# MINI PROJECT 2

Philip Thomson

**Problem Statement**:

In this project, the 'Card Transactions' datasets are used to identify credit card transactions as fraudulent or legitimate, as well as to learn how to fit all of the different types of models and produce performance curves for all of them.

**Approach**:

This study examines the prediction accuracy of probability of default among three data mining approaches in the instance of customers' default payments in Taiwan. In terms of risk management, the predictive accuracy of the projected likelihood of default will be more beneficial than the binary categorisation result - credible or not credible clients. Because the true chance of default is unknown, this study proposed the unique "Sorting Smoothing Method" to estimate the true likelihood of default. The simple linear regression result (Y = A + BX) with the real probability of default as the response variable (Y) and the predictive probability of default as the independent variable (X) shows that the forecasting model produced by artificial neural network has the highest coefficient of determination; its regression intercept (A) is close to zero, and regression coefficient (B) is close to one. As a result, among the three data mining approaches, only the artificial neural network can properly predict the true likelihood of default.

**Dataset**:

As the response variable in this study, a binary variable, default payment (Yes = 1, No = 0), was used. The 23 other attributes in the dataset were used as explanatory variables in this exploratory study.

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

**Training Models**:

In this project, a model-based classification approach was chosen.

To model, three train classification learning methods logistic regression, SVM, , decision tree and one ensemble learning technique random forests were chosen.

**Conclusion:**

The largest group of credit limits appears to be for amounts up to $50,000.Most of defaults are for credit limits 0-100,000 .

The proposed Random Forest model successfully predicts the target variable Default of Credit Card Clients in the test data by 86 percent.

**Random Forest's prediction model provides the highest accuracy and performance with the lowest error rate. As a result, Random Forest is a solid choice among the other data mining algorithms for accurately predicting the default payment.**