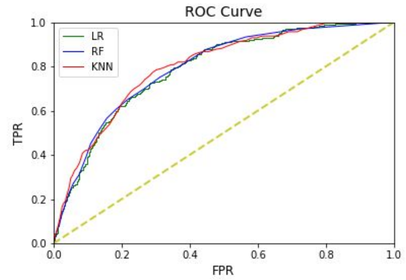
$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- AUC Score: Area under the ROC Curve, which is an aggregate measure of performance across all possible classification thresholds.
- TPR = TP / (TP + FN), FPR = FP / (FP + TN)

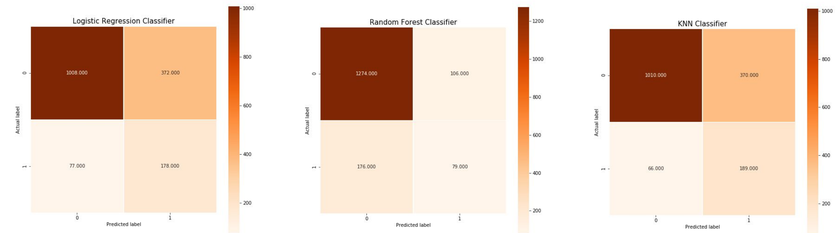
RESULTS

Random Forests have the best accuracy while KNN has the highest F1 score and the highest AUC score.

Classification Models	Evaluation Metrics (Validation Set)		
	Accuracy	F1 Score	AUC Score
Logistic Regression	72.54%	44.22%	79.00%
Random Forest	82.94%	36.16%	79.31%
K Nearest Neighbors	73.33%	46.44%	79.89%



For the confusion matrix, the TPR and FPR are 69.80% and 26.96% for Logistic Regression, 84.04% and 39.16% for Random Forests, and 74.12% and 26.81% for KNN respectively. Random Forests performs best on Confusion Matrix.



We selected KNN model to run on test set with:

Accuracy = 73.70% F1 score: 44.45%

FUTURE WORK

- Include more features of the companies, such as business description.
- Try more complex models, such as Neural Network and pre-trained word embedding.
- Try kernel method as moving the data to higher dimensional space.
- Explore some new questions, such as predicting the total funding size for a company (regression problem).

REFERENCE

- Wei CP, Jiang YS, Yang CS. Patent Analysis for Supporting Merger and Acquisition (M&A) Prediction: A Data Mining Approach[M]. Berlin: Springer, 2009: 187-200.
- Bento FRSR. Predicting Start-up Success with Machine Learning[D]. Lisboa: NOVA Information Management School, 2018. 9-83.