Georgia State University

J. Mack Robinson College of Business



Default of Credit Card Clients

Fall 2021

By: Ameya Deo | Gautam Jain | Sanchit Agrawal | Yash Mitkari

**Project Description:**

**Context**

This project is about predicting the Credit Card Defaults Using KNN, Naïve Bayes, Logistic Regression, Neural Networks, and Ensemble Models using the “Default of Credit Card Clients” Dataset from Kaggle (<https://www.kaggle.com/sophiazzj/default-of-credit-card-clients>) based on 30000 observations with 24 explanatory variables describing (almost) every aspect of credit card defaults.

**Goal:**

The project is aimed at implementing a model(s) to predict the credit card defaulters through the steps as shown below:

* Assemble the data and explore it
* Clean variables, build what is needed
* Models: Logistic Regression, KNN techniques, Naïve Bayes, Ensemble Learning & Neural Networks
* Choose the best model having best accuracy.

**Business Problem:**

In general, a banking institution faces issues related to distinguish between their borrowers that are likely to repay their credit card debt and borrowers that are likely to default on their debt. A credit card holder “defaults” on their debt when they make purchases with their credit card that they fail to repay or fail to pay before a given window (usually 30-45 days). Usually, most banks and lenders follow the same business model of defaulting, and it is even an aspect of how such institutions make money, there still must be a way to identify such borrowers, to lend responsibly. A bank that only lends to defaulters is likely to fail. At the same time, defaulters generate profits for banks in the late payments that they accrue and at the same time defaulters credit score/history affected. We leave the specific applications and uses of our model to those with domain knowledge.

Task is to perform exploratory data analysis and machine learning model for predicting default of credit card clients using R that can the prediction for the same.

**Data Exploration and Preprocessing:**

Our dataset includes information for 30,000 credit card users from Taiwan, whether they defaulted, and twenty-four independent explanatory variables

Identify the output variable and input variables: Default Payment Next Month is our target variable and the dependent variable for prediction. We have twenty-four independent features like Limit balance, sex, education, age, bill amount, pay amount etc.

We have divided the dataset further into training and validation as per 60:40 ratio.

**Data Cleaning:**

* Removed rows which were having unknown values for features like marriage and education.
* Dropped rows with Nan values.
* Dropped index column

We have also renamed the columns for Bill amount and Pay amount per months. We did some preliminary visualization and exploration of our data to get a sense for its overall shape and the distribution of various demographic categories.

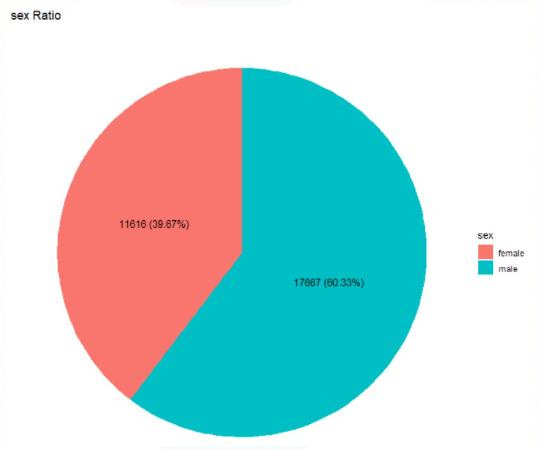
**Graphs or plots to explore the data:**

1. **Defaulter’s by Ratio:**

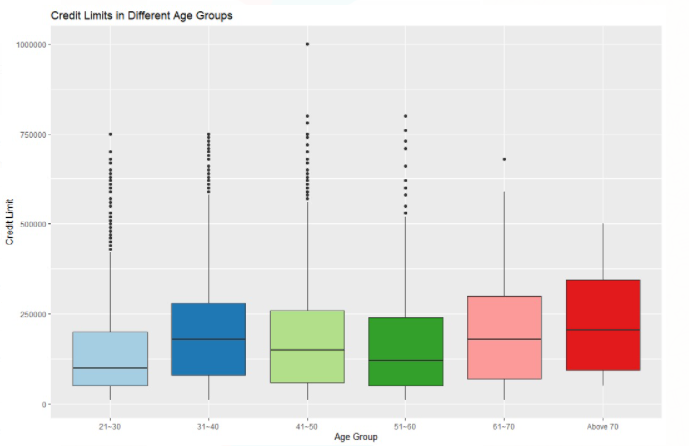
Chart, pie chart

Description automatically generated

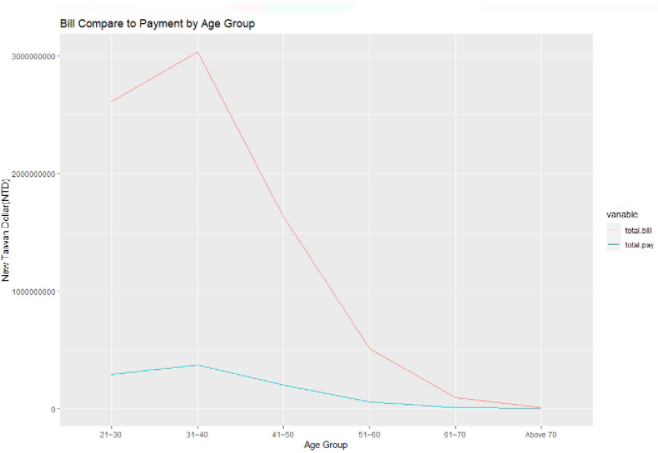
2. **Defaulter’s by Sex:**



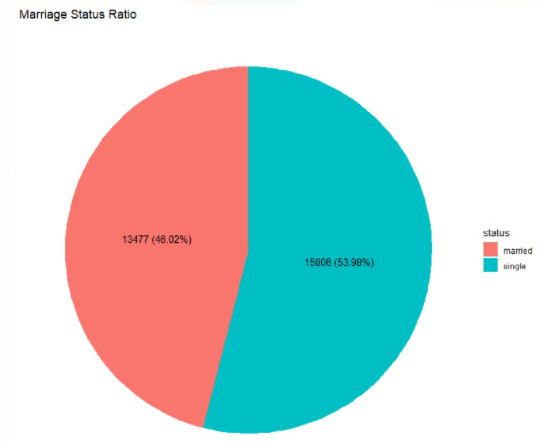
3. **Age Groups Difference:**



4. **Payment by Age Group:**



5. **Defaulter’s by Marriage:**



**Models and their comparison:**

We have implemented the below models:

* Logistic Regression
* Classification using K-Nearest Neighbors
* Neural Network
* NaïveBayes
* Ensemble method

## 

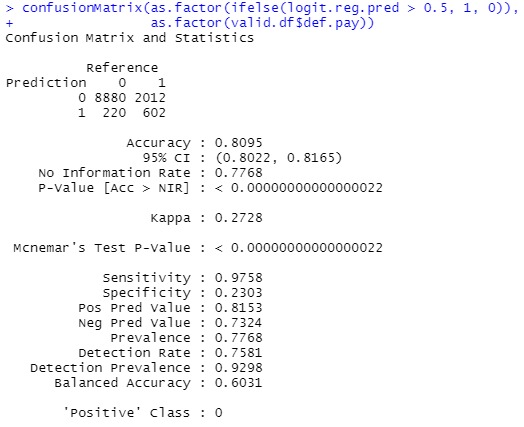
## Reasons for specific Model Selection:

## Logistic Regression:

Over the entire data set, around 22% were defaulters. We will be using Confusion Matrix as metrics for assessing Logistic regression model accuracy.

The Logistic Regression model on the testing data gives an **accuracy** value of **80.95%.**

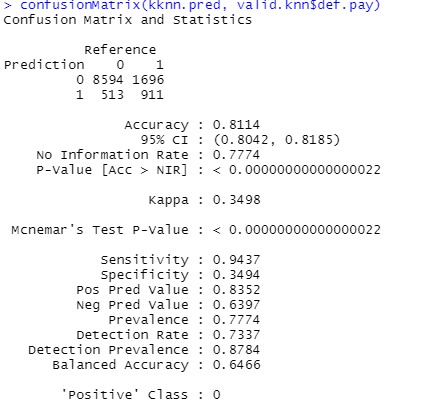
With 97.5% of customers that the model identified were paying credit card bills on time and 23.03% of customers that the model identified as defaulters.



**Classification using K-Nearest Neighbors:**

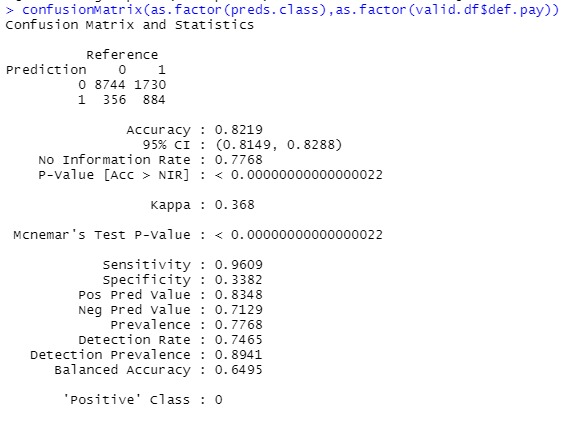
KNN stands for **K-Nearest Neighbors**. It is a supervised learning algorithm. It is often used as a benchmark for more complex classifiers such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM). We have used 21 independent features for KNN implementation. A robust implementation must consider feature engineering, data cleaning, and cross-validation. The **K-Nearest Neighbors** model on the testing data gives an **accuracy** value of **77.01%.**

We have implemented KNN with different optimal weights by changing k values and this time the **accuracy** we achieved is **81.14%.**



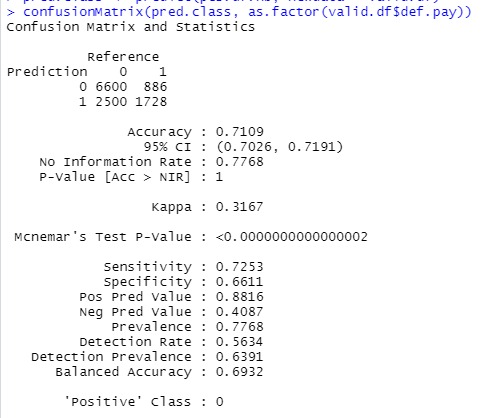
**Neural Network:**

Neural networks are a class of machine learning algorithms used for complex patterns in datasets using multiple hidden layers and non-linear activation functions. They are also known as artificial neural networks (ANNs) or simulated neural networks (SNNs). We have implemented in our scenario and the **accuracy** we achieved for the testing set is **82.19%.**



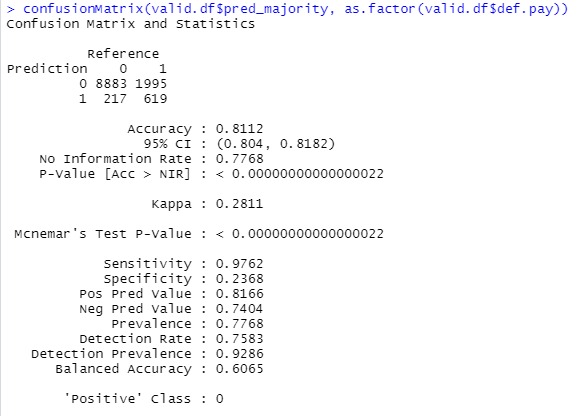
**Naïve Bayes:**

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. After testing with Naïve Bayes, we got an **accuracy** of **71.09%.** This is the lowest accuracy we have achieved in this project.



**Ensemble:**

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. We have implemented ensemble techniques with three models: Logistic Regression, Neural Network and KNN in this project. The **accuracy** we attained is **81.12%.**



## Model Comparison:

## Below is the accuracy for all the five models implemented in the project:

A picture containing text, transport, wheel

Description automatically generated

Neural Network performed the best with an accuracy of 82.19% followed by KNN and Ensemble with an accuracy of 81.14% and 81.12% respectively. Logistic regression performed good with 80.95% but naïve bayes performed worst in all the models with 71.09% accuracy.

# **Result:**

Results from Neural Network came out to be best followed by KNN and Ensemble. Naïve Bayes has the worst result, maybe this model is not fit for this scenario.