Credit Card Approvals

Oğuzhan Aydın
Middle East Technical University
Ankara, Turkey
aydin.oguzhan_01@metu.edu.tr

Abstract—**The main purpose of this project is to examine and analyze credit card approvals with various variables including demographics, years of employment, and so on. Different machine learning algorithms such as artificial neural network, xgboost, random forest, support vector machine, and regression models such as logistic regression, LASSO and ridge regression are developed to predict the credit card approvals, while the research questions are developed to help in understanding the data. To have a tidy, clean and discoverable dataset, data cleaning process is applied. The variable names are changed, outlier analysis is applied. Afterwards, statistical tests and models are performed. The model performances are tested and compared via confusion matrix, especially with accuracy, sensitivity, specificity and so on. R-Studio is used for this project with various libraries.**

Keywords—Artificial Neural Network, XGBoost, Random Forest, SVM, Lasso, Ridge, Logistic model, Data Cleaning, Credit Card Approval

I. INTRODUCTION

As a result of many changes in the world such as technological advances and globalization, credit cards that offer ease of payment to people came into life. Moreover, the use of credit cards spread quickly among the people. In 2019, approximately 2.8 billion credit cards were used actively in the global, while the world population was equal to 7.7 billion including babies and children. According to the statistics, half of the American people have more than one credit card on the average. However, it is not easy to get a credit card approval. It depends on many situations from debt to credit score.

The data set includes various variables that possibly influence the approval for the credit card. In this paper, the possible factors are analyzed, the credit card approvals are predicted by using several machine learning algorithms such as Support Vector Machine, XGBoost, Random Forest, Artificial Neural Network. All these models are compared via confusion matrix with the value of accuracy, sensitivity and specificity.

II. LITERATURE REVIEW

Significant research and extensive efforts have been dedicated to the study and analysis of Credit Card Approval. Numerous studies, investigations, and developments have been conducted in this field to enhance the understanding of the factors influencing credit card approval processes and to improve the accuracy and efficiency of credit card approval systems. In the study of IJEAT[1], the study on predicting credit card approval for customers was analyzed through customer profiling using both Decision Tree and K-Nearest Neighbors algorithms. In another study at IJRASET[2], the approval of credit card was analyzed using classification algorithms. The better result is achieved by Gradient Boosting Algorithm. Moreover, in the study[3], the best accuracy was achieved with Random Forest and Logistic Regression.

III. METHODOLOGY

A. Dataset:

The studied data set was taken from Kaggle. The raw data set is used instead of the cleaned version in the same file, since the missingness was omitted instead of imputation in the cleaned version. The data has 690 observations with 16 variables. Out of these 16 variables, 5 are numeric variables, and 10 are categoric variables. ZipCode is character variable. It is omitted and not used in the analysis. In the data set, there were no column names at first. Some of the observations were unmeaningful. The levels of the factor variables are changed based on the data description. The empty observations or unmeaningful observations were converted to NA's. The levels of the categorical variables are explained roughly in the table. The response variable is selected as Approved which represents the approval for credit. The dataset initially includes total of 54 missing observations under different variables. There is a list of variables which are used in the study below.

- **Gender**: Gender of applicant – Categorical variable with 2 levels.
- **Age**: Age of applicant – Continuous variable
- **Debt**: Debt of applicant - Continuous variable
- **Married**: Marital Status of applicant – Categorical variable with 2 levels.
- **Bank Customer**: If the applicant has a bank account or not - Categorical variable with 2 levels.
- **Industry**: Industry that the applicant works in - Categorical variable with 14 levels.

- **Ethnicity**: Ethnicity of the applicant - Categorical variable with 5 levels.
- **YearsEmployed**: Number of years worked – Continuous variable.
- **PriorDefault**: Categorical variable with 2 levels.
- **Employed**: Working status of applicant - Categorical variable with 2 levels.
- **CreditScore**: Credit Score of applicant – Continuous variable
- **DriversLicense**: If the applicant has a driver's license - Categorical variable with 2 levels.
- **Citizen**: How the applicant acquired the citizenship - Categorical variable with 3 levels.
- **Income**: Income of the applicant – Continuous variable
- **Approved**: If the applicant is successful in applying for credit card - Categorical variable with 2 levels.

B. Descriptive Statistics:

The tables below show the descriptive statistics of the variables that are used in the analysis of the research questions.

	Min	1st Q	Median	Mean	3rd Q	Max	NA's
Age	13.75	22.6	28.46	31.57	38.23	80.25	12
Debt	0	1	2.75	4.75	7.2	28	-

Table 1 Descriptive Statistics of Continuous Variables

Gender	Approved	Citizenship	Industry	Driver's License
Male:210 Female:468 NA's:12	1: 307 0: 383	ByBirth:625 ByOtherMeans:57 Temporary:8	Energy:137 Materials:78 Industrials:64 Other:402 NA's:9	1: 374 0: 316

Table 2 Descriptive Statistics of Categorical Variables

Table 1 shows the mean and the 5 number summary of the continuous variables that are used in the research questions. The range for the age variable is larger than the debt. However, there exists NA values in age variables. Since there is a noticeable difference between the 3rd quantile and the max value for both variables, we may say that there exists outlier.

Table 2 shows the number of observations for each level. For the *Approved* variable, 1 represents the getting approval for the credit card application while 0 represents the rejection. For the *Driver's License* variable, 1 represents the having the license while 0 represents not. Focusing the response variable, which is Approved, it can be said that the data set is balanced.

C. Exploratory Data Analysis:

In this section of the study, 5 research questions were generated and answered. These questions and answers of them helped to understand the data and the relationship between the variables within the data.

C.1 Is there an association between gender and the likelihood of being approved for credit card?

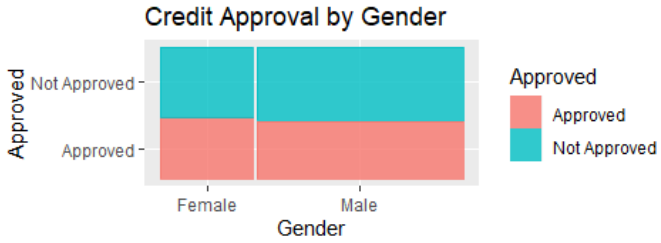

Figure 1 Mosaic Plot for Credit Card Approval by Gender

Mosaic plot is created to see whether there is an association between the gender and getting approval for the credit card, visually. Plot suggests that there is no association between them. To confirm it, a contingency table was created first. Then, the odds ratio is calculated. The output of the test shows that getting approved is 1.11 times higher for males compared to females. However, this is not statistically significant since the interval for odds ratio includes 1, which indicates that there is no association. Moreover, chi-squared test was conducted. Since the p-value of the test (0.54) is larger than 0.05, it can be concluded that there is no significant association between gender and credit card approval.

C.2 Does the citizenship status of individuals have an impact on their debt?

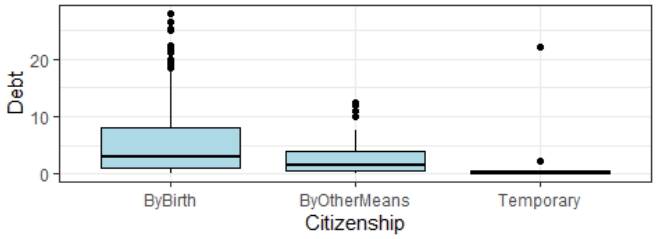

Figure 2 Boxplot for debt by citizenship status

To see whether there is a difference for the debt among citizenship status, boxplot was created. From the boxplot, it is clear that there is a difference among the citizenship status for median, max, IQR. Also, all the three has outliers. To determine whether the citizenship status of individuals have an impact on their debt, ANOVA assumptions were checked. The debt were not normally distributed. After the

transformation, normality assumption was satisfied. However, the assumption of homogeneity of variances is violated. Therefore, nonparametric Kruskal-Wallis test was considered. Since the p-value of Kruskal-Wallis test is smaller than 0.05, it can be concluded that there is a significant impact of citizenship status on debt.

C.3 Does gender and having driver's license have an interaction effect on years of employing?

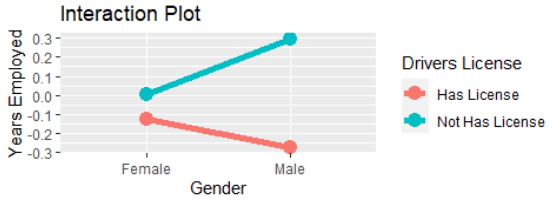

Figure 3 Interaction Plot

Interaction plot was created to see whether there is an interaction. Since the two lines are not parallel, there might be an interaction effect. To make sure whether driver's license and gender have an interaction effect on years of employing, ANOVA test was conducted. Firstly, the assumptions are checked. Years of employed variable seems not normal. Therefore, the ordered quantile normalization transformation, which is given as best transformation for this variable, was applied. After the transformation, Kolmogorov-Smirnov test suggests that the variable is distributed normally, and Bartlett test suggests the variances of the groups or samples being compared are equal. Then, ANOVA test was conducted. The test results indicates that there is no statistically significant interaction effect between gender and having driver's license on years of employing.

C.4 Is there a significant relationship between debt and Age?

To understand whether there is a linear relation and correlation between debt and age, scatter plot was created first.

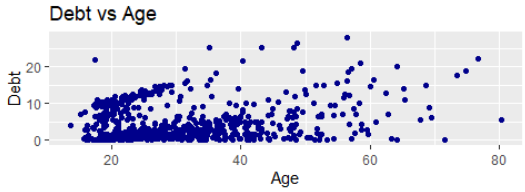

Figure 4 Scatter Plot of Debt and Age

There might be a positive correlation between these two variables. However, it is not strong. When the correlation coefficient is checked, it equals approximately 0.21, which is positive but not high. It seems there is an increasing trend. Also, variation is high. There might be a heteroscedasticity. Pearson's product-moment correlation was conducted formally. Since the p-value is less than 0.05, it can be concluded that there is a significant and positive correlation between debt and age.

C.5 Which 5 industries has the largest number of approval and largest probability of being approved?

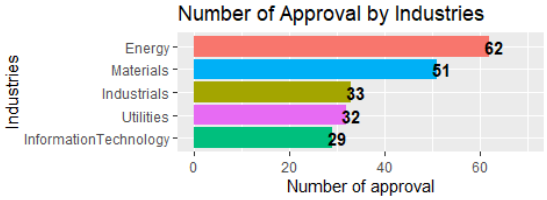

Figure 5 Number of Approval by Industries

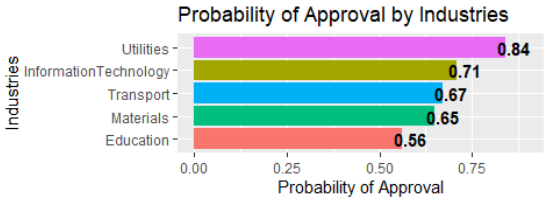

Figure 6 Probability of Approval by Industries

When the data set is checked to determine which industry has the largest number of approvals, it seems that the Energy industry has the largest number of approvals, as can be seen in Figure 4. Since the number of observations from each industry is not equal, this might be misleading. Therefore, the probability of approval for each industry is checked, Figure 5. According to the probabilities, the Utilities Industry has the largest probability for approval which is 4th in the number of approvals.

D. Missingness

Since the data already contains 54 NA values, there is no need to generate NA values. In order to comprehend the missingness mechanism, we can investigate the following plots and perform statistical tests.

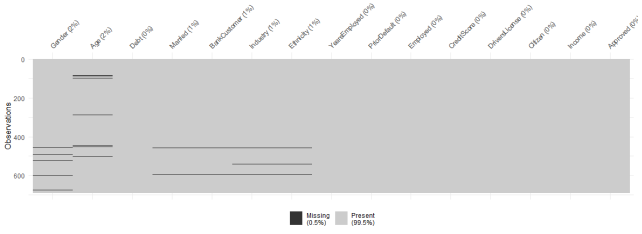

Figure 7 The Aggregation Plot of Missing

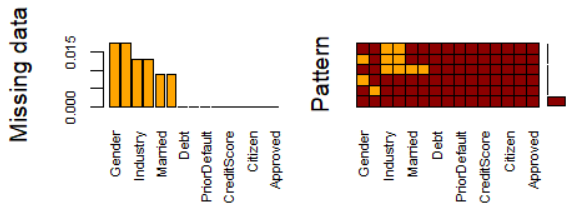

Figure 8 The Aggregation Plot of Missing

The missing values are represented as orange and non-missing observations are represented as red. The proportion of missingness in each variable is shown in the bar plot, which tells that gender and age variable have the highest number of missing values.

Variable	Number of Missing
Gender	0.017
Age	0.017
Industry	0.013
Ethnicity	0.013

Table 3 The missingness proportion of the variables

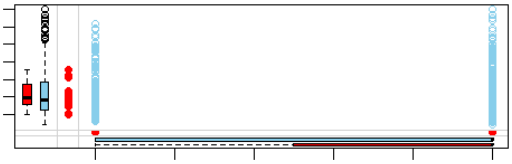

Figure 9 Margin Plot for the Missingness Mechanism

In the margin plot, the variables which have the highest missing percentages, Age and Gender, are plotted. Since there is not a significant difference between boxplots for age and gender, it can be said these variables have the same behavior. As a result, it can be concluded that the missingness mechanism of this data is Missing at Random (MAR). "mice" package was used for the imputation.

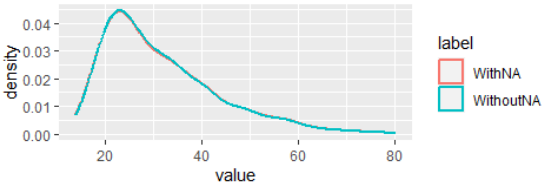

Figure 10 Density plot of Age with and without NA's

As seen from the density plot, the imputation for age variable does not cause a considerable change in the distribution of the variable. Therefore, imputation can be evaluated as appropriate.

E. Model Preparations

E.1: One-Hot Encoding

In order to apply machine learning algorithms without having a problem, one hot encoding was performed on the categorical variables which have at least 3 levels. In this way, categorical variables were converted into numerical representations.

E.2: Scaling

Scaling was performed in order to standardize all features to a similar scale and range. In this way, potential biases may be alleviated and the problems that might arise due to varying scales can be avoided.

E.3: Feature Selection

As a result of one-hot encoding, the dimension of the data increased from 15 to 34. In order to reduce the overfitting and enhance interpretability of the model, a feature selection method, Boruta, was applied. It selected the variables that are found important and eliminated the others. Finally, the studied data set has 18 variables.

E.4: Cross Validation

In the cross-validation part, the data set is divided into 2 sets: train and test. The train set has 80% of the data while the test set has 20%. While doing cross validation, different cross validation techniques such as leave-one-out cross validation, 10-fold cross validation, repeated 10-fold cross validation and validation set approach were performed. These techniques were compared according to accuracy and sensitivity scores. Since Repeated 10 Fold Cross Validation Approach has the highest accuracy, it is selected as CV Technique. It is followed by 10-Fold CV, LOOCV and Validation Set Approach.

F. Modelling

F.1 Multiple Logistic Regression:

Multiple Logistic Regression is specialized to evaluate the relationship between the response value, which is approval for the credit card in the data set, and the various predictor variables. It estimates the likelihood of approval by considering different variables such as debt, income, ethnicity and so on. The model of multiple logistic regression is:

$$logit(p) = \beta_0 + \beta_1 * x_1 + \cdots + \beta_p * x_p$$

Firstly, logistic regression model was conducted with all the variables. However, the model coefficients were not meaningful. Afterwards, VIF values were checked, and multicollinearity was conducted. Also, there were unmeaningful variables.

To overcome this problem, stepwise elimination method was applied. VIF values were checked and all the VIF values were less than 5.

	Coef	StdError	p-value
Intercept	-0.99	0.80	0.2
BankCustomer1	-0.54	0.34	0.11
PriorDefault1	-3.67	0.35	<0.05
`EmployedNot Employed	-1.41	0.31	<0.05
Industry.IT	-2.33	0.77	<0.05
Industry.Utilities	-3.09	0.89	<0.05
Ethnicity.Black	-0.63	0.37	0.09
Citizen.ByBirth	4.61	0.88	<0.05
Citizen.ByOtherMeans	4.52	0.98	<0.05
Income	-2.27	0.93	<0.05
YearsEmployed	-0.56	0.19	<0.05
Null deviance: 759.88 on 552 degrees of freedom Residual deviance: 321.84 on 542 degrees of freedom AIC: 343.84			

Table 4 Result of Multiple Logistic Regression

The overall model is statistically significant. Most of the variables are statistically significant. Focusing the estimated coefficient, being bank customer has a negative sign. It is expected to have a positive sign, and seems unmeaningful. The p-value of the variable indicates that the variable does not have statistical significance. It can be predicted that not working may have a negative effect on obtaining credit card approval. The coefficient of NotEmployed is negative. It is in the same way of our prediction. Having negative sign for the variable of Income and YearsEmployed is not meaningful since it is expected that as the income and employing years increase, the probability of getting approval increases. The presence of unexpected signs in the regression coefficients could be attributed to confounding factors that may have an influence on the relationship between the predictor variables and the outcome. Therefore, further investigation should be conducted to identify potential confounding factors and their impact on the results. Lastly, when the null and residual deviances are compared, it is possible to conclude that the logistic regression model has a reasonably good fit, as indicated by the relatively lower residual deviance than the null deviance.

F.2 Lasso Regression

Secondly, Lasso Regression was performed, which uses shrinkage. First of all, the expand grid is prepared for tuning the parameter. The best accuracy was reached with the value of 1 for alpha and 0.03 for lambda.

F.3 Artificial Neural Network

The third method for predicting credit card approval is Artificial Neural Network which is a machine learning algorithm. The grid search was applied by parameter tuning both for size and the decay. The optimal size was conducted as 1 while the decay was 0.2.

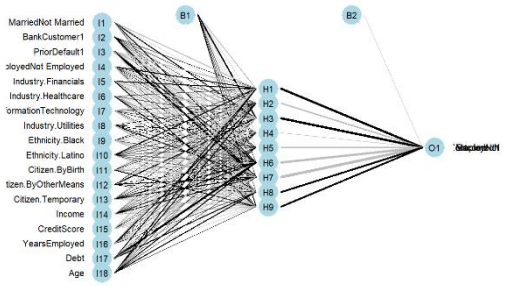

Figure 11 Neural Network Visualization of the Model

Figure 11 shows there are two layers NN model. The first layer has 18 neurons while the second layer has 9. Moreover, 20 weights were used in order to produce the final output.

F.4 Support Vector Machine

Fourth model is Support Vector Machine (SVM) which is a supervised machine learning method that can be used in the analysis for classification and regression. GridSearch is created to find optimal hyperparameter values for gamma values. The parameters for the cost was tuned from 0.1 to 10, while for gamma 0.01 to 1. Finally, The slight improvement in performance was observed.

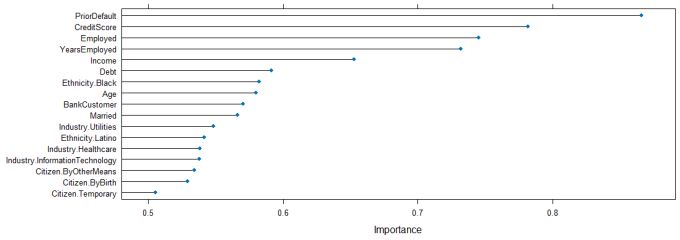

Figure 12 Importance of Variable Plot for SVM

Figure 11 represents the importance of the variable in the Support Vector Machine model which measures the contribution of each predictors towards the performance of the model. CreditScore, Employed, YearsEmployed, Income and Debt are important variables which were utilized in the research questions. PriorDefault is the most

important variable with 0.86 ROC curve variable importance.

F.5 Random Forest

As the fifth model, Random Forest, ensemble learning method, was performed. While creating random forest model, parameter tuning was applied on mtry by expand grid command. The optimal value is mtry equals to 2. Accuracy was used to select best model and optimal value.

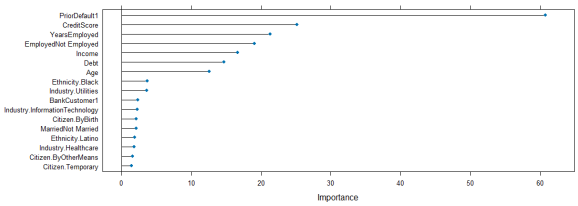

Figure 13 Importance of Variable Plot of Random Forest

Similar to the SVM, importance of variable plot was created. Although the contribution of the variables is different, the most important variables are similar, PriorDefault1, CreditScore, Employed, YearsEmployed, Income and Debt. PriorDefault is the most important variable as in SVM.

F.6 XGBoost

GridSearch is carried out with repeated 10 Fold cross-validation with the method "xgbTree". In order to do parameter tuning, eta, maximum depth, nround, min child weight and colsample bytree are changed progressively. At the end, the best accuracy for XGBoost method was achieved with 50 nround, 0.4 eta, 1 maximum depth, 0.6 colsample bytree, 1 min child weight, 1 subsample and 0 gamma.

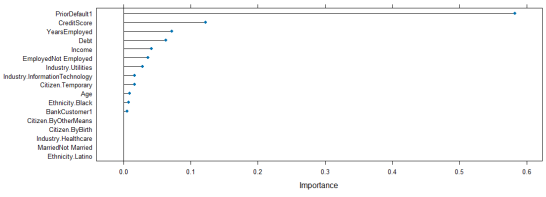

Figure 14 Importance of Variable Plot of XGBoost

The variable importance plot was created. Unlike the variable importance plot of SVM and Random Forest, variable importance plot of XGBoost claims that Citizen.ByOtherMeans, Citizen.ByBirth, Industry.Healthcare, MarriedNot Married, Ethnicity.Latino is unimportant. Similar to the other importance plots, XGBoost also claims that there are important variables such as CreditScore, Employed, YearsEmployed, Income, Debt and so on. PriorDefault is the most important variable as in SVM and Random Forest. However, the ROC curve of PriorDefault importance is the lowest compared to the others with 0.58.

G. Performance Comparison

The performance of the model for both train and test sets are evaluated. The train performance serves as a measure to assess the model's fit to the data, indicating how well it captures the patterns and relationships within the training dataset. On the other hand, the test performance evaluates the model's ability to make accurate predictions when presented with new, unseen data. The performances were compared based on accuracy, sensitivity, and specificity.

Model	Accuracy	Sensitivity	Specificity
Logistic Model	0.86	0.89	0.84
Lasso	0.85	0.92	0.79
ANN	0.89	0.89	0.89
SVM	0.89	0.90	0.88
Random Forest	0.92	0.88	0.95
XGBoost	0.89	0.89	0.89

Table 5 Model Performances on the train set

Model	Accuracy	Sensitivity	Specificity
Logistic Model	0.88	0.91	0.85
Lasso	0.87	0.93	0.82
ANN	0.86	0.86	0.85
SVM	0.88	0.93	0.84
Random Forest	0.87	0.83	0.90
XGBoost	0.87	0.87	0.88

Table 6 Model Performances on the test set

IV. DISCUSSION AND CONCLUSION

In this study, the main objective is to predict the credit card approvals. The first step towards this goal was to clean the data set. Adjustments have been made on the variable names and the observations. Then, focusing on the descriptive statistics, research questions related to data set were generated. In the stage of research question, the relations between the variables were figured out, and gained insights into the current state of knowledge and research findings. In some questions, transformations were applied in order to satisfy the assumption of related tests. Then, the missingness mechanism of the data set was estimated as Missing at Random. Imputation was performed in accordance with the missingness mechanism. Afterwards, the data set was prepared for the modelling part. One-hot encoding is applied on the categorical variables, and converted into numerical representation. Scaling was performed and all features were standardized to a similar scale and range. The increased number of variables as a result of one hot encoding has been reduced by Boruta. Then, the data set was splitted into train and

test sets by repeated 10-fold cross validation technique. Afterwards, various models were conducted with different techniques which are logistic regression, LASSO, random forest, XGBoost, SVM, and ANN. Finally, the performance of these models on test and train sets was compared. All models demonstrated strong performance, with high accuracy, sensitivity, and specificity. Random Forest has the highest accuracy and specificity, and performs consistently on the test set, also. On the other hand, XGBoost performed approximately the same and successful performance on both sets. To sum up, although all the models were successful, the best model can be selected as Random Forest since it performs well across all three metrics on both the train and test sets.

V. REFERENCES

[1] Hubert, A. C., Vimalesh, R., Ranjith, M., & Raj, S. A. (2020). Predicting credit card approval of customers through customer profiling using machine learning. International Journal of Engineering and Advanced Technology, 9(4), 552–557. https://doi.org/10.35940/ijeat.d7293.049420

[2] Dalsania, N., Punatar, D., & Kothari, D. (2022). Credit card approval prediction using classification algorithms. International Journal for Research in Applied Science and Engineering Technology, 10(11), 507–514. https://doi.org/10.22214/ijraset.2022.47369

[3] Peela, H. V., Gupta, T., Rathod, N., Bose, T., & Sharma, N. (2022). Prediction of credit card approval. International Journal of Soft Computing and Engineering, 11(2), 1–6. https://doi.org/10.35940/ijsce.b3535.0111222