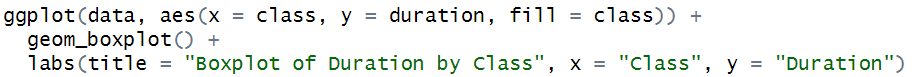
## 3.3 Objective 3: To investigate the relationship between duration and purpose with credit class (Douglas Chew Xi Zhi TP075339)

### 3.3.1 Analysis 3-1: How is the distribution between duration, purpose and credit class?

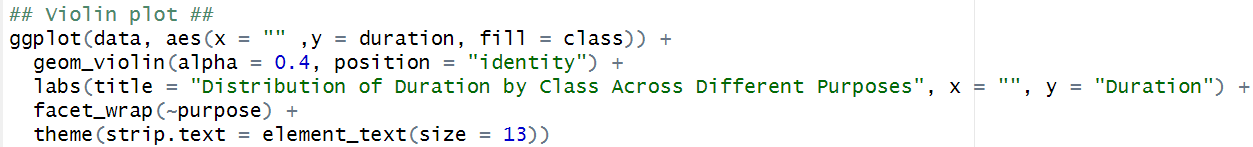
#### 3.3.1.1 Boxplot of Duration by Class (Aggregated purpose)

 A graph with red and blue squares

Description automatically generated

From the boxplot above, it shows that both distributions are slightly skewed to the right due outliers at the right side. Besides, it also shows that mean of duration for bad credit class is 25 which is higher than good credit class which is 19. This indicates that the higher the loan duration, the higher the credit risk.

3.3.1.2 Distribution of Duration by Class Across Different Purposes

A collage of images of different shapes

Description automatically generated

From the violin plot, it shows the density of good credit class become narrower across the loan duration in most of the purposes except other. This is because when loan duration increases, it indicates that loan borrower unable to repay the debt in short period. Moreover, the peak density of good credit class for each purpose mostly indicates that loans at that duration have lower risk. Besides, the extension of bad credit class for radio/tv indicates outliers exist in distribution of duration by class for radio/tv. The outlier is abnormal because longer loan duration means high credit amount but cost of radio/tv are mostly lower than car.

### 3.3.2 Analysis 3-2: Is there any association between purpose, duration with class?

#### 3.3.2.1 Hypothesis testing for purpose with credit class

A screen shot of a computer code

Description automatically generated A yellow and black text

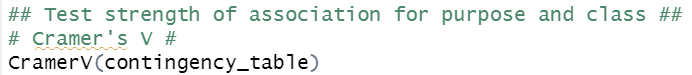
Description automatically generated

H0 : No association between purpose and credit class.

H1 : There are association between purpose and credit class.

The hypothesis testing is being conducted by using chi-square test of independence. Assume the significance level is 0.05. Since the p-value obtained is smaller than 0.05, we reject null hypothesis. Hence, we can conclude that there is association between purpose and credit class.

#### 3.3.2.2 Strength of association between purpose and credit class

A close up of a number

Description automatically generated

Cramér’s V is being used to investigate the strength of association between categorical variables. Since Cramér’s V is 0.3525031, we can conclude that the strength of association between purpose and credit class is medium.

#### 3.3.2.3 Hypothesis testing for duration with credit class

A close-up of a text

Description automatically generatedA screenshot of a computer

Description automatically generated

H0 : Duration has no effect on credit class.

H1 : Duration has effect on credit class.

The hypothesis testing is being conducted by using logistic regression. Assume the significance level is 0.05. Since the p-value obtained is smaller than 0.05, we reject null hypothesis. Hence, we can conclude that duration has effect on credit class.

#### 3.3.2.4 Strength of association between duration and credit class

A close-up of a computer code

Description automatically generated A close-up of a white background

Description automatically generated

Point-Biserial Correlation is being used to investigate the strength of association between continuous variable and binary variable. Since Point-Biserial Correlation is 0.3016942, we can conclude that the strength of association between duration and credit class is medium.

### 3.3.3 Analysis 3-3: Which model performs with duration and purpose as predictors for credit class?

A close-up of a computer code

Description automatically generated

In this analysis, bad credit class will act as positive class.

3.3.3.1 Train test split

A screen shot of a computer code

Description automatically generated

Train test split is a technique that splits dataset into training set and testing set. It allows us to evaluate if the predictive model that trained will training set overfitting occurred on unseen data.

#### 3.3.3.2 Setup K-fold cross-validation

A close up of a number

Description automatically generated

K-fold cross-validation is a technique used to reduce bias and prevent overfitting. Since the data entry is 6000, the number of folds is set to 10 for more reliable estimates.

#### 3.3.3.3 Logistic regression

A screenshot of a computer program

Description automatically generated

Logistic regression is supervised machine learning algorithms that can be used for predicting the probability of certain occurrences. In this analysis, it is being used to predict if credit class is bad based on duration and purpose.



From the performance metric of logistic regression above, the accuracy of 70.3% with unseen data is good and the difference between training accuracy and testing accuracy is 0.5%. It indicates that the model is neither overfitting nor underfitting. Besides, the model has good precision of 70.6% and moderate recall of 69.7% which means 70.6% of predicted bad credit are correct while only 69.7% of actual bad credit is being identified by the model. This resulted in 70.1% of F1 score for logistic regression. Lastly, the Area under the receiver operating characteristics (AUROC) of the model is 75.7% means there is 78% chance the model distinguishes bad and good credit class correctly.

#### 3.3.3.4 Random forest

A screenshot of a computer program

Description automatically generated

In this analysis, random forest is being used as predictive model to predict credit class based on duration and purpose.



From the performance metric of random forest, the accuracy of model is 76.9% with unseen data and the difference between train accuracy and test accuracy is small which indicates that the model is neither overfitting nor underfitting. The model had good precision of 80% and recall of 71.8% means the model 80% of predicted bad credit class is correct while 71.8% of actual bad credit class being identified. The F1 score of random forest is 75.7% and AUROC of 82.3% means there is 82.3% chance for the model to distinguish bad and good credit class correctly.

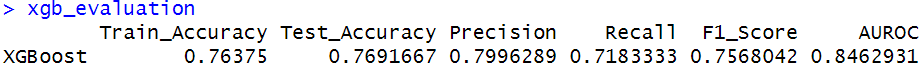
#### 3.3.3.5 EXtreme Gradient Boosting

A screen shot of a computer code

Description automatically generatedA screenshot of a computer program

Description automatically generated

EXtreme Gradient Boosting (XGBoost) is a powerful technique that combines gradient boosting, regularization and computational optimization. In this analysis, it is used to predict the credit class with purpose and duration with L2 regularization and cross validation to prevent overfitting.

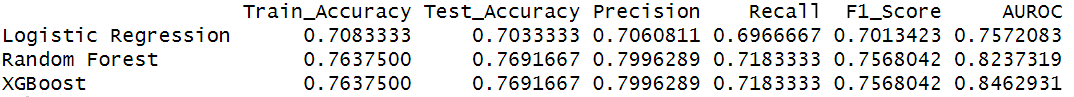


From the performance metric of XGBoost, the model has good accuracy of 79.9% when tested with unseen data and the difference between train accuracy is test accuracy is close which indicate the model is neither overfitting nor underfitting. The model had good precision of 80% and recall of 71.8% means the model 80% of predicted bad credit class is correct while only 71.8% of actual bad credit class being identified. The F1 score of the model is 75.7% and the AUROC of 84.6% means the model have 85% of chance in classify the credit class correctly.

#### 3.3.3.6 Overall model evaluation

A screenshot of a computer code

Description automatically generatedA graph of different colored lines

Description automatically generated

With duration and purpose as predictors of credit class, random forest and XGBoost had the best accuracy is 76.9% without hyperparameter tuning for any predictive models. With the aid of performance metric table and ROC curve plot, XGBoost has better performance in credit risk classification because it had higher AUROC and closer to top-left corner in the ROC curve plot than random forest. In credit risk classification, recall of the model is more important than accuracy because it is costly when predicting bad credit class as good credit class. Hence, the performance for XGBoost in credit risk classification is about moderate. To minimize the likelihood of missing bad credit risk, predictors with high feature importance value should be included and perform hyperparameter tuning to optimize the model.

### 3.3.4 Analysis 3-4: Does credit amount improve the performance of XGBoost?

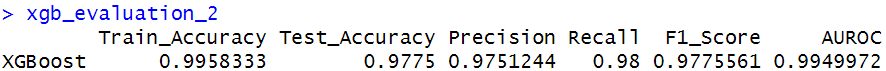
#### 3.3.4.1 Train new XGBoost with duration, purpose and credit amount as predictors

A computer screen shot of a code

Description automatically generatedA screenshot of a computer program

Description automatically generated

From the code above, credit amount is being added into the feature lists and new XGBoost model is being trained.

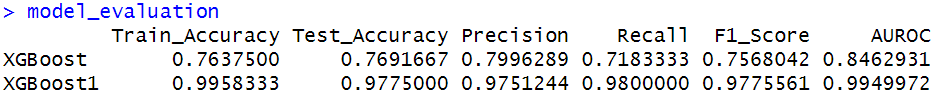


From the performance metric of new XGBoost, the model has good accuracy of 97.8% when tested with unseen data and the difference between train accuracy is test accuracy is about 1.8% which can still indicate the model is neither overfitting nor underfitting. The model had good precision of 97.5% and recall of 98.0% means the model 97.5% of predicted bad credit class is correct while only 98.0% of actual bad credit class being identified. The F1 score of the model is 97.8% and the AUROC of 99.5% means the model has 99.5% of chance in classify the credit class correctly.

#### 3.3.4.2 Overall model evaluation

A screenshot of a computer code

Description automatically generatedA graph of a curve

Description automatically generated

After adding credit amount as predictor, the XGBoost predictive model had improved significantly. The ROC curve of the new XGBoost model is very close to the top-left corner. Besides, the accuracy of XGBoost model increased from 76.9% to 97.8%. Moreover, recall of the XGBoost model increased from 71.8% to 98.0% which means only 2.0% of the actual bad credit class are not being identified. This significantly minimized the risk of missing any actual bad credit class. In conclusion, the newly trained XGBoost model performs very well and is suitable for credit risk classification.

### 3.3.5 Analysis 3-5: What is the optimal loan duration for lowest credit risk if business and credit amount is given?

#### 3.3.5.1 Optimization Technique

A screenshot of a computer program

Description automatically generated

Optimization technique is being applied by finding the loan duration with minimum probability of bad credit risk with the XGBoost model with accuracy of 97.8%. It will create dummy data with conditions given to predict the optimal duration to minimize the credit risk. If the risk of loan is higher or equal to 50%, it will be rejected.

A close-up of a car

Description automatically generatedA white background with blue text

Description automatically generated

From the result above, loans with same credit amount but different purposes have different optimal duration and credit risk. Loan for business have higher optimal duration and risk than loan to buy a car. This is because operation of business cost money constantly, to reduce the risk and burden of monthly payment, the optimal duration for business loan is longer than loan for buying used car although the risk is still higher than loan for buying used car.

A white background with blue text

Description automatically generatedA white background with blue text

Description automatically generated

From the result above, both loan purposes is education have the same optimal duration, but the loan with credit amount of RM3000 has higher risk than education with credit amount of RM4000. It seems counterintuitive but borrowers that request loans with low credit amount might have weaker financial stability and profiles.



From the result above, the loan with amount of RM300000 is being rejected because the minimum probability of risk reached 0.63 which might cause financial loss when the loan is approved.