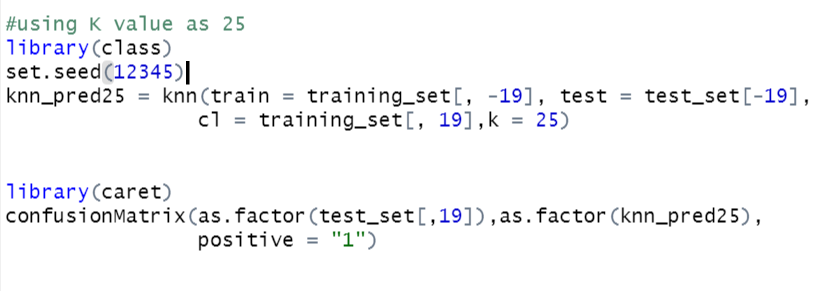
# MACHINE LEARNING ANALYSIS

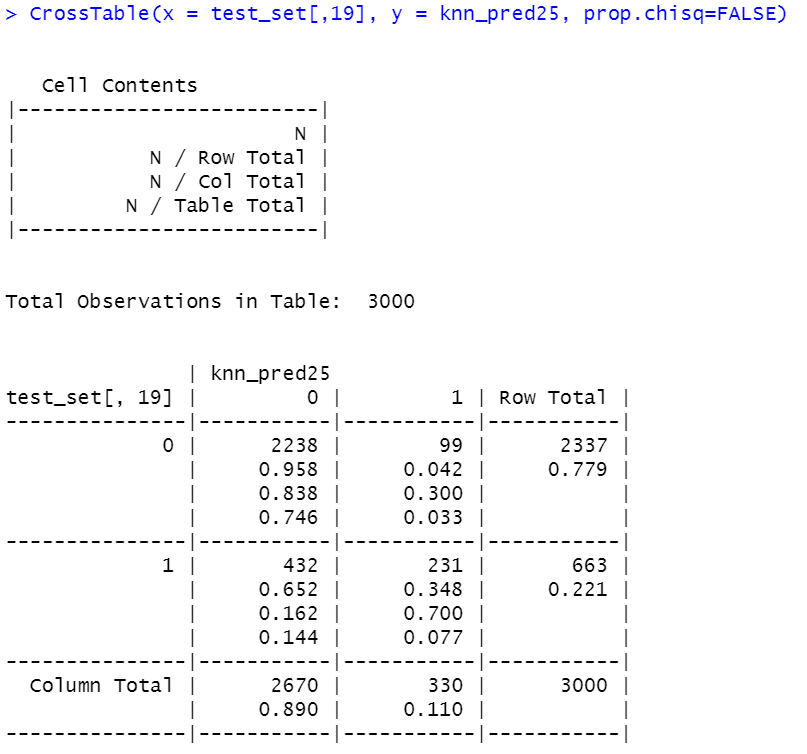
**Assumptions:** The dataset shows that the data has been collected over a period of six months and does not contain any null values. Additionally, we assume that the data was collected in accordance with a consistent schedule and that it is accurate.(Alam *et al.*, 2020) Throughout our Supervised learning technique, the confusion matrix was used to evaluate our work.  Refer Appendix D.

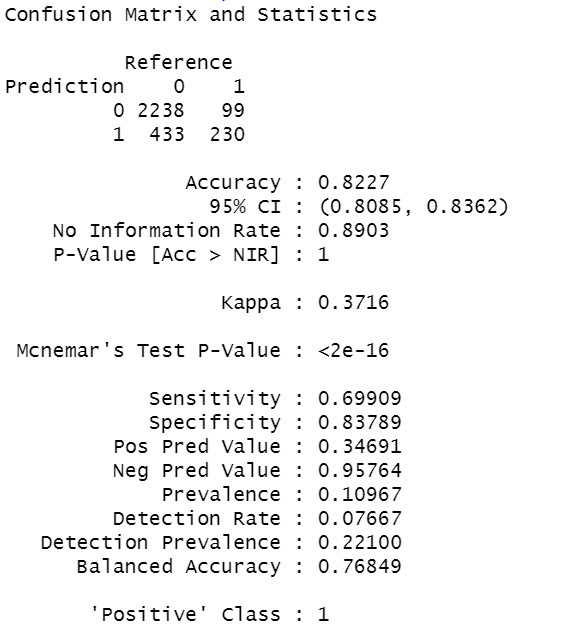
## K- NEAREST NEIGHBOUR

KNN was selected for this analysis since it is one of the most basic classification methods. In the correlation matrix, the top 19 predictors were selected and the attributes were then scaled using the min-max scaling technique, meaning that the ranges of the variables varied from 0 to 1. Various K values were then chosen in order to get the optimal result in the code.

K value was initially set to 124, which is the root square of the training set, which is 21000. The results were quite appreciable but the Kappa value was only 0.36. Additionally, there is a possibility of overfitting of the model. We reduced K and experimented with different values of K. K = 25 produced the best results.





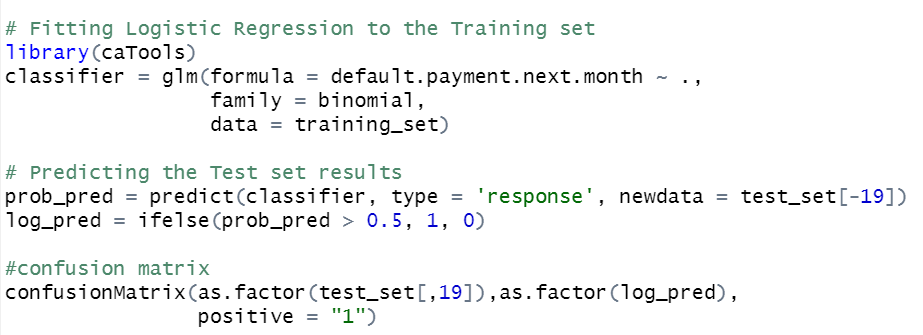


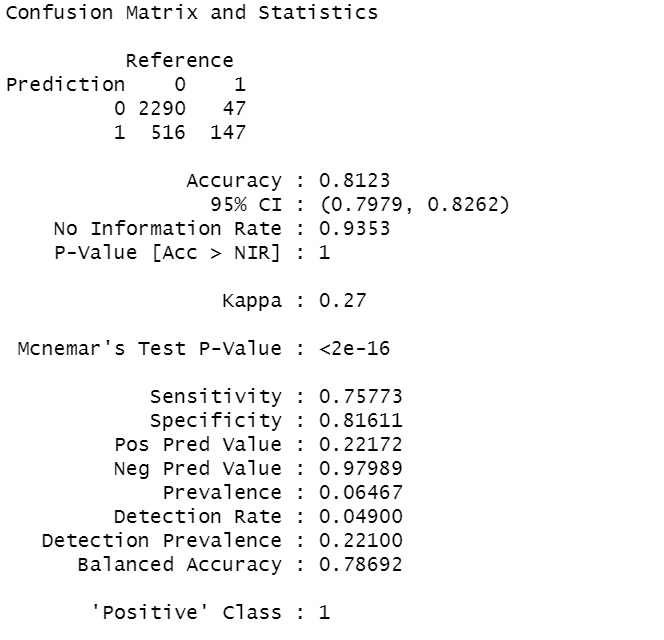
*Fig 15: The confusion matrix of KNN with the K value as 25*

## 

## LOGISTIC REGRESSION

Since this is a binary classification problem, we ran a simple logistic regression. The output of our analysis is a probability range ranging from 0 to 1, which reflects default probabilities. As a result, we divide them into halves where less than 0.5% goes into the non-defaulter category, and more than 0.5% is classified as a defaulter. The reason to choose Logistic Regression for this analysis was that it gives good accuracy and also provides high training efficiency compared to other models. Due to the nature of the problem, there was no requirement to normalise the data.

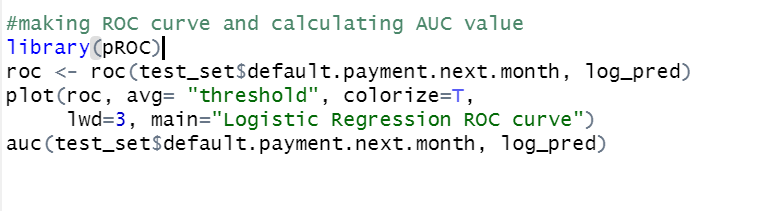


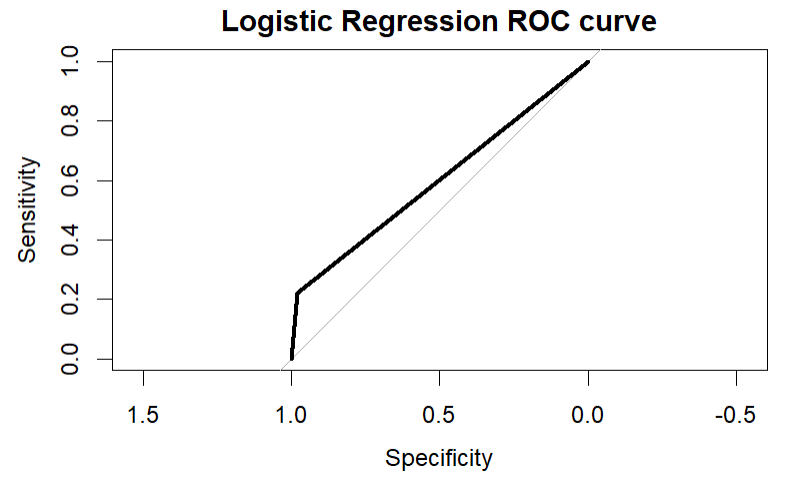


*Fig 16: The confusion matrix of Logistic Regression*

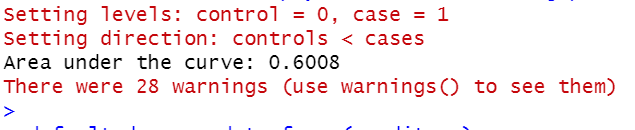
The model was further evaluated by plotting the ROC curve. The Receiver Operating Characteristics (ROC) is one of the performance measures that can be used to measure the efficiency of a model. (Hamori *et al.*, 2018b) By calculating ROC, it can be determined how well a model is capable of discriminating between classes. On the graph below, it can be seen that the True Positive Rate (TPR) is plotted on the y-axis, and the False Positive Rate (FPR) is plotted on the x-axis.

Tests can be evaluated by measuring their area under the ROC curve (AUROC), which is a measure of their discriminatory ability. Area Under Curve reflects how well a model differentiates between the positive and negative classes, so the higher the AUC, the better the performance. AUC = 1 means that the classifier is able to differentiate the Positive from the Negative perfectly. The classifier would have predicted all Negatives as Positives, while all Positives would have been predicted as Negatives if the area under curve had been 0. There is a high probability of being able to distinguish between positive and negative class values when 0.5<AUC<1 is used. This is because the classifier can detect many more True positives and True negatives than False negatives and False positives.





*Fig 17: ROC curve for Logistic Regression*

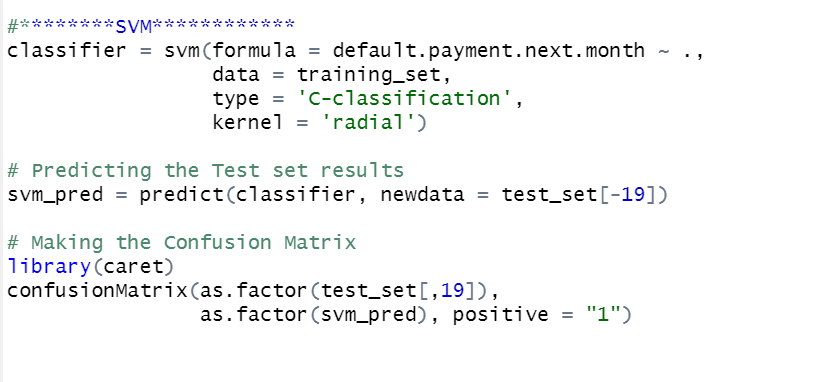


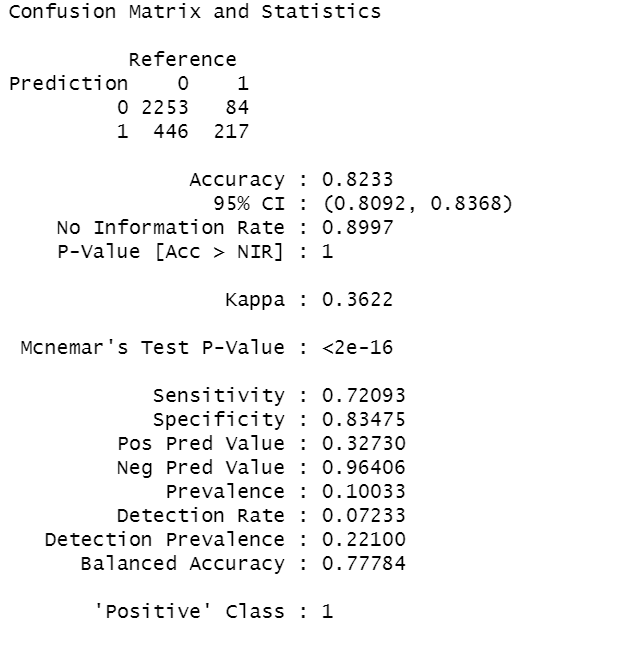
## SUPPORT VECTOR MACHINE

SVM is regarded as one of the best classification models as it gives the maximum accuracy. This particular method chooses "radial" as the kernel. When it comes to non-linear data in SVM, Gaussian Radial Basis Function (RBF) is considered a top kernel.(Kečo et al., 2018)

 There is a need to normalise the data as this is a model based on linearly separable distances. This will ensure that all the components have the same range bound. As mentioned in the KNN classifier, the Min-max scaler is also used here for normalisation.

**Confusion Matrix**



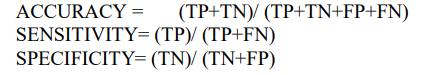


*Fig 16: The confusion matrix of Radial SVM*

# EVALUATION

Throughout our Supervised learning technique, the confusion matrix was used to evaluate our work.   An analysis of a classification model is conducted by analyzing the confusion matrix. The number of correct and incorrect predictions are broken down by class and summarized with count values. The key terms used in the prediction of the evaluation metrics are: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Main metrics used to evaluate the model are:

* Accuracy - A model's accuracy is the proportion of predictions it correctly classifies.
* Sensitivity -An indicator of proportion of actual positive cases which are predicted to be positive (TP) (Gupta *et al.*, no date)
* Specificity - the ratio of actual negatives to those predicted as negatives(TN)
* Kappa value- A metric that compares observed accuracy to expected accuracy. This is important factor here as this is an imbalanced dataset.



Using the following metrics, the model is compared.

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics | KNN | Logistic Regression | SVM |
| Accuracy | 0.822 | 0.812 | **0.823** |
| Sensitivity | 0.699 | **0.757** | 0.720 |
| Specificity | **0.837** | 0.816 | 0.834 |
| Kappa | **0.371** | 0.27 | 0.362 |

**Accuracy**: SVM has the best accuracy of the three models. But it differs from KNN only by 0.1%, making it equally good on this metric.

**Kappa:**Since the dataset is imbalanced, the Kappa value is an important factor to consider. According to Landis and Gary, Kappa values between 0.2 and 0.4 are considered fair. The model with the highest Kappa value is the KNN.

**Specificity:**The higher the specificity, the higher the chances of the model correctly predicting the defaulter. In this case, a KNN model is considered good because of its high specificity.

**Sensitivity:**The higher the specificity, the better chance of predicting a non-defaulter, client who will pay on time. Logistic Regression wins here.

**Machine Performance**: There are a few other parameters to consider, such as model execution time. Compared to other models, SVM provided high accuracy but took longer to execute. This dataset consisted of only 30,000 records. However, in the real world, banks run millions of records to find defaulters, making SVM a failure.

**AUC for ROC:**Although the sensitivity for Logistic Regression was higher, the AUC turned out to be poor, making this model unfit. In general, an AUC value near 1 is considered a good model for a dataset. Nevertheless, the value in this case was only 0.6, which makes it unreliable.

The K-Nearest Neighbor (KNN) model is the best predictor of default credit card client after carefully analyzing all the models and weighing the pros and cons.

# CONCLUSION

Default payment in Taiwan's credit card industry was investigated in this study. Machine learning techniques were employed in order to predict the pattern of the client's behaviour in paying off the credit card bill. As a start, it's worth mentioning that there are several ways to approach this classification problem. When the complexity of the approach is varied, different approaches can yield similar results, as demonstrated within this paper. It is primarily because the problem's data isn't highly complex. In order to reach a conclusion on the best model, various steps were implemented including data cleaning, feature scaling, feature selection, and training the data on machine learning algorithms.  According to accuracy, Kappa values, Area Under the Curve of ROC, and other statistical values, K-Nearest Neighbour was the best model to predict the outcome.

After conducting this analysis, a conclusion was drawn that while selecting a model a myriad of other factors should be take into consideration much more than just accuracy. Focus should be put on non-technical metrics like machine performance and the time that it takes to execute a model in order to make a decision about the suitability of the model, because in the real world there can be limitations.

On further analysis it was evident that the dataset was highly imbalanced, which was one of the biggest limitations of the study. This imbalance reduced the performance of the model. As a next step towards making a better accurate model, a resampling technique can be implemented in order to implement various machine learning algorithms in order to not only classify defaulters but also predict the probability of default in a way that will be more suitable for banks when formulating various policies.