Credit Card Approval System

Project Report

Abstract

In the banking system, there is a major task to determine if the person is eligible for the credit card. Profit to the credit card providers largely depend on if person is paying back.By predicting correct credit card customers ,the banks can maximize the profits.We here try to automate this process by applying machine learning techniques. Here we need to make sure to target the right customers.

1 Introduction

The correct assessment for credit card approval is very important for banks and organisations who lend credit card to the people. The recent years has seen huge growth in credit cards and loans. The exact judgement of person to be approved for credit cards allows the organisations to minimize losses and same time make suitable credit arrangements as per requirement.Due to huge growth in number of applicants there is need for more sophisticated method to automate the process and speed it up.

2 Importance of Project

Credit card approval can very helpful for organisations that lend credit cards and due to increase in huge number of applicant there is need to automate the task and classify the applicants into if they are eligible for credit card or not. This helps to avoid organisation losses by avoiding potential defaulters. Here we are not just looking into bank balance but into there personal attributes like gender,married,age,Occupation etc.We account for these personal attributes to evaluate if given is applicant is good customer.This can also help cut down weeks long process into few days.This gives benefit by cutting down costs on credit analysis and faster credit decisions.

3 Literature Survey

In past various models have been proposed to determine the and evaluate the credit scoring criteria. Techniques can be further classified into parametric and non-parametric models.The most well liked parametric model were logistic regression and Linear discriminant analysis.LDA has been criticized because of categorical nature of data. Logistic regression overcomes these problems later turn out to be common credit scoring tool for credit lending organisations.The prediction is taken into consideration after applying sigmoid function into it.[1]

Non-parametric techniques that can be used are K-nearest neighbour,decision tree and support vector machines. SVM can be combined with backpropogation neural network to get better accuracy.The results displayed SVM's accuracy comparable to that of backpropogration neural network[2] Further there have been improvement in various other hybrid data mining techniques to get better results at these kinds of problems.Various ensemble method have also been to used to get improved accuracy by aggregating scores.

Credit card approval is also being done through genetic programming, this paper examines the usage of strong typed genetic programming for automated credit approval.Eight different genetic programming approaches where applied and compared.[5]

In another paper implementation based on logistic regression has been compared with XGBoost algorithm.It is found that XGBoost algorithm has significantly higher model discrimination and model stability than that of logistic regression.[6]

4 Data Description

4.1 Data Source:

The Data Set we have obtained from UC Machine Learning data repository [1] submitted by quinlan '@' cs.su.oz.au. The data contains various variables such as Gender, Age, Education, Employed ....

4.2 Data Description:

The data have been encoded to some special text to protect the confidentiality of the persons. But the encoding does not affect our purpose. In the dataset there are 690 instances out of them 70% are used for training and rest 30% used for validation. There are continous values as well as categorical values. There are 16 columns first 15 are credit application attributes and the last one is the approved columns which contain boolean value either the credit card application is approved or rejected.

5 Data Analysis Processing

The data comprises of 690 instances having 16 attributes the last column is dependent variable while other 15 are independent variables. $Gender$, $Married$, $BankCustomer$, $Education$, $Ethnicity$, $Employed$, $PriorDefault$, $Employed$, $DrivingLicense$, $Citizenship$, $Approved$ are categorical attributes Age, $Debt$, $YearsEmployed$, $CreditScore$, $ZipCode$, $Income$ are continuous attributes.

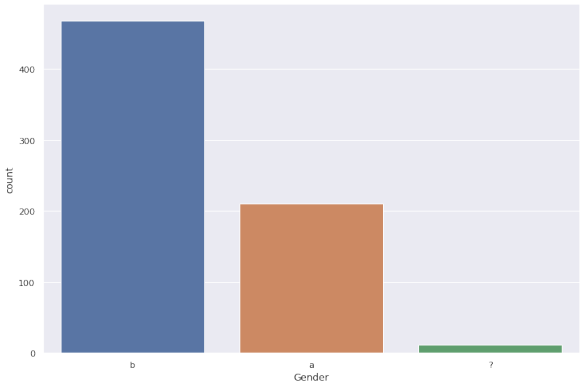
5.1 Anomalies

Some of the attributes having some outliers in continuous data i.e. they contain some abnormal values that must be replaced with other to predict the output correctly. To reduce the anamoly affect, we have replaced those with mean value of that column.

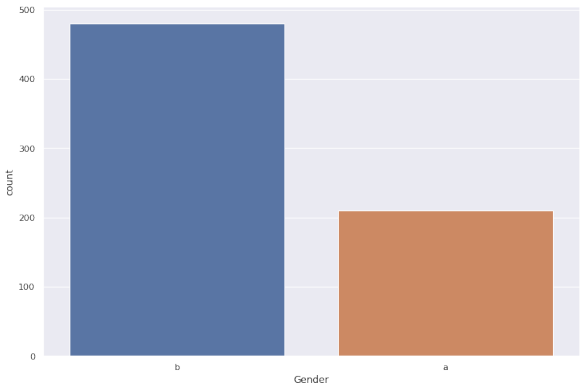
5.2 Missing Data Handling

Some of the instances does not have any data of a particular attribute and is represented by ?. In categorical value we replaced the missing value by the highest occuring value of that attribute, while in the numeric data type it is replaced by the mean of that attribute. For example attribute Gender consist a:480, b:214, ?:12 so the ? is replaced by a.

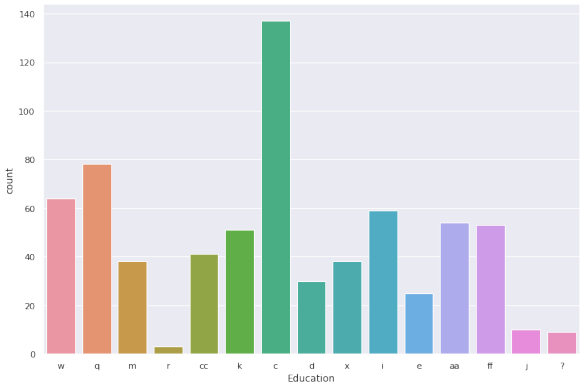
Gender Attribute Before Mapping



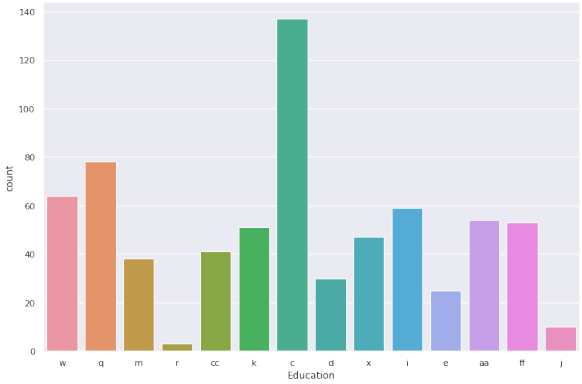
Gender Attribute After Mapping



Education Attribute Before Mapping



Education Attribute After Mapping



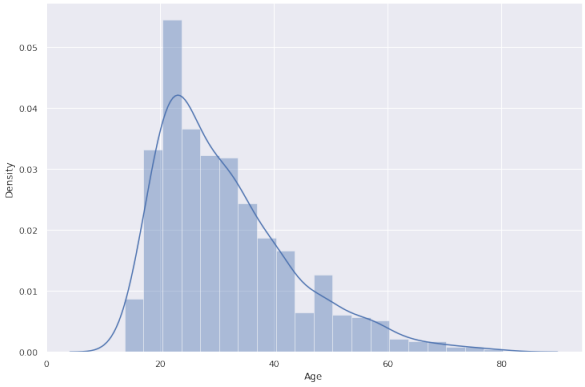
5.3 Data Encoding

The categorical values needed to be encoded in some numerical value type before passing the dataset to the model for training, we counted the

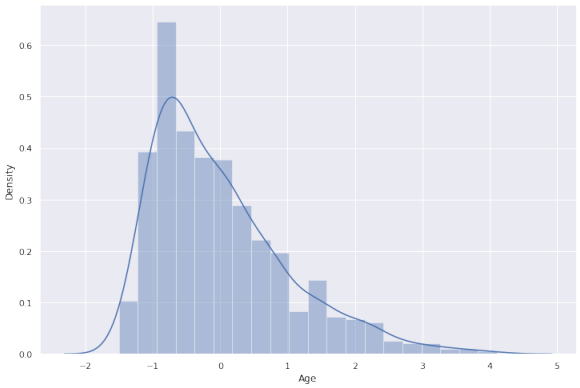
distinct values of that attribute and then replaced them with corresponding integer.

5.4 Standardization of Data

Some of the numerical attributes vary with a very large difference, so they need to be Standardized to better prediction of our model, we Standardized the value by subtracting mean from it and then dividing by the standard deviation of that column.
Age before Standardization



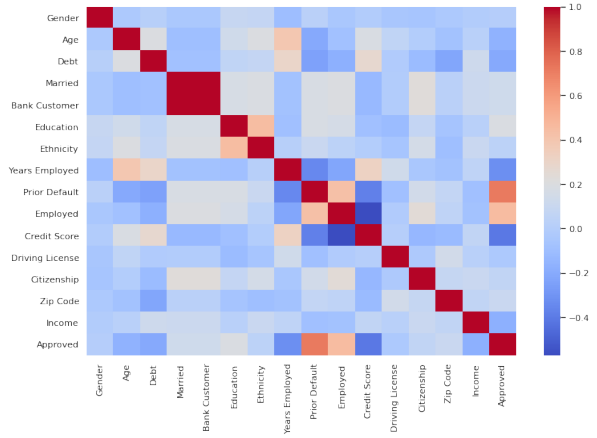
Age after Standardization



5.5 Analysis

Some of the attributes affect the prediction most, while some does not strongly participate in the prediction but sometimes they can also reduce the accuracy of our system, we can also remove those features for example the gender does not affect the approval of the credit card, so we can also remove this feature from our dataset.

This is the correlation matrix of all the features



As we can see that the attribute affects output the most are Age, Debt, Years Employed, Prior Default, Employed, Credit Score, Income. Attributes which contributes the least in affecting the output are Married, Bank Customer, Citizenship, Gender.

6 Model Selection

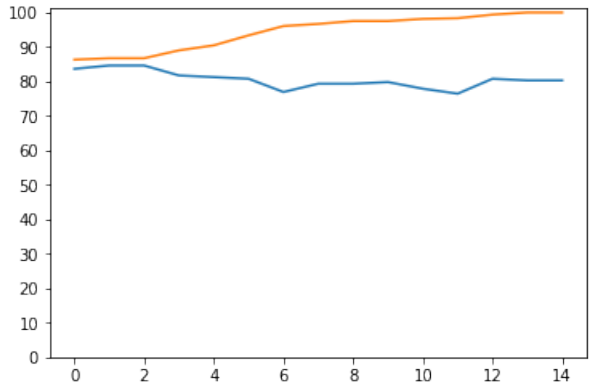
In this project we have to predict only categorical values, moreover our task is to only predict the binary values, so simply we cannot use the Linear Regression here, so we decided to use the Decision Tree Classifier.

6.1 Decision Tree Classifier

We used Decision Tree Classifier and then tuned its depth to observe the best testing accuracy, we obtained that under the depth of 3 levels the testing accuracy reaches 84% that is the highest while the training accuracy remains 86% and if the depth larger than 13 levels then training accuracy reaches 100% while testing accuracy touches 80%

Below are the prediction accuracy table of testing and training data using Decision Tree Classifier.

Depth	Training	Testing
1	0.8631	0.8029
2	0.8672	0.8365
3	0.8672	0.8462
4	0.89	0.8462
5	0.9046	0.8173
6	0.9336	0.8125
7	0.9606	0.8077
8	0.9668	0.7692
9	0.9751	0.7933
10	0.9751	0.7933
11	0.9813	0.7981
12	0.9834	0.7788
13	0.9938	0.7644
14	1	0.8077
15	1	0.8029



The graph that shows the training and validation accuracies for given depth of the tree.

6.2 Logistic Regression

We also used Logistic Regression in our project, the Logistic Regression gives the training accuracy 86.93% while the testing accuracy remains 80.29% same as the previous we got in Decision Tree Classifier, also Logistic Regression is not working well with our the data as the testing accuracy fluctuates between 72% to 84% we need to tune our parameters.

We have used grid search to find optimal parameters for logistic regression with parameters:
tolerance = [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5]
Max Iterations = [50, 100, 150, 200, 250, 300]
Obtained optimal parameters as:
Tolerance : 0.0001
Max Iterations : 150
For the same parameters, following accuracies obtained

Train Accuracy : 84.1393%
Test Accuracy : 86.7052%

6.3 Gradient Boosting Classifier

We have further implemented Gradient Boosting Classifier which is a type of ensemble algorithm. The training accuracy obtained in this algorithm was 95.5%.While the testing accuracy obtained was 87.28%.

This is the best result we have obtained from a model so far. Therefore we conclude this model to be our best model.

Train : 95.5513%
Test : 87.2832%

7 Conclusion

We tried several models to get maximum accuracy. We used Decision Tree Classifier which giving accuracy of 84%, then we used Logistic Regression with optimal parameters and obtained accuracy of 86.7%.
We also implemented Gradient Boosting Classifier to further improve accuracy and got 87.2% accuracy, which is better than all the other models used.

8 Further Improvements

Logistic Regression gives accuracy 86% with optimal parameters while the Decision Tree Classifier gives accuracy above 80%. The Gradient Boosting Classifier gives accuracy of 87.28% which is best model so far we have obtained.

Further model could be more optimized and could give better accuracy if Neural Network could be implemented. Also we could use AdaBoosting technique to further make a better model out of this.

9 References

[1] M. A. Sheikh, A. K. Goel and T. Kumar, ”An Approach for Prediction of Loan Approval using Machine Learning Algorithm”

[2]Zhang Lei-lei,HUI Xiao-feng,WANG Lei,”Application of Adaptive Support Vector Machines Method in Credit Scoring”

[3] Z. Qiu, Y. Li, P. Ni and G. Li, ”Credit Risk Scoring Analysis Based on Machine Learning Models”

[4] A. Gahlaut, Tushar and P. K. Singh, ”Prediction analysis of risky credit using Data mining classification models”

[5]Sum Sakprasat,Mark C. Sinclair,”Classification Rule Mining for Automatic Credit Using Genetic Programming”

[6]Yu Li,”Credit Risk Prediction Based on Machine Learning Methods”