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.

### Logistic Regression

Logistic regression is a 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 trees. The model is as follows:

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

\[ P(y = j | x = x) = \frac{1}{K} \sum_{i=1}^{K} I(y^{(i)} = j) \]

### Model Selection

We present three metrics:

- **Accuracy**: The proportion we have predicted right.
- **F1 Score**: \[ \text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]
- **AUC Score**: Area under the ROC Curve, which is an aggregate measure of performance across all possible classification thresholds.

### 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 used to model a binary dependent variable with many independent variables.

**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 trees. The model is as follows:

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

\[ P(y = j | x = x) = \frac{1}{K} \sum_{i=1}^{K} I(y^{(i)} = j) \]

### 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
  - funding_total_usd
  - country_code
  - funding_rounds
  - Num_of_investor
  - funding_duration
  - first_funding_at_UTC
  - last_funding_at_UTC
  - label
- 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>29428</td>
<td>1015</td>
<td>1015</td>
</tr>
<tr>
<td>Upsample</td>
<td>50324</td>
<td>9135</td>
<td>9781</td>
</tr>
</tbody>
</table>

### Model Selection

- **Accuracy**: The proportion we have predicted right.
- **F1 Score**: \[ \text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]
- **AUC Score**: Area under the ROC Curve, which is an aggregate measure of performance across all possible classification thresholds.

### RESULTS

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

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