MMU
MULTIMEDIA UNIVERSITY

TDS 3301
DATA MINING

PROJECT

QUESTION 3: Intelligent Decision-Making for Loan Application

Prepared by

Ong Shuoh Chwen, 1171102212, 018-2992362
Yong Wen Kai, 1171101664, 016-3963254

Contents

1 Exploratory Data Analysis (EDA)

The dataset used in this project is from Question 3: Intelligent Decision-Making for Loan Application, *'Banking_CS.csv'.* The dataset contained the basic personal data of the person applying for loan, ranging from employment type, loan amount, credit card type and etc.

The EDA process starting off by with filling in the missing values of the data, pre-processing the data, oversampling the data, and performing the EDA tasks to understand the data even more in-depth and better.

1.1 Missing Values Treatment

1.1.1 Checking for missing values

First and foremost, empty or missing values are required to be either get rid off or replace by assigning new values to it. In this section, we are going to use *isnull()* function to check if there's any missing or null values in the dataset.

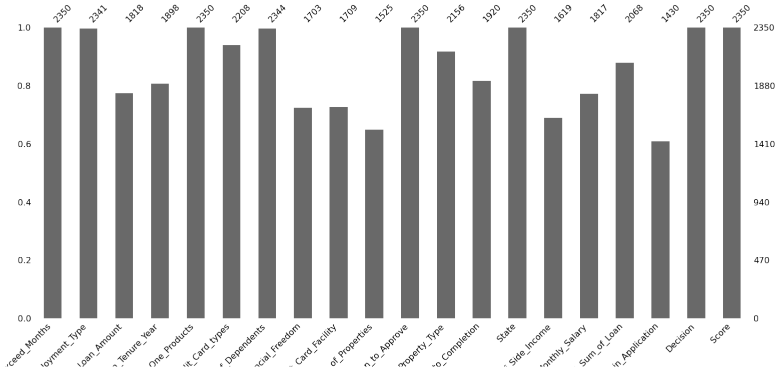


Figure 1: The results of *isnull()*

Based on Figure 1, we found out that the dataset contained a lot of missing values, such as Loan_Amount, Loan_Tenure_Year, Credit_Card_types, Years_to_Financial_Freedom, Card_Facility, Number_of_Properties, Property_Type, Years_for_Property_to_Completion, Number_of_Side_Income, Monthly_Salary, Total_Sum_of_Loan and Total_Income_for_Join_Application.

1.1.2 Dealing with missing values

From the missing values we found in the previous subsection, we are going to fill those missing values with mode values from each corresponding column. We felt that removing those missing values are going to impact hard on the upcoming data preprocessing, as some of the data are very limited. For instance, data from Kedah state are 1 out of 10 compared to Johor state which are almost 8 out of 10.

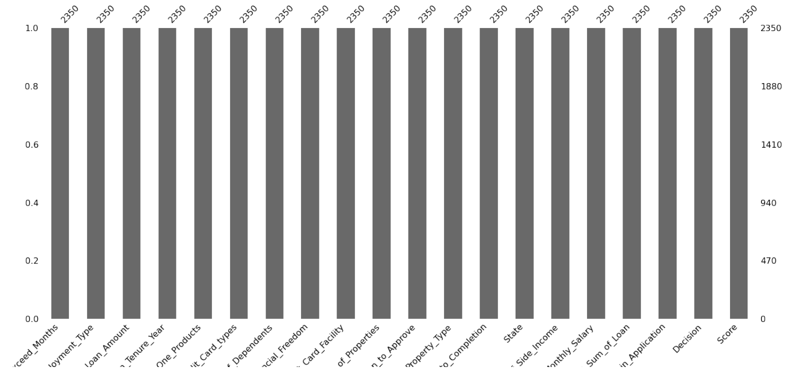


Figure 2: The results of *isnull()* after filling the mode values

Based on Figure 2, we had successfully filled in the mode values to replace the missing values. All of the columns in the dataset are now containing 2350 records.

1.2 Data Pre-processing

1.2.1 Noisy Data

We are going to apply regular expression (Regex) to the noisy data. Here are some of the example that we replace our data column in the dataset using Regex.

Table 1: Noisy Data

Before Regex	**After Regex**
P.Pinang	Penang
Johor B	Johor
K.L	Kuala Lumpur
N.Sembilan	Negeri Sembilan
SWK	Sarawak
Trengganu	Terrenganu
Self_Employed	Self Employed
government	Government

1.2.2 Categorizing Data

We are going to categorize some of the data from float64 type to categorical type. We implemented *.astype("category")* to categorize the data. Loan_Amount and Monthly_Salary are chosen to be categorized.

1.3 Synthetic Minority Over-sampling Technique (SMOTE)

1.3.1 Imbalance Data

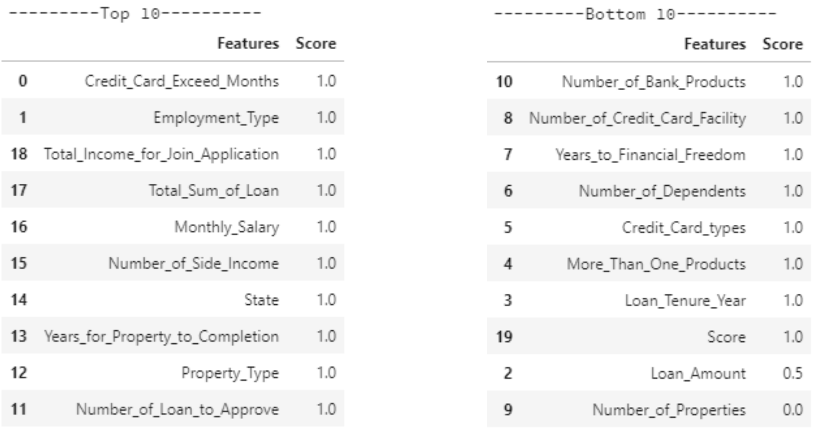
In this section, we are going to oversample the imbalance data using SMOTE. We chose oversampling method is because undersampling will reduce the size of the data, thus making the dataset smaller and not enough useful data to be take into account for prediction later.

First and foremost, we duplicated a new dataframe for SMOTE. In the new dataframe, 'Decision' is dropped and all the values are encoded using *LabelEncoder()*. X is the features dataframe which includes all columns except 'Decision' while y is the target variable which is referring to 'Decision'. After that, X and y are fitted into SMOTE.

2 Feature Selection

For feature selection to work, the original dataset has been transformed into an analytical dataset that is able to fit into the prediction model later on.

Feature selection algorithms used for the dataset are Boruta and Recursive Feature Elimination (RFE). Both of the algorithms are feature ranking and selection algorithms that based on Random Forest Classifier. Due to Random Forest Classifier, both of the feature selection algorithms are able to return good accuracy and result.



```
---------Top 10----------
```

	Features	Score
0	Credit_Card_Exceed_Months	1.0
1	Employment_Type	1.0
18	Total_Income_for_Join_Application	1.0
17	Total_Sum_of_Loan	1.0
16	Monthly_Salary	1.0
15	Number_of_Side_Income	1.0
14	State	1.0
13	Years_for_Property_to_Completion	1.0
12	Property_Type	1.0
11	Number_of_Loan_to_Approve	1.0

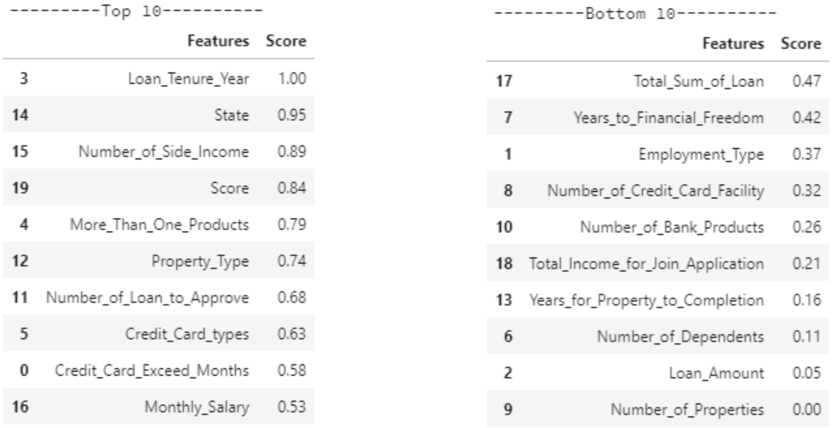
(a) Boruta Top 10 Features

```
---------Bottom 10----------
```

	Features	Score
10	Number_of_Bank_Products	1.0
8	Number_of_Credit_Card_Facility	1.0
7	Years_to_Financial_Freedom	1.0
6	Number_of_Dependents	1.0
5	Credit_Card_types	1.0
4	More_Than_One_Products	1.0
3	Loan_Tenure_Year	1.0
19	Score	1.0
2	Loan_Amount	0.5
9	Number_of_Properties	0.0

(b) Boruta Bottom 10 Features

Figure 3: Results of Boruta Features Selection



---------Top 10----------

	Features	Score
3	Loan_Tenure_Year	1.00
14	State	0.95
15	Number_of_Side_Income	0.89
19	Score	0.84
4	More_Than_One_Products	0.79
12	Property_Type	0.74
11	Number_of_Loan_to_Approve	0.68
5	Credit_Card_types	0.63
0	Credit_Card_Exceed_Months	0.58
16	Monthly_Salary	0.53

(a) RFE Top 10 Features

---------Bottom 10----------

	Features	Score
17	Total_Sum_of_Loan	0.47
7	Years_to_Financial_Freedom	0.42
1	Employment_Type	0.37
8	Number_of_Credit_Card_Facility	0.32
10	Number_of_Bank_Products	0.26
18	Total_Income_for_Join_Application	0.21
13	Years_for_Property_to_Completion	0.16
6	Number_of_Dependents	0.11
2	Loan_Amount	0.05
9	Number_of_Properties	0.00

(b) RFE Bottom 10 Features

Figure 4: Results of RFE Features Selection

Based on both of the feature selection results of Boruta and RFE, we found out that LoanAmount and NumberofProperty are ranked last and having the least amount of score. We dropped both of the features and proceed into the next step.

3 Data Mining Techniques

3.1 Association Rule Mining

Association rule mining is performed to extract unique rules of each individual who applies for the loan, based on the current loan duration, credit card information and amount of property the applicant have. This technique is used to investigate the combination of past loan records of the applicant. It observes the loan records that are frequently accustomed together by the applicant and calculates their strength of relations. The algorithm is suitable for this purpose because the dataset is mainly in non-numeric data. With this technique, the antecedents and consequences can be clearly identified, alongside with their scores of strength, which is measured by lift. A lift score with 1.0 and above means the combination of loan records is happening often.

3.2 Classification Model

Multiple classification models are implemented to compare their performances, which are Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). These models are evaluated with multiple metrics, such as accuracy, precision, recall and F1-score. However, the metric we are going to focus on will be F1-score, as accuracy has the issue of accuracy paradox, where a model that predicts everything to be the major class will have a high accuracy. On the other hand, precision and recall only takes into consideration two classifier labels. We came to a conclusion that F1-score will be used for comparison, while other metrics can still be used as reference for more detailed breakdown of the performance of the models.

3.3 Clustering

The dataset after the feature selection remained only categorical data, so that the clustering technique used is K-modes clustering. K-modes clustering defines clusters based on the number of matching categories between the data points, which in contrast to the K-means clustering technique that clusters the data based on a mean Euclidean distance. The Elbow Method is used in order to determine the optimal number of clusters, K. Figure **??** is plotted to compare the cost against each K. The cost is the sum of dissimilarities of all data points with their closest centroids. Based on Figure **??**, we can see that the elbow point is when K=3. Hence, the number of clusters that we used is 3.

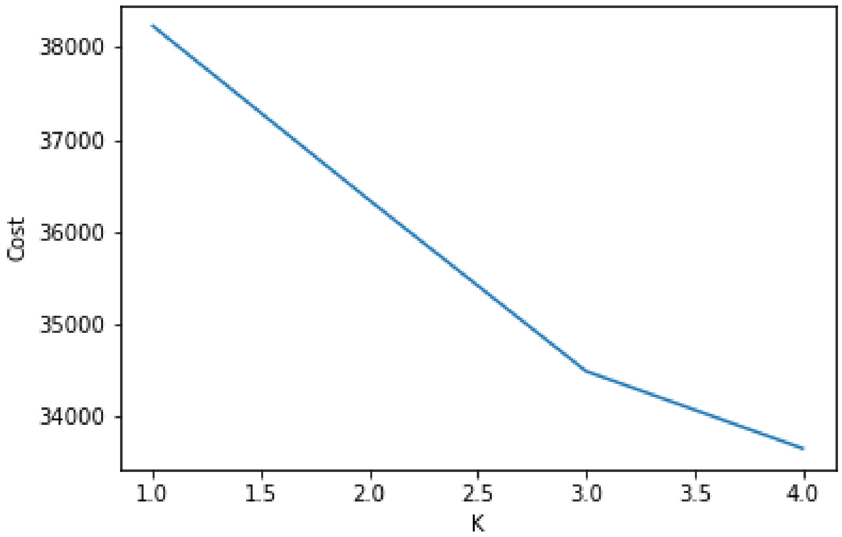


Figure 5: Finding the Elbow for K-Modes Clustering

4 Results & Discussion

4.1 Exploratory Data Analysis

4.1.1 Exploratory Data Analysis Before SMOTE

In this section, we will show the Exploratory Data Analysis before applying SMOTE on the dataset.

Question: Which type of employment has the highest chance of getting the loan accepted?

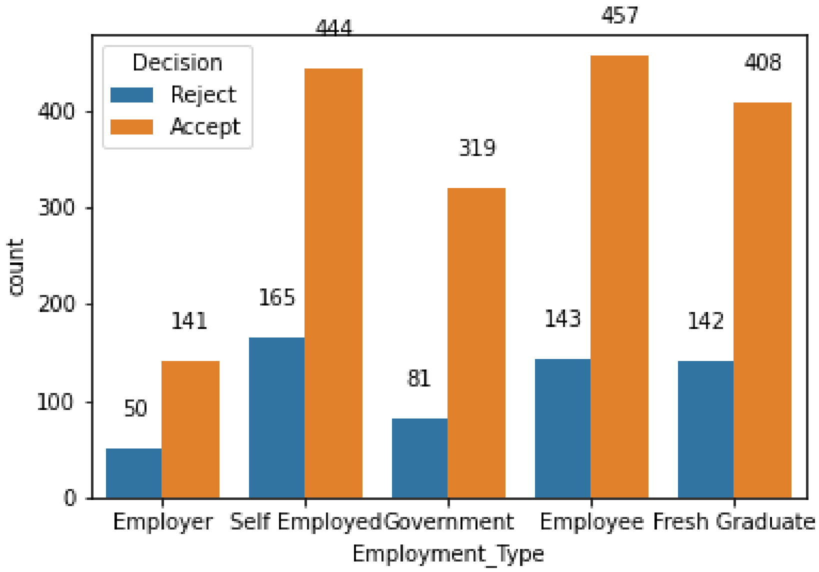


Figure 6: Employment Type before SMOTE

Figure above shows that Self Employed people have the highest count of loan application. In our opinion, this is most likely that self employed people will need the loan to start up a business.

Question: Which credit card type is the most?

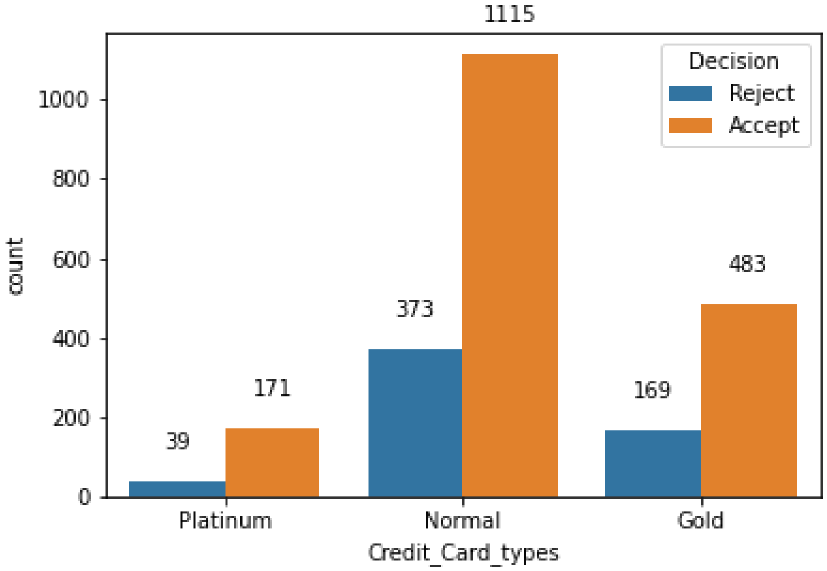


Figure 7: Credit card Type before SMOTE

Figure above shows that normal type of credit card user have the most loan application among platinum and gold user. In our opinion, this is most likely that normal type of credit card user will have the least amount of income compared to gold and platinum users.

Question: Which type of property type has the highest chance of getting the loan accepted?

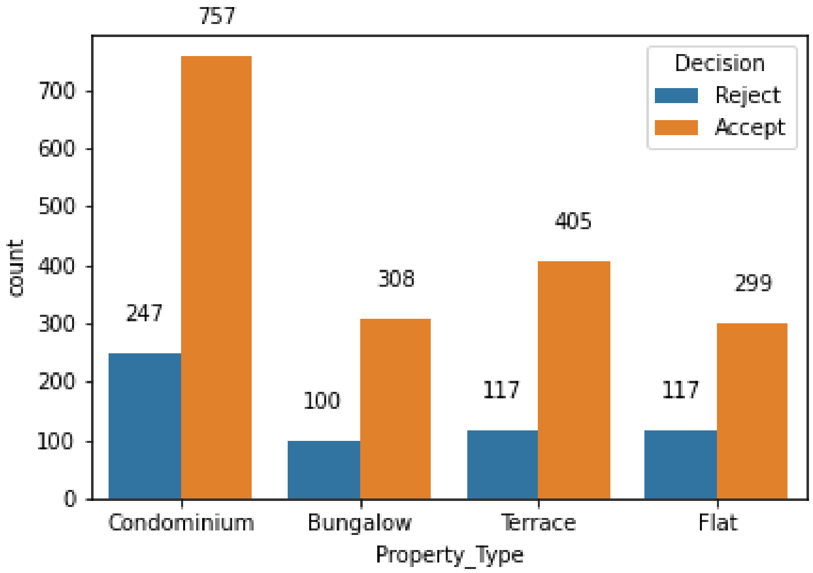


Figure 8: Property Type before SMOTE

Figure above shows that condominium type of property have the most loan application among the the other types. In our opinion, this is because condominium have the most average rate of income compared to other property types. For example, people who are able to afford condominium will have a more reasonable purpose to apply a loan compared to people who are staying in bungalow.

Question: Which category of monthly salary has the highest chance of getting the loan accepted?

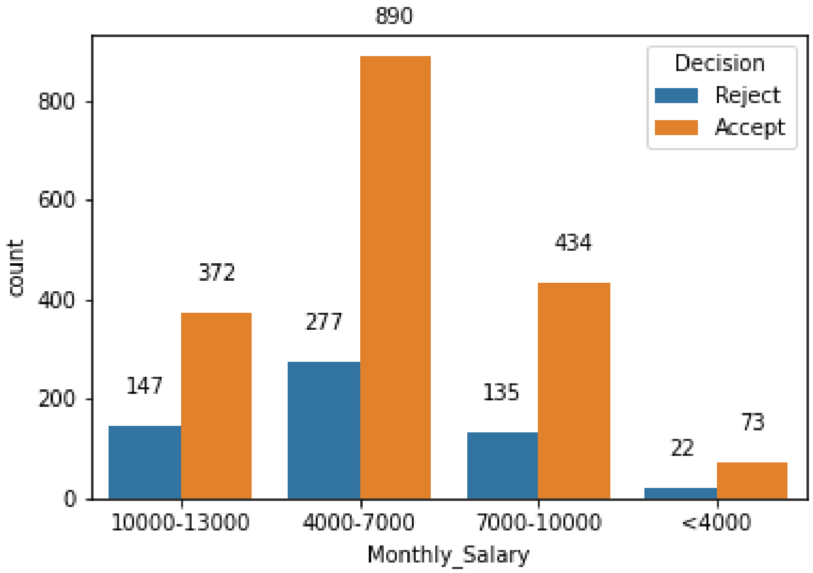


Figure 9: Monthly salary before SMOTE

Figure above shows that monthly salary category of RM4000 - RM7000 have the most loan application. In our opinion, this is because that the loan will be more helpful compared to the other categories. In the other hand, if compared to the category of less than RM4000, they will be much more helpful compared to category RM4000 - RM7000. But due to their low amount of income, they cannot guarantee that they are able to pay back the loan.

Question: Which type of employment has the highest chance of getting the loan accepted?

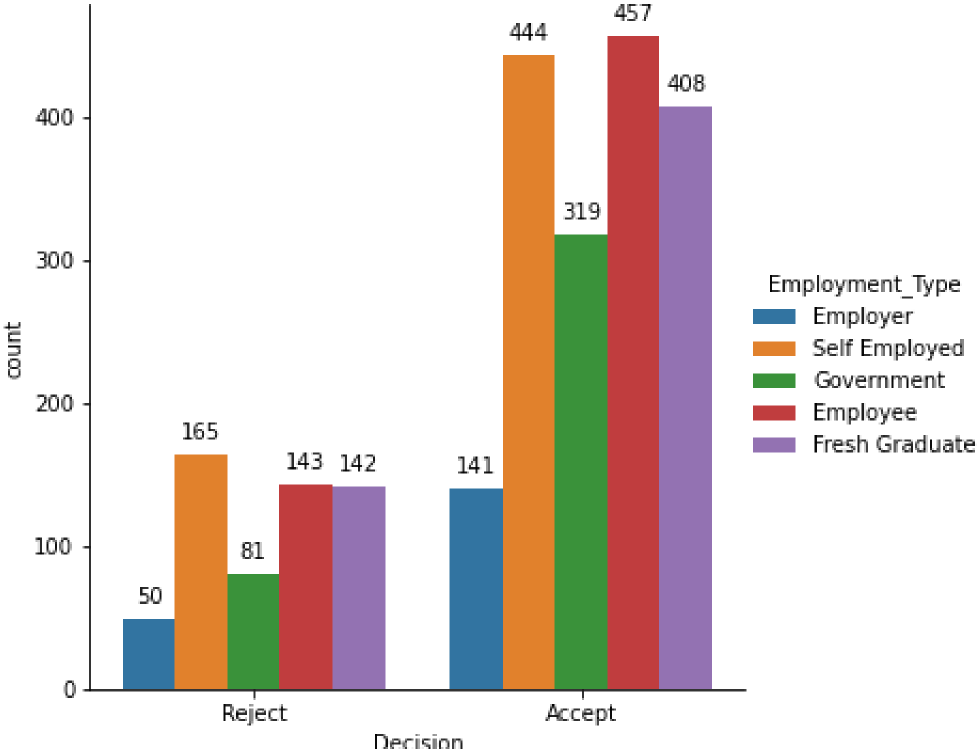


Figure 10: Bank decision on employment type before SMOTE

Figure above shows that Employee have the most acceptance counts compared to the other employment type. In our opinion, this is because that employee have a more stable income compared to the other employment type. For example, self employed people will have a higher risk of losing their own business.

Question: Which type of credit card has the highest chance of getting the loan accepted?

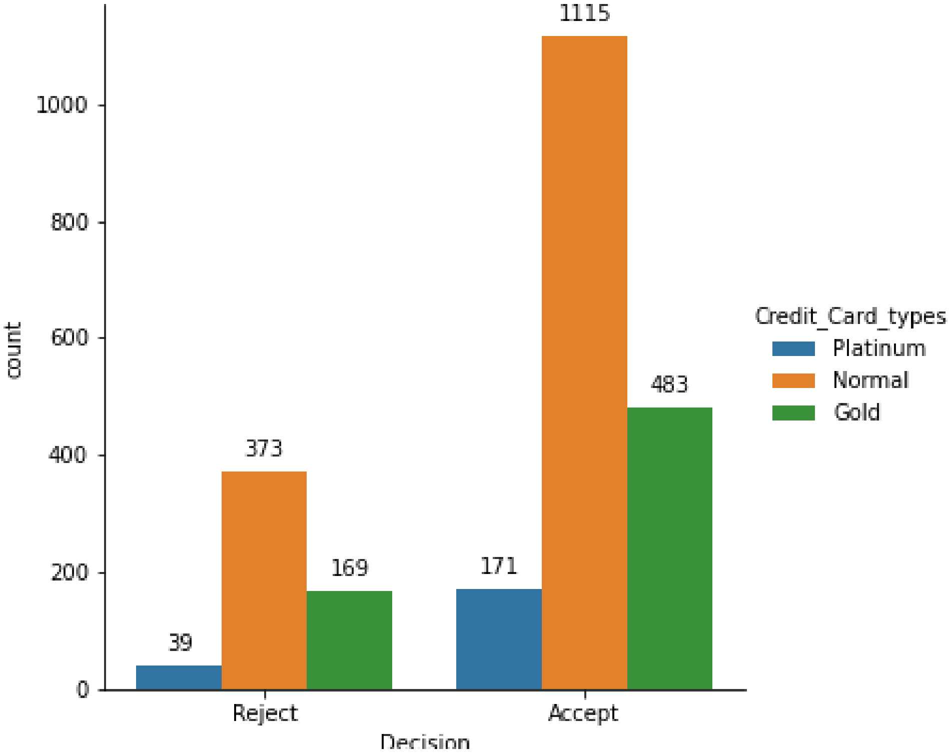


Figure 11: Bank decision on credit card type before SMOTE

Figure above shows that normal type of credit card users have the highest acceptance count compared to the other types. In our opinion, this is because that accepting normal credit card user will be more helpful for them them compared to the other types of credit card users.

Question: Which type of property has the highest chance of getting the loan accepted?

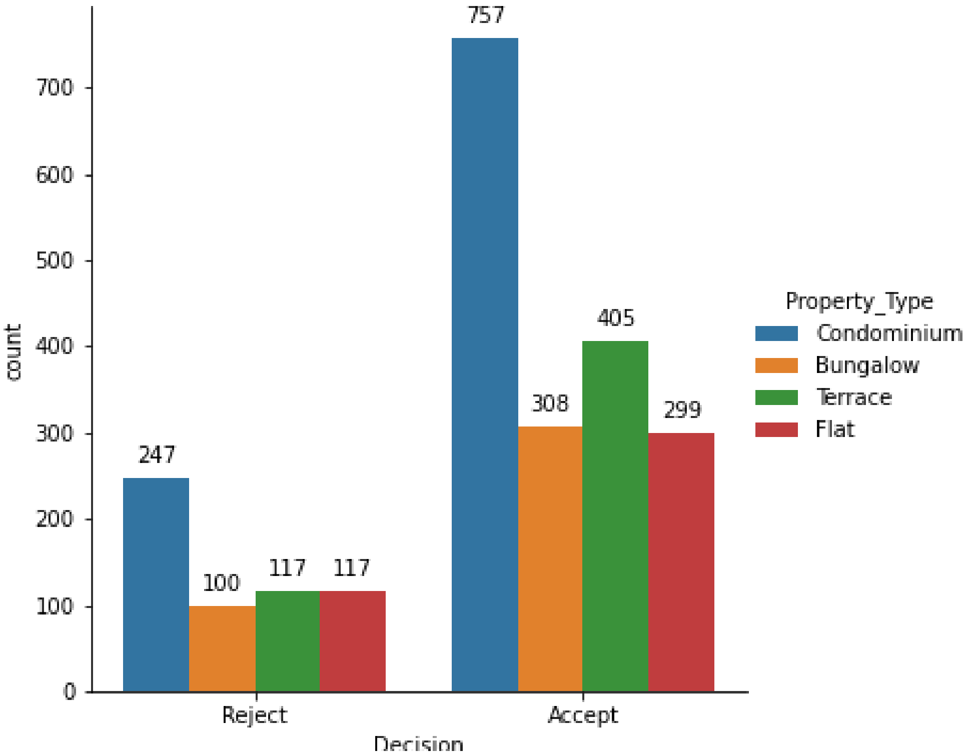


Figure 12: Bank decision on property type before SMOTE

Figure above shows that condominium property type have the highest acceptance count compared to the other property type. In our opinion, this is because that people who are able to afford condominium will have an average and stable income compared to those who are staying in terrace and flat. In the other hand, people who are living in bungalow are probably those who does not need loan and people who are living in terrace and flat have the risk of not able to pay back the loan on time.

Question: Which category of monthly salary has the highest chance of getting the loan accepted?

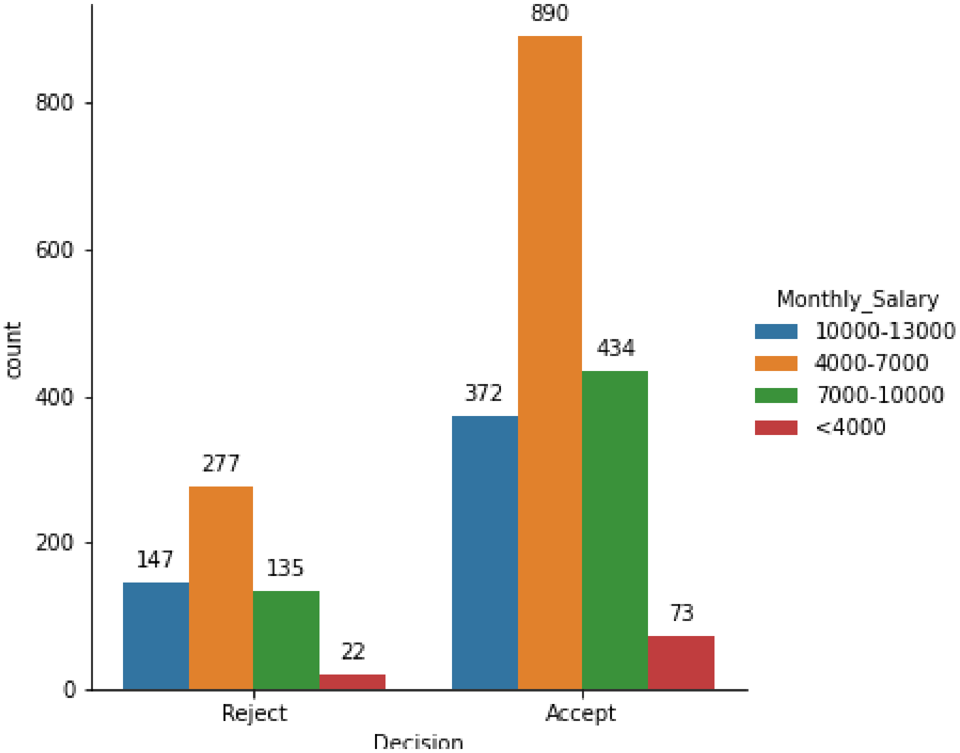


Figure 13: Bank decision on monthly salary category before SMOTE

Figure above shows that monthly salary category of RM4000 - RM7000 have the most acceptance rate compared to the other categories. In our opinion, this is because that accepting those who are in the category of RM4000 - RM7000 will be more helpful compared to the other categories. In the other hand, for the bank to accept those who are in the category of lesser than RM4000 will have the risk of them could not payback the loan because of their low monthly salary.

Question: Which decision is made most?

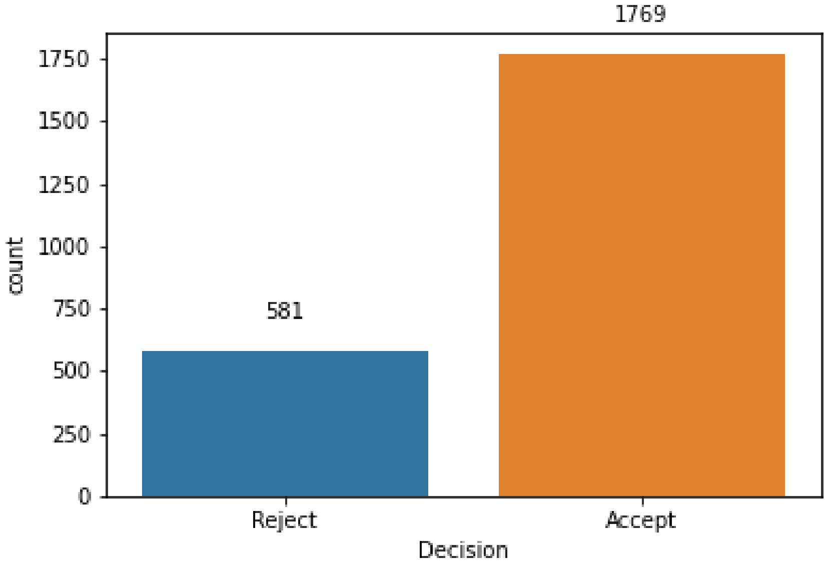


Figure 14: Bank decision before SMOTE

Figure above shows that the bank accept most loan application. Based on the dataset and the data we visualized, we found out that the in each of the features, such as employment type, monthly salary, credit card types and property type. All of these features are being treated as the factors for loan acceptance. Thus, resulting in the bank having higher acceptance rate compared to rejection rate in terms of decision-making in loan application.

4.1.2 Exploratory Data Analysis After SMOTE

In this section, we will show the Exploratory Data Analysis after applying SMOTE on the dataset.
Question: Which type of employment has the highest chance of getting the loan accepted?

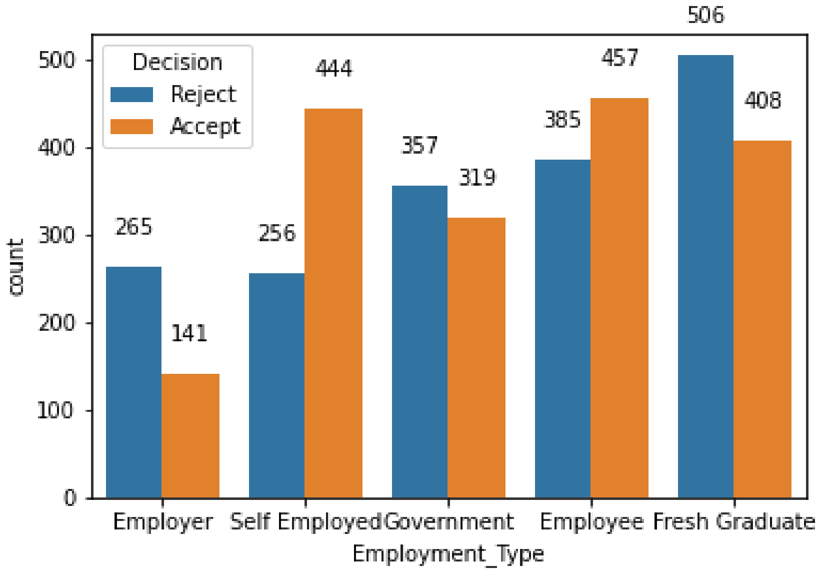


Figure 15: Employment Type after SMOTE

Figure above shows the employment type after applying SMOTE. Comparing with figure 3, we are able to observe that the number of fresh graduates had increased and became the most among the other employment type.

Question: Which type of credit card has the highest chance of getting the loan accepted?

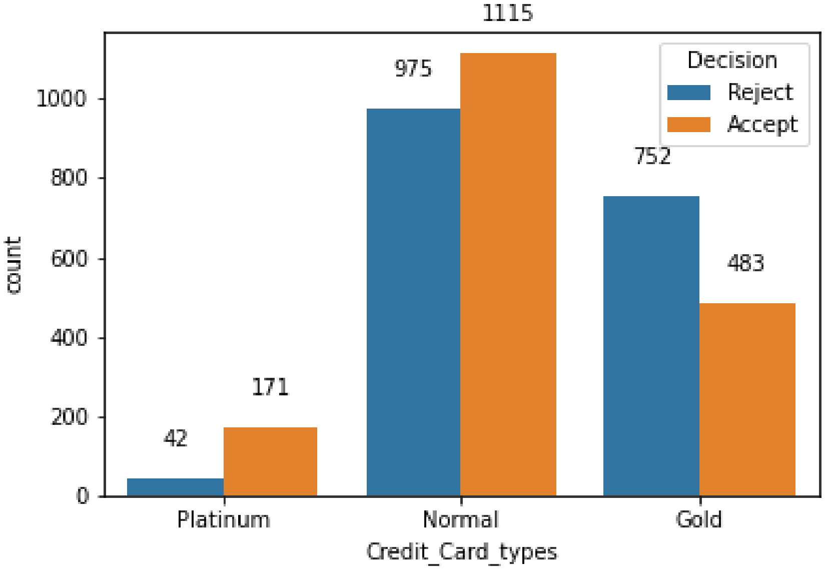


Figure 16: Credit card Type after SMOTE

Figure above shows the credit card type after applying SMOTE. Comparing with figure 4, we are able to observe that there is a light increment on normal and gold users.

Question: Which type of property has the highest chance of getting the loan accepted?

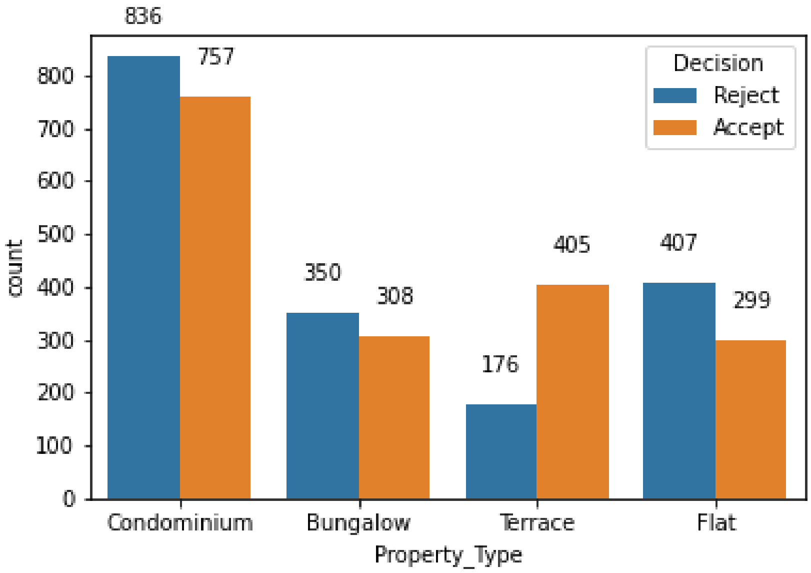


Figure 17: Property Type after SMOTE

Figure above shows the property type after applying SMOTE, Comparing with figure 5, we can observe that the number of terrace had increased by a little.

Question: Which category of monthly salary has the highest chance of getting the loan accepted?

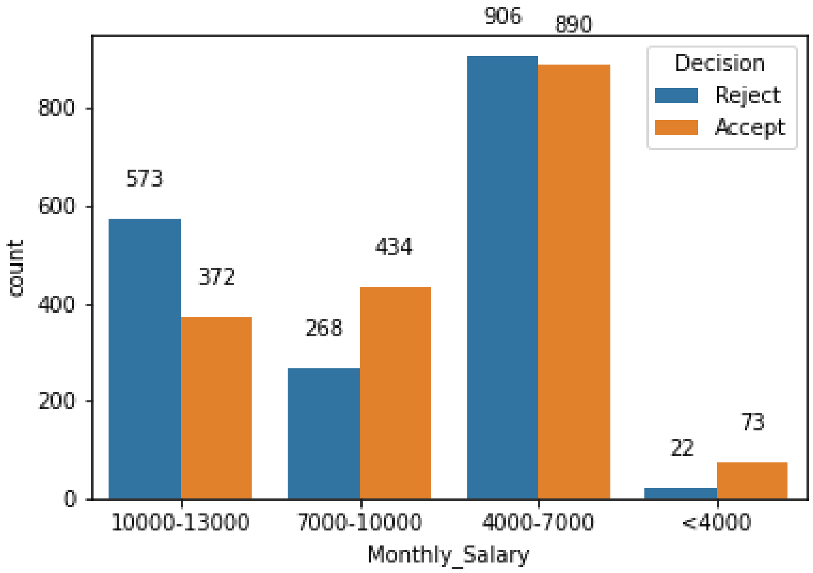


Figure 18: Monthly salary after SMOTE

Figure above shows the property type after applying SMOTE. Comparing with figure 6, we are able to observe that there is an increment at the category RM4000 - RM7000 and RM10000 - RM13000.

Question: Which type of employment has the highest chance of getting the loan accepted?

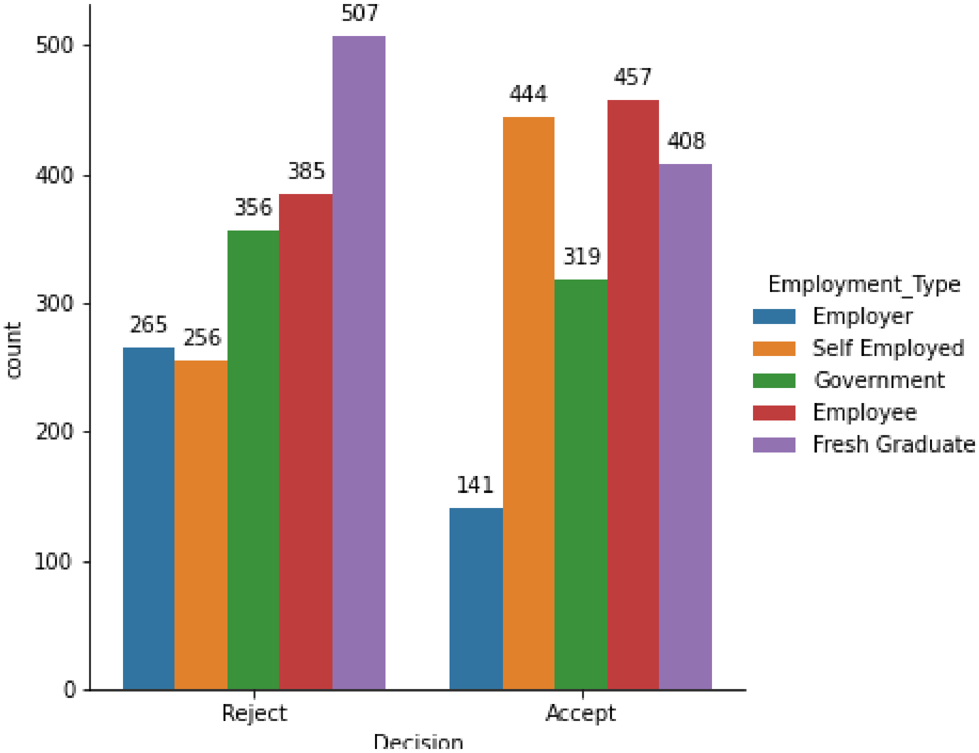


Figure 19: Bank decision on employment type after SMOTE

Figure above shows the loan acceptance and rejection by the employment type. Comparing to figure 7, we can observe that the number of rejected loan applications had greatly decreased.

Question: Which type of credit card has the highest chance of getting the loan accepted?

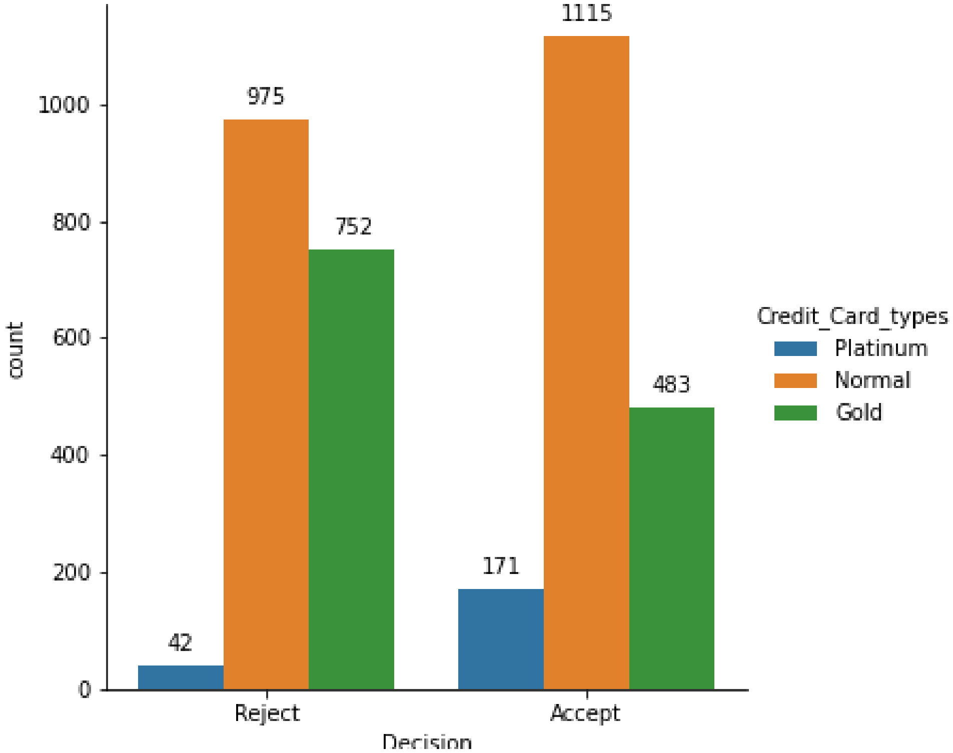


Figure 20: Bank decision on credit card type after SMOTE

Figure above shows the loan acceptance and rejection by credit card types. Comparing to figure 8, we are able to observe that the number of rejected loan applications had greatly decreased.

Question: Which type of property has the highest chance of getting the loan accepted?

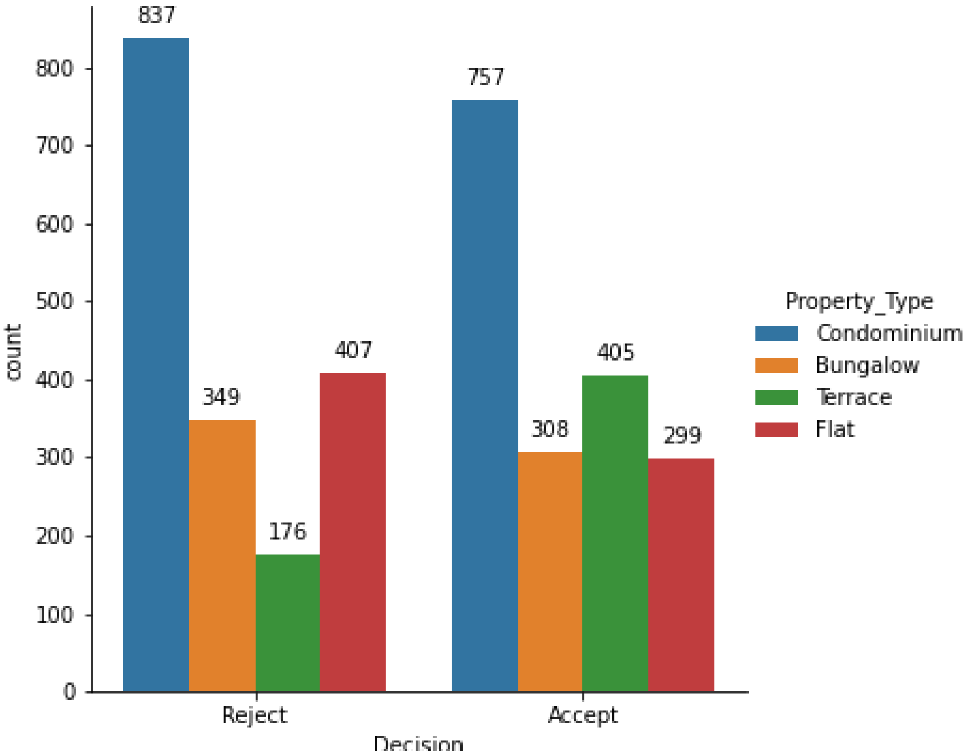


Figure 21: Bank decision on property type after SMOTE

Figure above shows the loan acceptance and rejection by property types. Comparing to figure 9, we can observe that the number of rejected loan application on condominium had greatly decreased.

Question: Which category of monthly salary has the highest chance of getting the loan accepted?

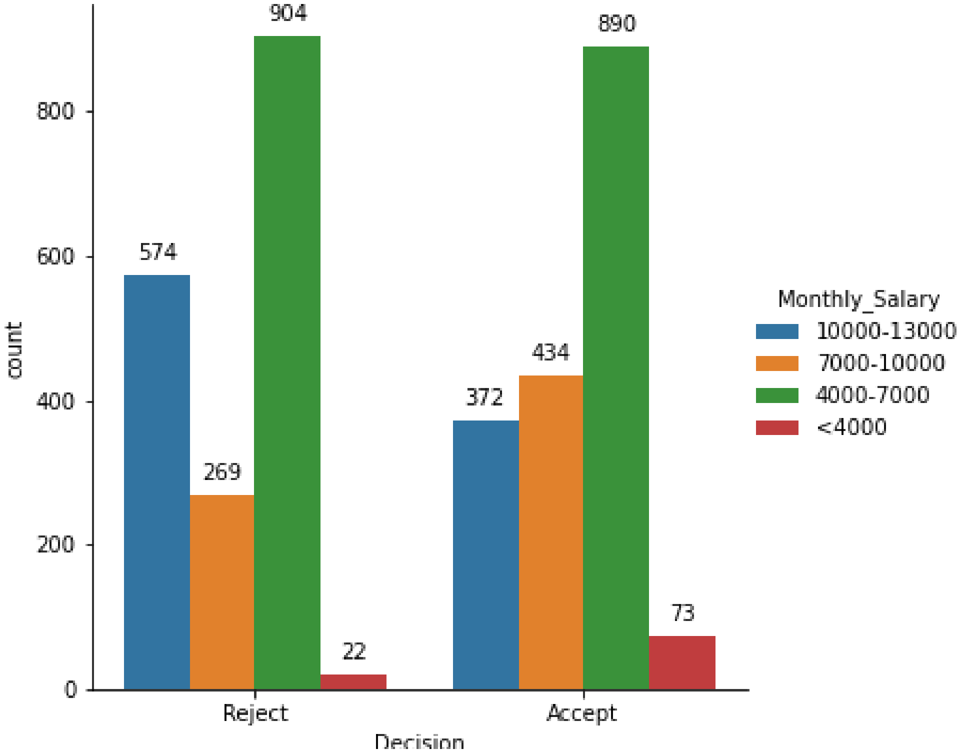


Figure 22: Bank decision on monthly salary category after SMOTE

Figure above shows the loan acceptance and rejection by monthly salary. Comparing to figure 10, we are able to observe that the number of rejected applications on category RM4000 - RM7000 had greatky decreased.

Question: Which decision is made most?

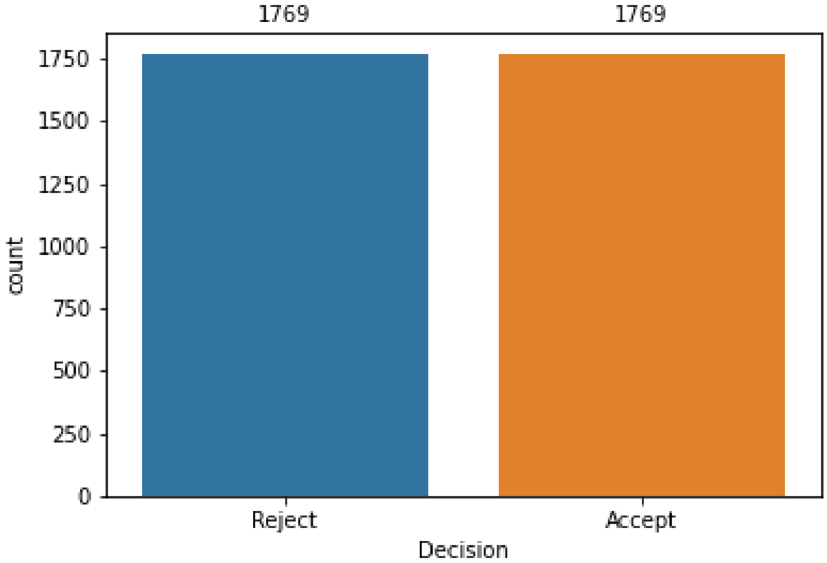
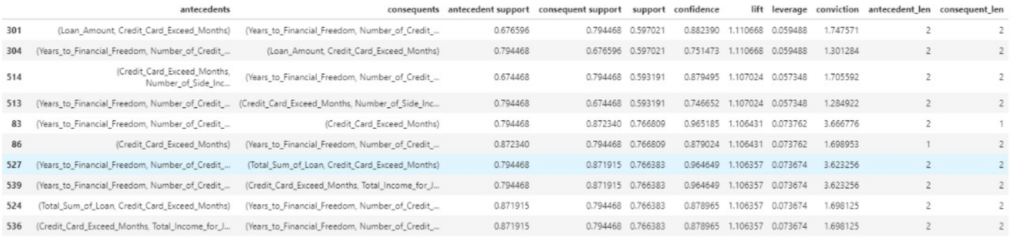


Figure 23: Bank decision after SMOTE

Figure above shows the total number of loan acceptance and rejection by the bank. Comparing to figure 11, we can observe that the number is balanced out on both reject and accept.

4.2 Association Rule Mining



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	antecedent_len	consequent_len
301	(Loan_Amount, Credit_Card_Exceed_Months)	(Years_to_Financial_Freedom, Number_of_Credit_...	0.676596	0.794468	0.597021	0.882390	1.110668	0.059488	1.747571	2	2
304	(Years_to_Financial_Freedom, Number_of_Credit_...	(Loan_Amount, Credit_Card_Exceed_Months)	0.794468	0.676596	0.597021	0.751473	1.110668	0.059488	1.301284	2	2
514	(Credit_Card_Exceed_Months, Number_of_Side_Inc...	(Years_to_Financial_Freedom, Number_of_Credit_...	0.674468	0.794468	0.593191	0.879495	1.107024	0.057348	1.705592	2	2
513	(Years_to_Financial_Freedom, Number_of_Credit_...	(Credit_Card_Exceed_Months, Number_of_Side_Inc...	0.794468	0.674468	0.593191	0.746652	1.107024	0.057348	1.284922	2	2
83	(Years_to_Financial_Freedom, Number_of_Credit_...	(Credit_Card_Exceed_Months)	0.794468	0.872340	0.766809	0.965185	1.106431	0.073762	3.666776	2	1
86	(Credit_Card_Exceed_Months)	(Years_to_Financial_Freedom, Number_of_Credit_...	0.872340	0.794468	0.766809	0.879024	1.106431	0.073762	1.698953	1	2
527	(Years_to_Financial_Freedom, Number_of_Credit_...	(Total_Sum_of_Loan, Credit_Card_Exceed_Months)	0.794468	0.871915	0.766383	0.964649	1.106357	0.073674	3.623256	2	2
539	(Years_to_Financial_Freedom, Number_of_Credit_...	(Credit_Card_Exceed_Months, Total_Income_for_J...	0.794468	0.871915	0.766383	0.964649	1.106357	0.073674	3.623256	2	2
524	(Total_Sum_of_Loan, Credit_Card_Exceed_Months)	(Years_to_Financial_Freedom, Number_of_Credit_...	0.871915	0.794468	0.766383	0.878965	1.106357	0.073674	1.698125	2	2
536	(Credit_Card_Exceed_Months, Total_Income_for_J...	(Years_to_Financial_Freedom, Number_of_Credit_...	0.871915	0.794468	0.766383	0.878965	1.106357	0.073674	1.698125	2	2

Figure 24: Antecedents and Consequences of Association Rules Mining

The association rule mining algorithm can be adjusted by setting different minimum thresholds for support, confidence and lift as well as setting the numbers of antecedents and consequences. Based on Figure 24, the minimum support and confidence is set at 0.5, with a lift score at least 1.0, and at most 2 antecedents and consequences. The rules are also sorted by lift score. Based on the results in Figure 24, we found out that most of the applicant which are experiencing financial freedom in recent years are tend to have different credit card facilities. This is probably due to those applicant had stable income, so that they can apply different credit cards.

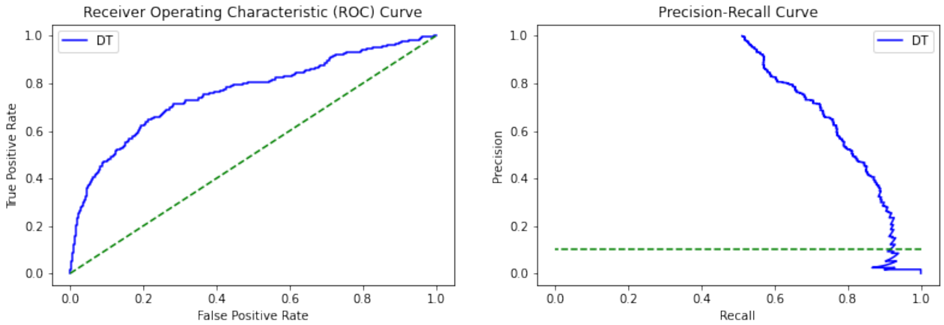
4.3 Classification Models

In this section, we are going to explore each classification model in detail. We will perform evaluation scores on each classification model and seek for the best F1-score to be used for our prediction model.

4.3.1 Decision Tree (DT)

Table 2: Evaluation Scores of the Decision Tree model

	Accuracy	**Precision**	**Recall**	**F1-Score**
DT	0.684557	0.683186	0.712177	0.697380



(a) ROC Curve

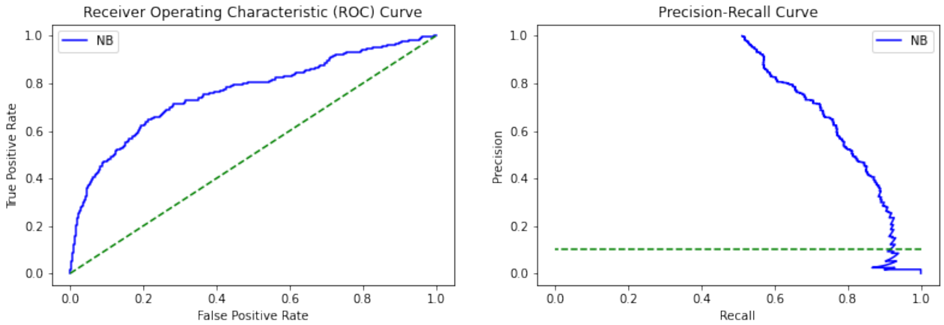
(b) Precision-Recall Curve

Figure 25: Curve Scores of Decision Tree

4.3.2 Naive Bayes (NB)

Table 3: Evaluation Scores of the Naive Bayes model

	Accuracy	**Precision**	**Recall**	**F1-Score**
NB	0.713747	0.722846	0.712177	0.717472



(a) ROC Curve

(b) Precision-Recall Curve

Figure 26: Curve Scores of Naive Bayes

4.3.3 K-Nearest Neighbor (KNN)

Table 4: Evaluation Scores of the K-Nearest Neighbor model

KNN	Accuracy	Precision	Recall	F1-Score
K=3	0.661016	0.646774	0.739852	0.690189
K=6	0.621468	0.643442	0.579335	0.609708
K=9	0.612052	0.610544	0.662361	0.635398

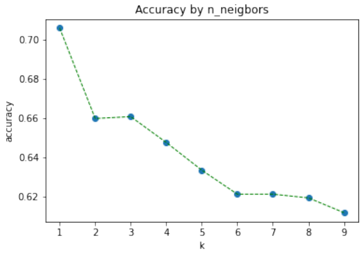
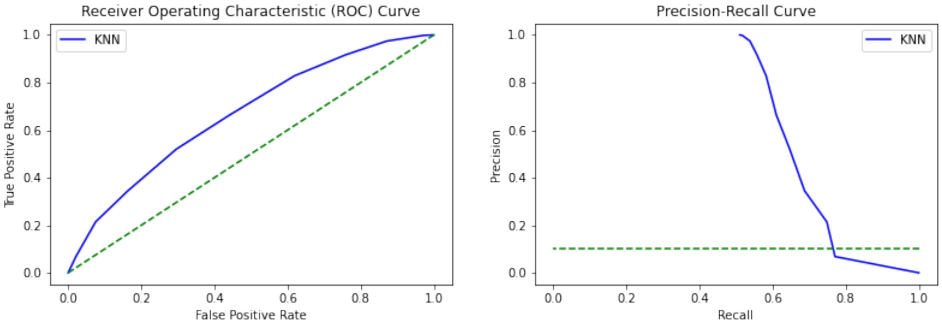


Figure 27: Accuracy by n-neighbors



(a) ROC Curve

(b) Precision-Recall Curve

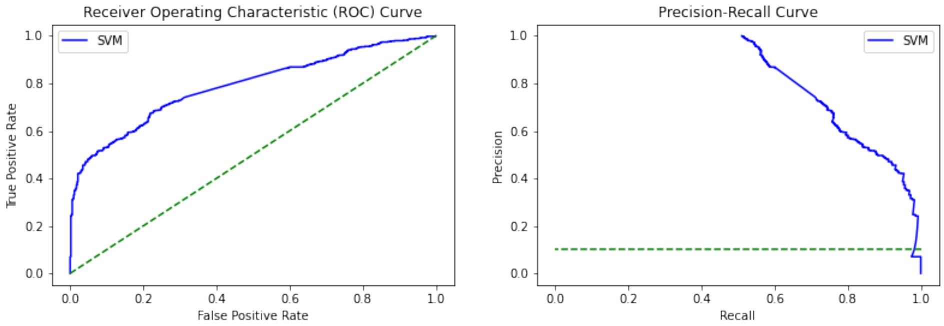
Figure 28: Curve Scores of K-Nearest Neighbor

4.3.4 Support Vector Machine (SVM)

Table 5: Evaluation Scores of the Support Vector Machine model

SVM Kernel	Accuracy	Precision	Recall	F1-Score
Rbf	0.694915	0.932539	0.433579	0.592686
Linear	0.671374	0.689587	0.647601	0.667935

Due to hardware and system constraint, we are unable to perform Linear and Poly kernel for Support Vector Machine. For Linear kernel, we are able to retrieve the evaluation score after 2 hours. As for Poly kernel, we failed to retrieve any evaluation score even after one day. Due to the execution time is high, we are utilizing Rbf kernel as our representation for Support Vector Machine.



(a) ROC Curve

(b) Precision-Recall Curve

Figure 29: Curve Scores of Support Vector Machine

4.3.5 Classification Evaluation Model

Table 6: F1-Score of the classification models

Classification Model	F1-Score
Decision Tree	0.697380
Naive Bayes	0.717472
K-Nearest Neighbor	0.690189
Support Vector Machine	0.592686

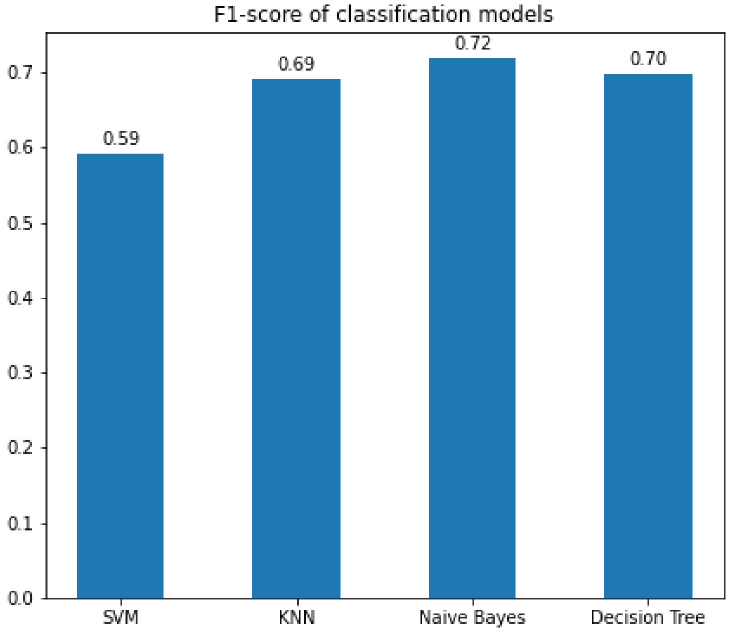
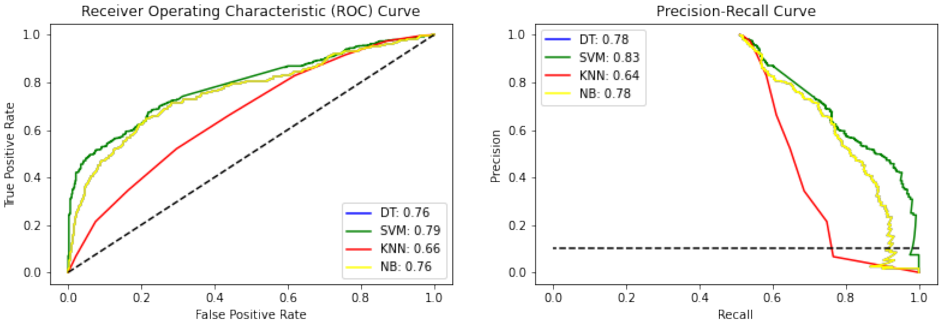


Figure 30: F1-Score of Classification Models

Based on Figure 30, we found out that Naive Bayes have the highest score in terms of F1-Score. So, we implemented Naive Bayes as the machine learning model for our prediction model later on.

Table 7: Table of Comparison between all AUC and Precision-Recall Curves

Classification Model	**AUC**	**Precision-Recall Curve**
Decision Tree	0.76	0.78
Naive Bayes	0.76	0.78
K-Nearest Neighbor	0.66	0.64
Support Vector Machine	0.79	0.83



(a) ROC Curve

(b) Precision-Recall Curve

Figure 31: Curve Scores of all Classification Models

4.4 Clustering

4.4.1 K-Modes Clustering

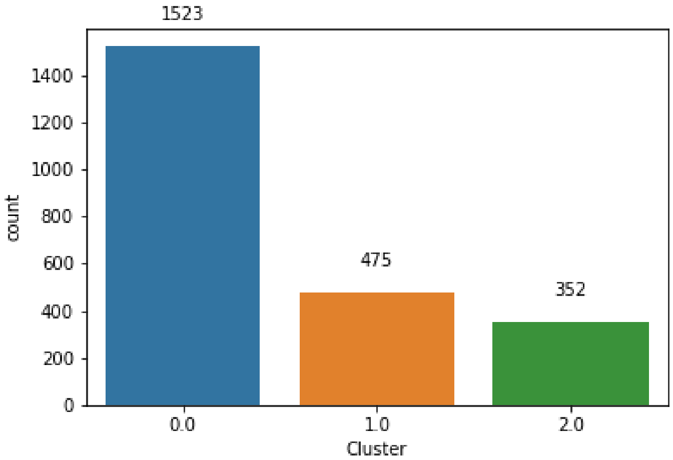


Figure 32: Cluster after performing K-Modes Clustering

Based on Figure 32, we can then visualize more in depth into the features we will be using in the prediction model. Cluster 0 occupied the most proportions compared to Cluster 1 and Cluster 2. All of the Cluster 0s have a larger proportion compared to their respective Cluster 1 and Cluster 2.

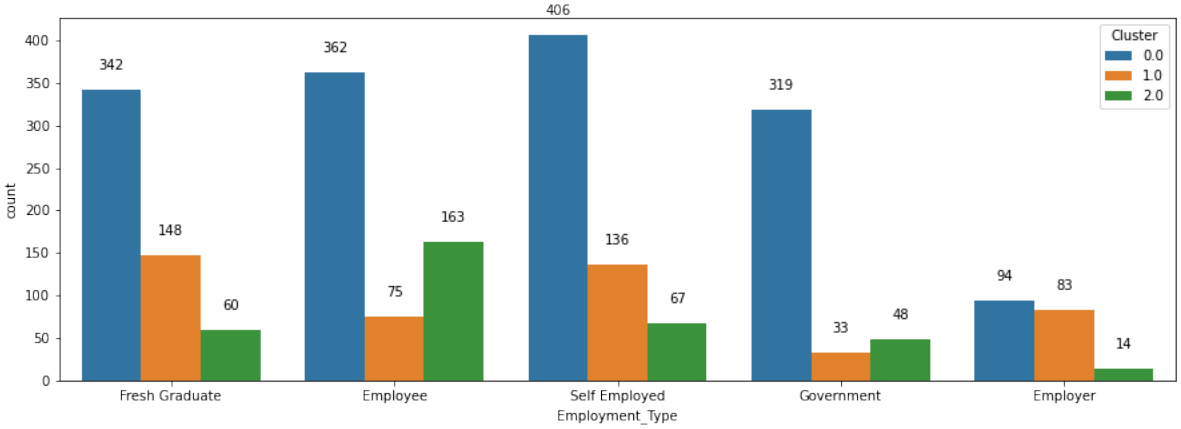


Figure 33: K-Modes Clustering of Employment Type

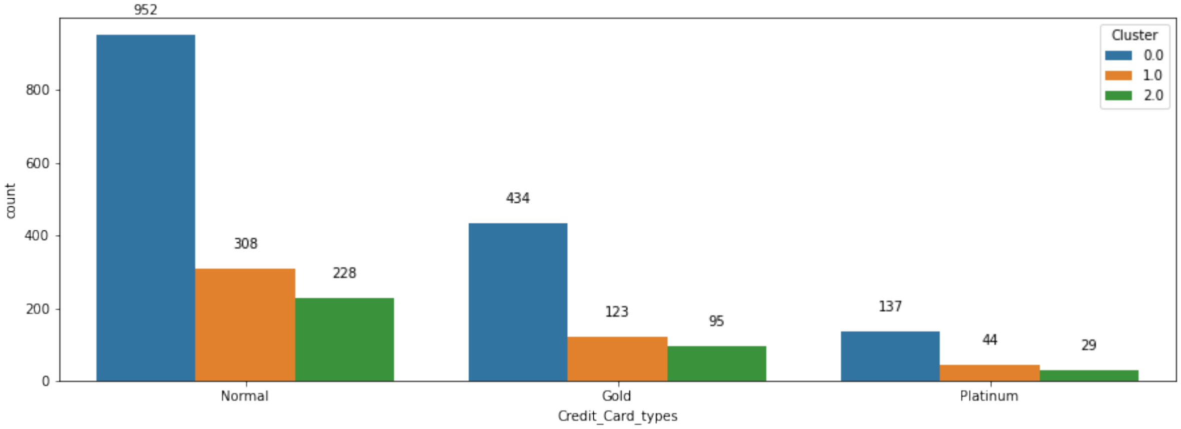


Figure 34: K-Modes Clustering of Credit Card Type

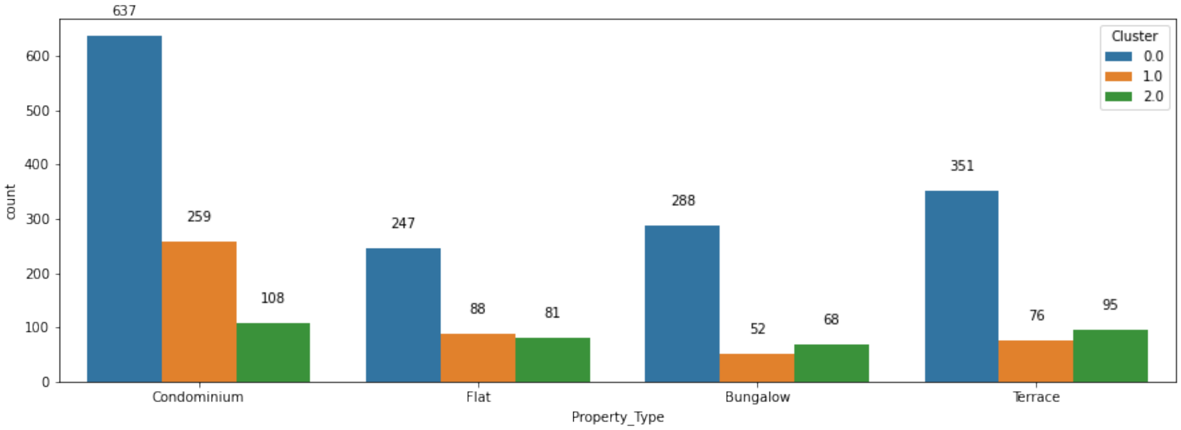


Figure 35: K-Modes Clustering of Property Type

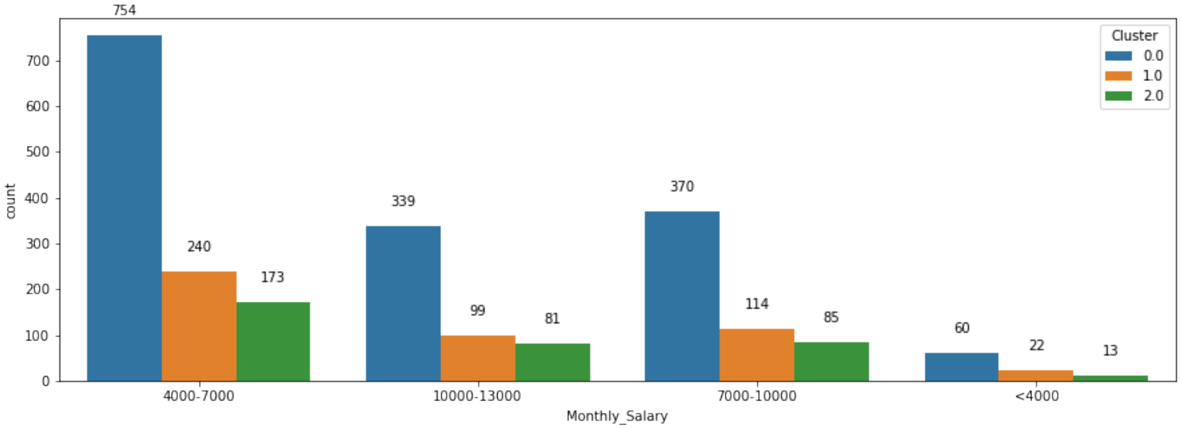


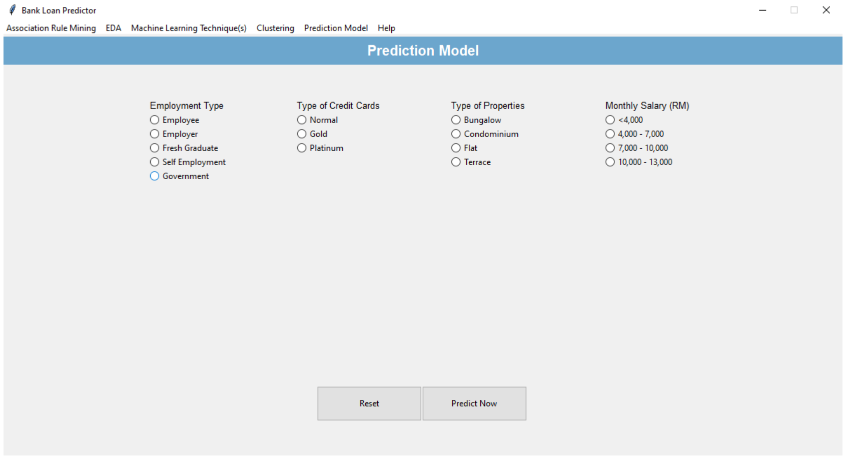
Figure 36: K-Modes Clustering of Monthly Salary

4.5 Conclusion

From the data mining tasks executed above, multiple findings about the banking data have been discovered, such as which employment type apply for loan the most, how much of the monthly salary required to apply for loan, which type of credit card user will apply for loan the most and much more. Besides that, multiple classification models have been built to predict the 'Decision' based on the applicant's loan records. At the end, the result of the Naive Bayes model provides the highest F1-score. Thus, for our prediction model, we will be implementing Naive Bayes as our model for training and predicting the dataset. Besides that, the K-modes clustering technique is also performed on the dataset.

5 Deployment

For deployment, we implemented Tkinter to create our prediction model. Tkinter is an integrated library which comes along with python. Upon landing page, the user is able to select ARM, EDA, Machine Learning Techiniques, Clustering and Prediction Model from the title bar. To run the program, the user would need to open it up in the command prompt and type in **python3 Project.py**. After that, Figure 37 will be displayed.



Employment Type
○ Employee
○ Employer
○ Fresh Graduate
○ Self Employment
○ Government

Type of Credit Cards
○ Normal
○ Gold
○ Platinum

Type of Properties
○ Bungalow
○ Condominium
○ Flat
○ Terrace

Monthly Salary (RM)
○ <4,000
○ 4,000 - 7,000
○ 7,000 - 10,000
○ 10,000 - 13,000

Figure 37: Prediction Model

When the user click on the EDA tab, the user will be able to view the visualization of the data before and after applying SMOTE.

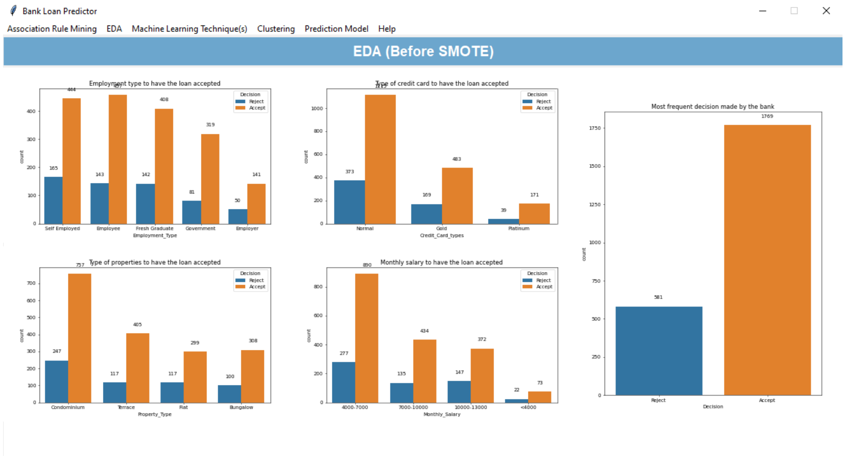


Figure 38: Before Smote

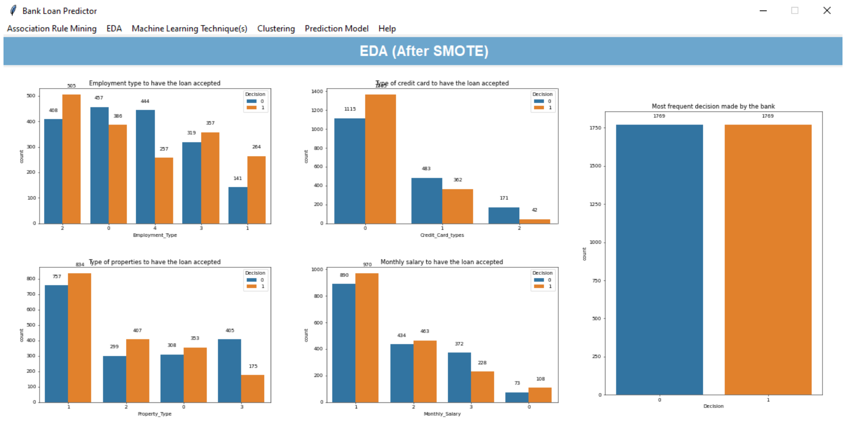
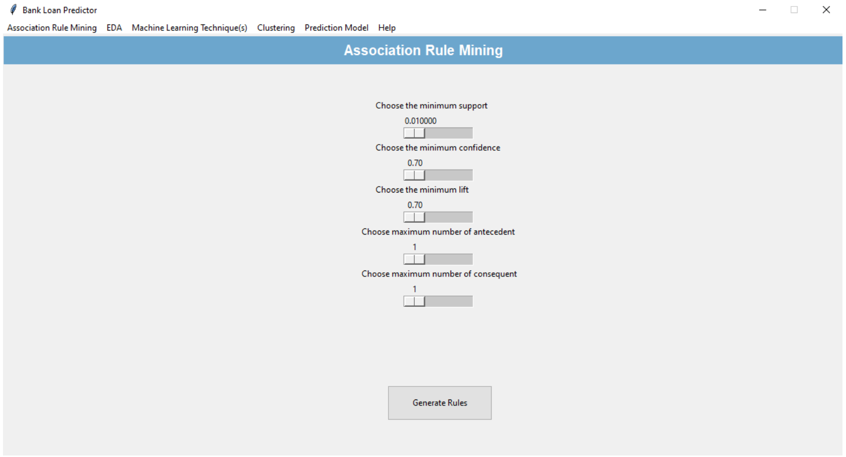


Figure 39: After Smote

On the Association Rule Mining tab, the user will be able to set the minimum support, minimum confidence, minimum lift, number of antecedent and number of consequent. After selecting all those parameters, the user can generate the rule by clicking the ”Generate Rules” button.



Choose the minimum support
0.010000

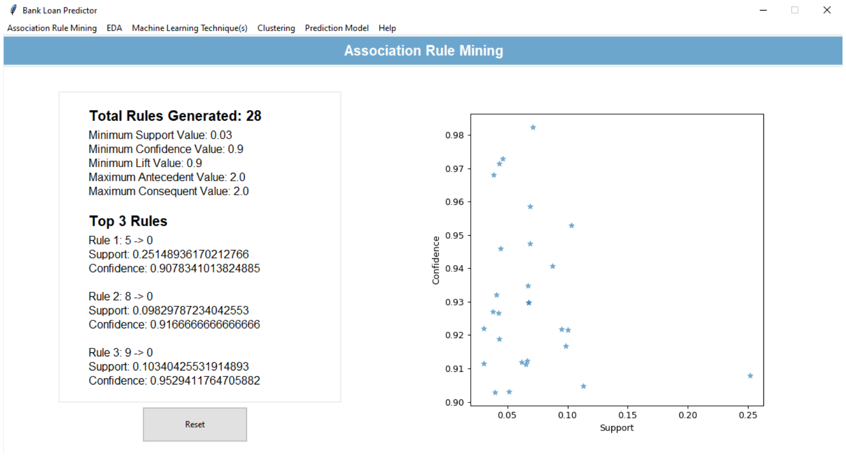
Choose the minimum confidence
0.70

Choose the minimum lift
0.70

Choose maximum number of antecedent
1

Choose maximum number of consequent
1

Figure 40: Association Rule Mining



Total Rules Generated: 28

Minimum Support Value: 0.03
Minimum Confidence Value: 0.9
Minimum Lift Value: 0.9
Maximum Antecedent Value: 2.0
Maximum Consequent Value: 2.0

Top 3 Rules

Rule 1: 5 -> 0
Support: 0.25148936170212766
Confidence: 0.9078341013824885

Rule 2: 8 -> 0
Support: 0.09829787234042553
Confidence: 0.9166666666666666

Rule 3: 9 -> 0
Support: 0.10340425531914893
Confidence: 0.9529411764705882

Figure 41: ARM after

On the Machine Learning Technique tab, the user are able to choose machine learning techniques such as SVM, KNN, Naive Bayes and Decision Tree. After selecting the ML techniques, the user will be able to view the Receiver Operating Characteristic curve, Precision-Recall Curve and F1-score of classification models.

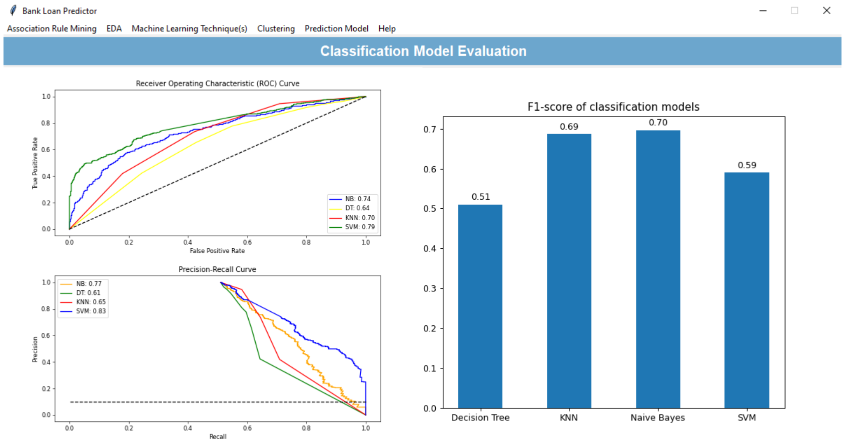
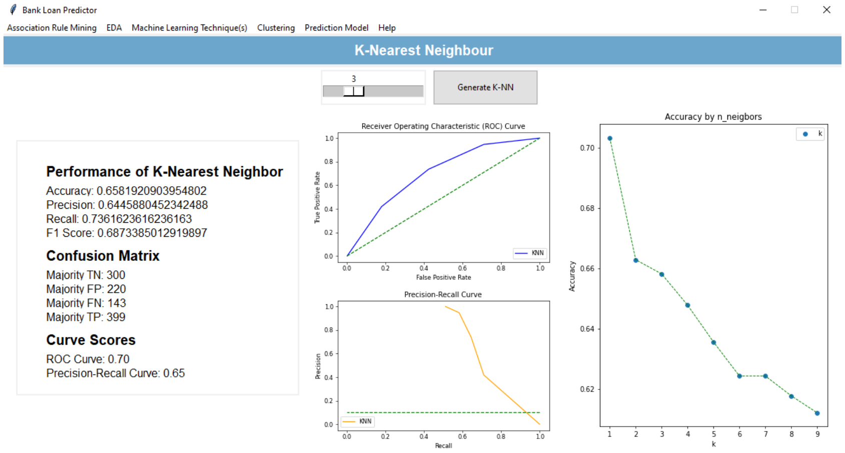


Figure 42: Classification Model Evaluation



Performance of K-Nearest Neighbor

Accuracy: 0.6581920903954802
Precision: 0.6445880452342488
Recall: 0.7361623616236163
F1 Score: 0.6873385012919897

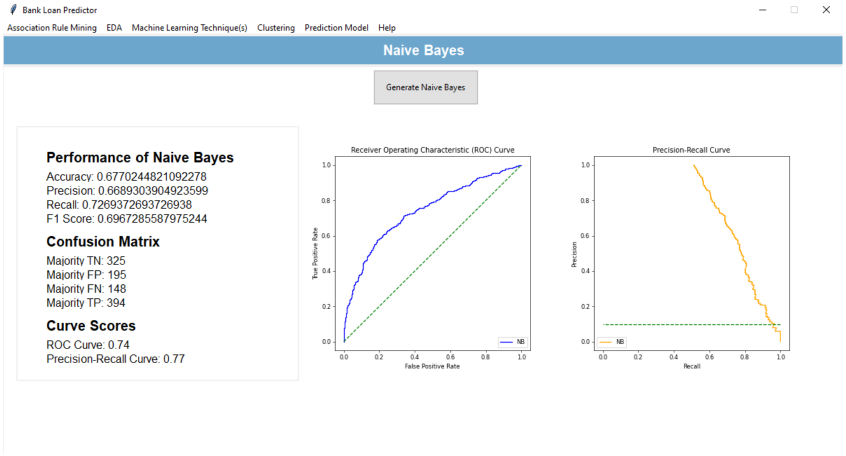
Confusion Matrix

Majority TN: 300
Majority FP: 220
Majority FN: 143
Majority TP: 399

Curve Scores

ROC Curve: 0.70
Precision-Recall Curve: 0.65

Figure 43: K-Nearest Neighbour



Performance of Naive Bayes

Accuracy: 0.6770244821092278
Precision: 0.6689303904923599
Recall: 0.7269372693726938
F1 Score: 0.6967285587975244

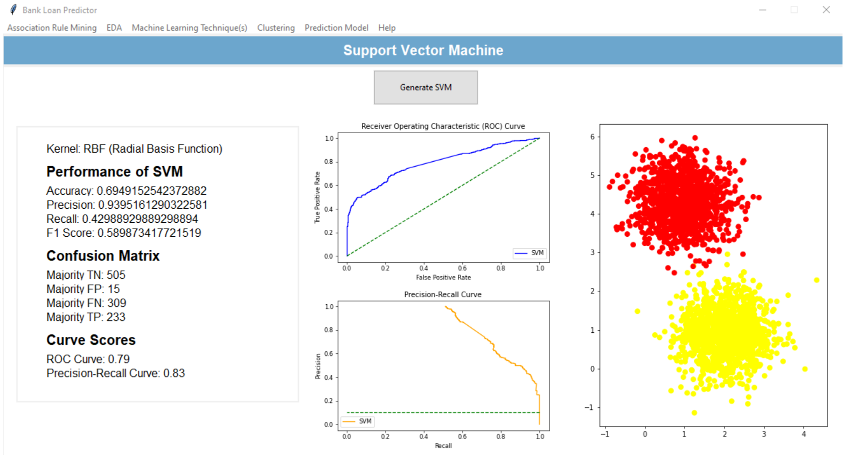
Confusion Matrix

Majority TN: 325
Majority FP: 195
Majority FN: 148
Majority TP: 394

Curve Scores

ROC Curve: 0.74
Precision-Recall Curve: 0.77

Figure 44: Naive Bayes



Kernel: RBF (Radial Basis Function)

Performance of SVM

Accuracy: 0.6949152542372882
Precision: 0.9395161290322581
Recall: 0.42988929889298894
F1 Score: 0.589873417721519

Confusion Matrix

Majority TN: 505
Majority FP: 15
Majority FN: 309
Majority TP: 233

Curve Scores

ROC Curve: 0.79
Precision-Recall Curve: 0.83

Figure 45: Support Vector Machine

On the clustering tab, the user will be able to view the visualization of the K-Nearest NeighBour. By changing the sliding bar, the visualized data will be able to change immediately.

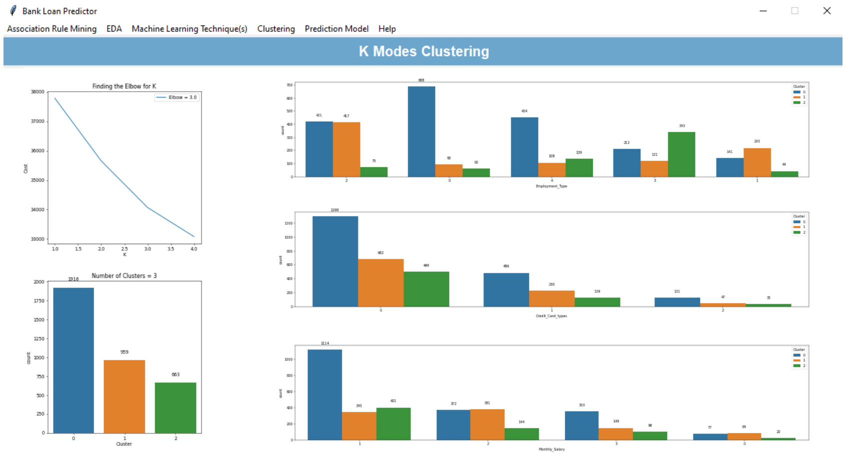


Figure 46: K-Modes Clustering