**CREDIT SCORING PROJECT REPORT**

**MARY MAINA**

## BACKGROUND INFORMATION

This project aims at creating a prediction model for credit card scoring. The machine learning model is to predict if an applicant is a good or bad client based on the applicant’s personal information and a record on their behaviour of credit cards. In this project we have two datasets; the application record dataset and the credit record dataset. The ID column is the unique identifier for the datasets. The Application Record dataset contains the respondents’ personal information while the Credit Record dataset has the respondents’ credit behaviour.

The total number of applicants is two hundred and seventy thousand, eight hundred and five (270805). Total number of columns for the Application Record table is eighteen (18) and three (3) for the Credit record table.

Project source: Kaggle

Link: <https://www.kaggle.com/rikdifos/credit-card-approval-prediction>

Date of Submission: 24th/May/2024

## METRIC OF SUCCESS

1. Perform data wrangling and exploratory data analysis to understand the data provided and write a report.
2. To come up with a highly accurate machine learning model that can do predictions on real time incoming data and write a report.

## EXPERIMENTAL DESIGN

1. **Data collection.**

Extracting data from Kaggle website.

1. **Data cleaning.**

This includes checking for missing values, removing duplicates and irrelevant variables and merging data sets.

1. **Exploratory Data analysis.**

Exploring the data through visualizations to give insights on the structure of the project.

1. **Machine learning model.**

A working ML model that can do predictions on real-time incoming data.

1. **Summary and recommendations**

## DATA CLEANING.

**Application Record dataset**

All column names were converted to lowercase, then some of the columns were renamed.

All categorical data were then converted to numerical form for easy analysis and modelling using sklearn LabelEncoder. Sklearn provides a very efficient tool for encoding the levels of categorical features into numeric values. LabelEncoder encodes labels with a value between 0 and n\_classes-1 where n is the number of distinct labels. If a label repeats it assigns the same value to as assigned earlier.

There were duplicate records in the dataset which had the same values except for the ID column. The duplicate records were dropped so as to remain with a customer having just one record.

Checked for missing values in the record and dropped the occupation\_type column since it had many missing values that may not be easy to fill.

**Credit Record dataset**

Renamed the columns names and converted to lowercase.

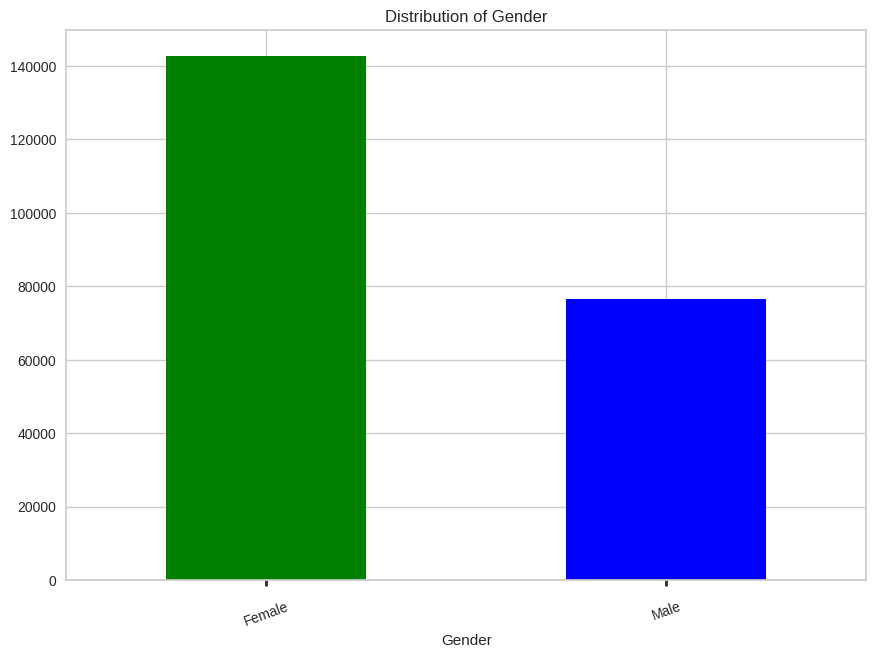
## DATA ANALYSIS

The Credit Record dataset and Application Record dataset were merged to make the analysis easier. They were merged using the unique identifier which was the ID column.

Checked all the categorical data to look at the distribution.

**Gender**

Out of the 219,173applicants, (65%, n=142,682) were female and (35%, n=76,491) were male.



**Owns car and/or property**

More than half of the applicants owned cars and property. From the total applicants, 136,870 had cars and 82,303 did not own cars. 144,422 applicants had property and 74,751 did not have property.

**Owns mobile, Work phone and Owns phone**

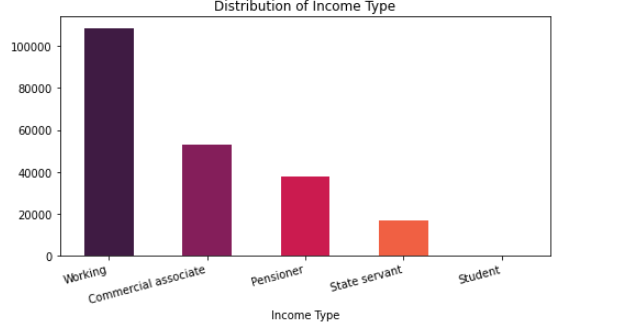
All applicants owned mobiles hence the column was dropped because the feature may not be of much help in analysis or modelling.

169,786 of all the applicants had work phones and 49,387 did not have work phones.

155,059 of the applicants owned phones while 64,114 did not own phones.

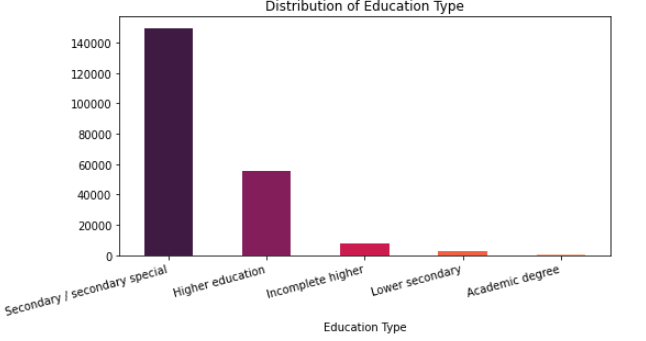
**Income type**

As shown in the diagram below, out of the total applicants, (50%, n=110,071) were working, (24%, n=53,265) were commercial associates, (18%, n=38,934) were pensioners, (7%, n=16,838) were state servants while (1%, n=65) were students.



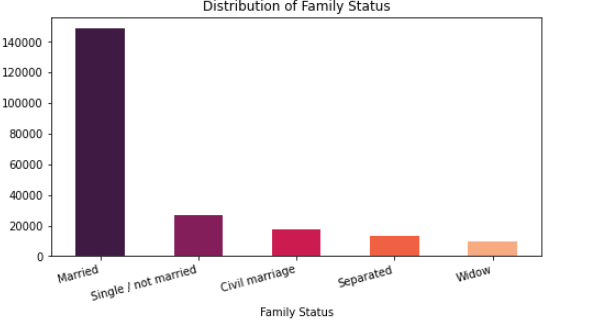
**Education type**

Out of the total applicants, (69%, n=152,062) had Secondary / secondary special education, (26%, n=56,113) had higher education, (3%, n=7,998) had incomplete higher education, (1%, n=2813) had lower secondary education and (1%, n=187) had academic degree.

****

**Family status**

Married respondents were (69%, n=150,186), (13%, n=28,209) were single/not married. The other (18%, n=40,778) were either from civil marriage, separated or widows.

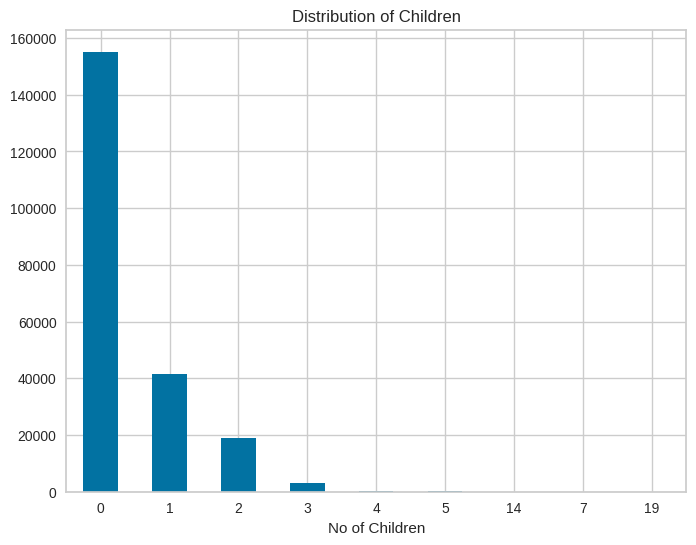


**Housing type**

The housing type was as follows;

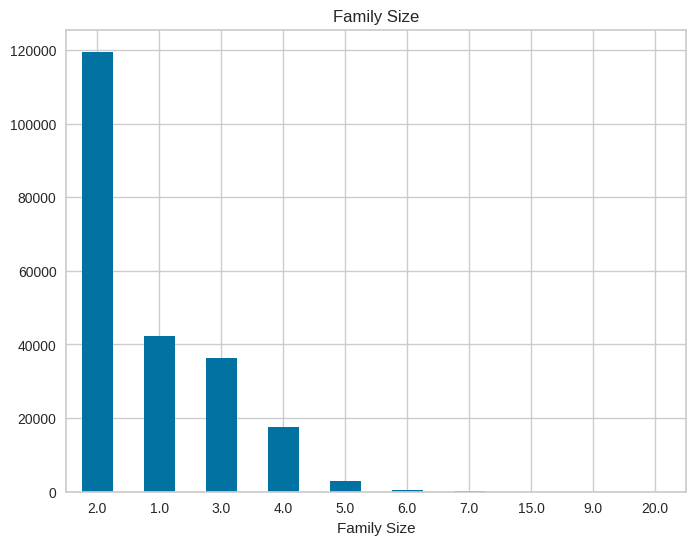
| **Housing type** | **Frequency** | **Percentage** |
| --- | --- | --- |
| House / apartment | 197,748 | 90.2% |
| With parents | 9,045 | 4.1% |
| Municipal apartment | 7,099 | 3.2% |
| Rented apartment | 2,729 | 0.8% |
| Office apartment | 857 | 0.4% |

**Number of Children**

Majority of the applicants had no children.

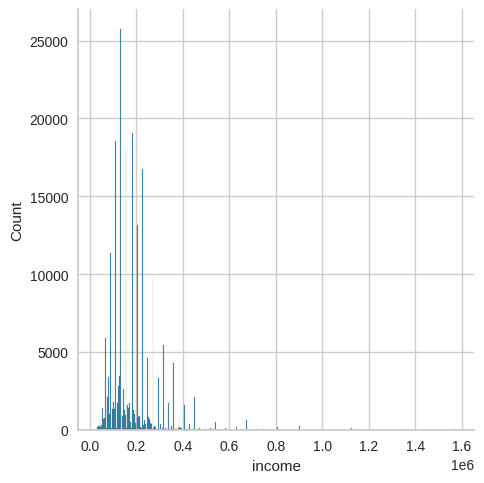
**Family size**

The applicants’ majority family size was two family members.



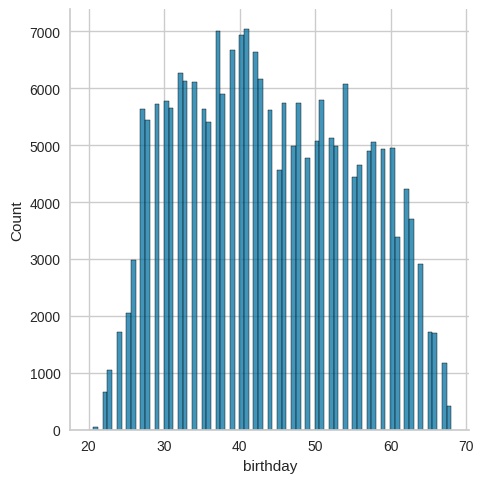
**Income Distribution**

The respondents’ income distribution is shown in the chart below.



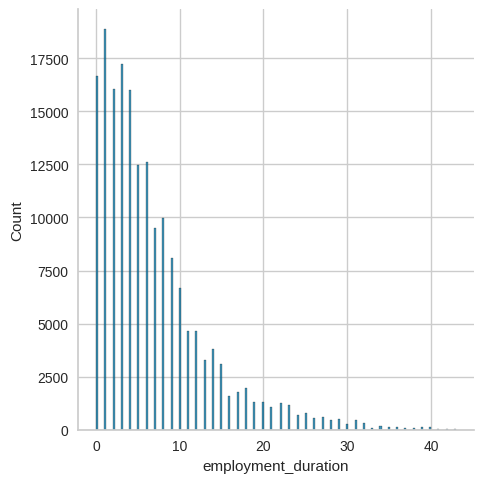
**Birthdays**

Their birthday distribution was as follows.



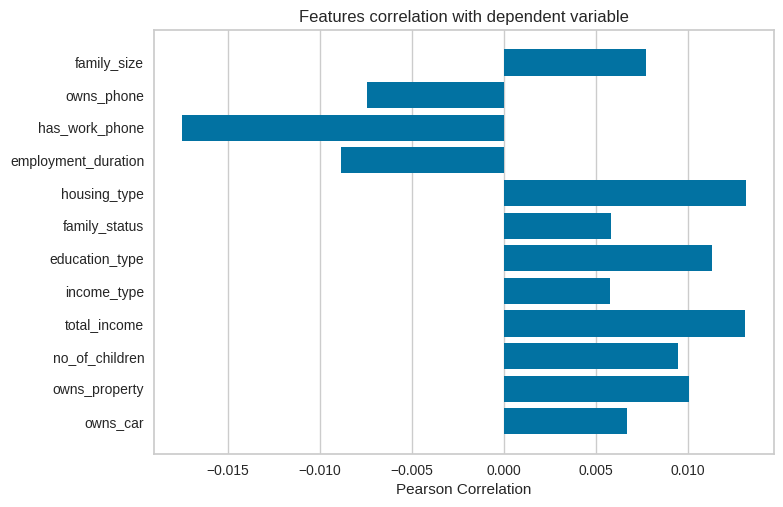
**Employment period**

The distribution is as follows.



## MACHINE LEARNING MODEL

**Feature Correlation with dependent variable**



The diagram is a representation of the variables in relation to the target or dependent variables. Before we make any type of algorithmic model, we have to select the best features for the model in order to get an accurate model. The highest correlation with the target variable (Status) is **education\_type** and **owns\_property** among many others. Statistically, pearson is the best formula derived to find correlation among variables, others that can be used are **Spearsman** and **Kendall** correlation formulas.

**Feature selection for our model**

The features used for the modelling were based on the following variables:

[ 'gender','no\_of\_children','owns\_car','owns\_property','education\_type','income\_type','total\_income','employment\_duration'

] which is represented by the diagram above.

**Training data and Testing data**

Training data size used: 75% (164,379)

Testing data size: 25% (54,794)

**Confusion Matrix for model testing.**

-**’0’** and **’1’** represent the target or dependent variable which is the status used to calculate whether or not one is a good or bad bank client.

**‘0’-** this represents people who are at a lower risk meaning they either paid their bank debts or loans or do not have any loans for the month.

**‘1’-** this represents the number of people who have loans that are overdue the time granted therefore having a higher risk.

**Model 1: Logistic Regression Predictions**

Predicted values: [0 0 0 ... 0 0 0]

Accuracy Score is **0.61932**

|  | 0 | 1 |
| --- | --- | --- |
| 0 | 33935 | 0 |
| 1 | 20859 | 0 |

**Decision Tree Classifier Predictions**

Predicted values: [0 0 0 ... 0 0 0]

Accuracy Score is **0.61932**

|  | 0 | 1 |
| --- | --- | --- |
| 0 | 33935 | 0 |
| 1 | 20859 | 0 |

**Random Forest Classifier Predictions**

Prediction values: [0 0 0 ... 0 0 0]

Accuracy Score is **0.67126**

|  | 0 | 1 |
| --- | --- | --- |
| 0 | 33379 | 556 |
| 1 | 17457 | 3402 |

**Best model is Random Forest Classifier** with:

Test score: **67.126 %**

**Description**

The test score shows how accurate the model is, the confusion matrix shows a summary of correct and incorrect predictions in a dataset and then breaks them down by each class. In relation to the classifier used, we see that the correct predictions made for the **‘0’** class is **33,379 a**nd for the **‘1’** class is **3,402.**  The rest of the predictions were wrongly classified, resulting in the other percentage error of the classifier.

## SUMMARY AND RECOMMENDATIONS

1. All the applicants had mobiles.
2. More than half of the applicants owned cars and properties.
3. 50% of the applicants were working.
4. The highest count level of education was secondary education.
5. Majority of the applicants had no children.
6. The best model to use for predictions is Random Forest Classifier.