SMU SINGAPORE MANAGEMENT UNIVERSITY

HONG LEONG FINANCE

SMU-X

Hong Leong Finance - Credit Evaluation

Hong Leong Finance, the largest financial company in Singapore, seeks to implement data analytics to identify the top seven variables to predict if a loan would default

1 Data Cleaning

- 3 years (2015, 2016 & 2017) of quarterly data reflecting the characteristics of the loan was received
- Variables with only one value, no values, zero or NaN were removed
- Variables with more than 10% of NaN values were removed
- **Defining Defaulted Loans:** Loans labelled positively in either variables "Non-Performing Loan" or "Written Off"
- The data was **balanced** by replicating defaulted loans to ensure that models are not skewed towards identifying certain outcomes to avoid the accuracy paradox
- Data was analyzed quarterly to preserve the **time factor** in financial data
- Data was then split into training and testing data sets;
    - **Training Dataset**: 2015 & 2016; **Testing Dataset**: 2017

2 Feature Selection

Light Gradient Boosting	Pearson's Correlation Coefficient	Chi-Squared Distribution	Recursive Feature Selection	LASSO	Random Forest

The above six feature selection models were then ran on all the 2015 and 2016 quarters (training set) and based on the aggregate scores as per run in the model, the relevant variables earn a point each. Subsequently, the quarterly scores of each variable are then summed up and ranked decreasingly from the highest scoring variables to the lowest. The **top scoring variables** are then identified as the **top predictors**. The purpose of **ensembling six different feature selection methods** is to provide a more **dynamic model** and generate **objective results**.

3 Result Analysis

- To ensure the selection of the best predictors, the **multicollinearity** possibility needs to be thoroughly examined and reviewed. This is essential as most model functions assume that the identified variables are independent.
- The multicollinearity checks were conducted through the utilization of the **Variance Inflation Factor (VIF)**. If the **VIF score < 10**, it indicates low correlation which is ideal, whereas **VIF ≥ 10** indicates potentially high correlation and in the ideal scenario, is solved by replacing the highly correlated variable with the next best variable.
- However, due to the **limitations of real-life data** and **tradeoffs of data losses**, it is inevitable that certain variables with high predictive power will still possess a VIF of higher than 10.
- As such, after the analysis of multicollinearity, the seven identified variables, though not necessarily the top seven scoring variables, were still in the **top 15** out of hundreds of variables.

4 Evaluation

Five predictive models were then deployed to **evaluate the predictive power** of the seven variables by running the models on the 2017 data (testing set). As **false negatives** (i.e. loans that defaulted but not identified as defaulted by the model) have serious implications on the client, the **recall score** and the **overall accuracy score** are prioritized. The predictive models used, alongside their relevant test set results are as follows:

Random Forest	Logistic Regression	XGBoost	Naïve-Bayes	K Nearest Neighbours
Recall: 0.463	Recall: 0.894	Recall: 0.945	Recall: 0.922	Recall: 0.740
Accuracy: 0.595	Accuracy: 0.823	Accuracy: 0.861	Accuracy: 0.825	Accuracy: 0.797

5 User Interface

- A web-designed user interface was constructed using commonly available tools to ensure better integrability. The tools used include the following: HTML, CSS, JavaScript, PHP
- **Purpose of User Interface**: To allow for easy generation of credit risk levels. The scores generated aims to provide the user with a more mathematically supported analysis and should be treated as supplements to the employees' discretion/business decision.
    - **Home Tab**: Allows for mass prediction of a large amounts of data through a file upload
    - **Individual Entry Tab**: Allows for individual checking of customer credit rating through a single data entry
- Probability of default is calculated by finding the average of the accuracy score of all five algorithms multiplied by the binary score of the respective predicted result of default.

Ekaphat Mathiprechakul | Lim Wei Jie | Malcolm Lai Chun Hao | Vanessa Gan Hui Qi | Wendy Choa Yen Yi | Wong Chiu Theng