Word Count: 1768

Plagiarism Percentage 40%

Matches

1 **World Wide Web Match**
View Link

2 **World Wide Web Match**
View Link

3 **World Wide Web Match**
View Link

4 **World Wide Web Match**
View Link

Suspected Content

Loan Status Prediction Sumanth Prasad - PES2201800092 Gokul K M - PES2201800517 Y R Pavan Sai - PES2201800484 ABSTRACT:- Loan prediction analysis is one such analysis which banks opt to sort which customers would be able to repay the loan if provided to them. So, the banking company should be able

> to identify the factors that are eligible for granting the loan. 1

Customers first apply for a loan which can be of any type. After that the banking company validates the customer eligibility for the loan. We have classified the customers based on the Loan_Status variable present in the dataset. I. Introduction This analysis is used to uncover the underlying structure of a relatively larger set of variables present in the dataset.. Almost all banks deploy this analysis scheme to ensure guaranteed return of the loan given. This also avoids any scams related to loans and ensures safe transactions for customers demanding the loan. The dataset comprises multiple attributes such as loan id, gender, married, education, loan amount.etc. which would directly affect the income of a person in 1994, as explained in our report which says that married people earned more than single people during the year 1994. We have used different methods such as logistic regression, K nearest neighbours and decision tree classifiers to compare the results II. Dataset We have used the customer loan dataset for our problem statement. We use 2 datasets here: one for training and one for testing. There are altogether 25 columns in our dataset. We use 13 columns for the training dataset and 12 columns for the test dataset. 3)

> Married -> Applicant who is married is represented by Y and not married is 1

represented as N. 4) Dependents ->The

number of people dependent on the applicant who took the loan. 1

5)

Education -> It is either non -graduate or graduate. 1

6)

Self_Employed -> As the name suggests Self Employed means , he/she is employed for himself/herself only. 1

An applicant who is self employed is represented by Y and the one who is not is represented by N. 7) Applicant Income -> Applicant Income suggests the income by Applicant. We assume that the 1

one who earns more has a high probability of clearing the loan amount and would be highly eligible for loan. 8) Co Applicant income -> this represents the income of co-applicant. 1

9)

Loan Amount -> This amount represents the loan amount in thousands. 1

10)

Loan_Amount_Term -> This represents the number of months required to repay the loan. 11) Credit_History 1

->

A credit history is a record of a borrower’s responsible repayment of debts. It suggests → 1 denotes that the credit history is good and 0 otherwise. 12) Property_Area -> The area 1

can be

> Urban or Semi Urban or Rural. 13) Loan_Status -> If the applicant is eligible for loan it's yes represented by Y else it's no represented by N. 1

1)

> Loan ID -> As the name suggests each person should have a unique loan ID. 2) Gender -> In general it is male or female. 1

The third gender is not included. Figure 1: Columns in the dataset III. Single variable analysis Now, we perform exploratory data analysis using a single variable. Figure 2: For variable "Gender" Figure 3: For variable "Married" Figure 4: For variable "Dependents" Figure 5: For variable "Education" Figure 6: For variable "Self_Employed" Figure 7: For variable "Property Area" Figure 8: For variable "Credit History" Figure 9: For "Loan Amount Term" Figure 10: Correlation among numeric values From the analysis done, we can come to the following conclusions: 1)

> We can see that approximately 81% are Male and 19% are female. 2) Percentage of applicants with no dependents is higher. 3) There are more number of graduates than non graduates. 4) Semi Urban people are slightly higher than Urban people among the applicants. 5) Larger Percentage of people have a good credit history. 6) The percentage of people that the loan has been approved for has been higher rather than the percentage of applicants for which the loan has been declined. 1

7)

> We see that the most correlated variables are: (ApplicantIncome - LoanAmount) 2

with correlation coefficient of 0.57 (Credit_History - Loan_Status) with correlation coefficient of 0.56 LoanAmount is also correlated with CoapplicantIncome with correlation coefficient of 0.19. IV. Preprocessing Steps included for preprocessing: 1)

> Before we go for modeling the data, we have to check whether the data is cleaned or not. 1

We use the isnull( ) function to check for missing values. 2) The missing values were replaced with median for continuous variables and mode for categorical variables. 3) As we know that a loan id must be unique to all customers, we check for the uniqueness as well for the Loan_Id variable using the unique( ) function. 4)

For modelling, we have to convert any categorical variables to numeric ones. Here, we apply the get dummies ( ) function using the regular one-hot encoding method. These preprocessing steps are enforced on both, the training and the test data sets. As we are interested in the Loan_Status variable, we analyze numerical variables by plotting boxplots against the Loan_Status variables. Figure 10: Loan_Status vs Applicant Income Figure 11: Loan_Status vs CoApplicant Income Figure 12: Loan_Status vs Loan_Amount From the plots, we see that there is no significant relation to the Loan_Status variable. ●

> Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. ● Logistic regression is an estimation of Logit function. Logit function is simply a log of odds in favor of the event. ● This function creates a s -shaped curve with the probability estimate, which is very similar to the required step wise function 2

●

> Logistic Regression takes only numeric values as input 2

For our problem statement,

> we will make use of dummy variables for the categorical variables. For example, “Gender” variable. 2

It has two classes, Male and Female.

> Once we apply dummies to this variable, it will convert the “Gender” variable into two variables (Gender_Male and Gender_Female), one for each class, i.e. Male and Female. Gender_Male will have a value of 0 if the gender is Female and a value of 1 if the gender is Male 2

> Now, we train the model using the training data set and make predictions using the test 2

data set. Figure 13: Test-Train Curve 1 From the above test-train curve, we could keep a min threshold of 0.4. V. Model Building: As our problem statement is a classification problem, there can be several models put to use. Our team has used 3 models to check which fits the best for our problem statement. These are:

> Logistic Regression, Decision Tree Classifier and Random Forest Classifier. 4

A) Logistic Regression Figure 14: Confusion Matrix 1 Test Accuracy: 0.8617886178861789 Test F1 Score: 0.9081081081081082 Based on the above 2 metrics, we find that logistic regression gave an approximate 86% accuracy. From the above graph, we can conclude that keeping 'Max_Depth' = 3 will yield optimum Test accuracy and F1 score Optimum Test Accuracy ~ 0.805; Optimum F1 Score: ~0.7. B)

> Decision Tree Classifier Decision tree is a type of supervised learning algorithm(having a predefined target variable) that is mostly used in classification problems. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on the most significant splitter / differentiator in input variables. Decision trees use multiple algorithms to decide to split a node in two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable. 2

We use Gini-Index to find the best node to split. We get the following results after training the model and testing the predictions made. Figure 15: Min-Samples leaf plot From the above plot, we can consider the minimum number of leaf samples = 35 to improve test accuracy. Figure 16: Confusion matrix 2 Test Accuracy: 0.8536585365853658 Test F1 Score: 0.903225806451613 Although we got a good test accuracy, from the confusion matrix, we see that there have been misclassifications. Some customers are granted loans but actually, they shouldn't. Figure 14: Max Depth Plot C)

> Random Forest Classifier ● RandomForest is a tree based bootstrapping algorithm wherein a certain no. of weak learners (decision trees) are combined to make a powerful prediction model. ● For every individual learner, a random sample of rows and a few randomly chosen variables are used to build a decision tree model. ● Final prediction can be a function of all the predictions made by the individual learners. ● In case of a regression problem, the final prediction can be mean of all the predictions. 2

● Usually the bigger the forest the better, there is a small chance of overfitting here. The more estimators you give it, the better it will do. takes a lot of time, we have automated

> the loan eligibility process (real time) based on customer information. 1

There are many more attributes which can be taken into account, but we have considered the most important and widely used attributes. All banks try their best to segregate customers based on whether they can repay the loan or not, but there are still many cases where the customers do not repay the loan but satisfy all requirements for the banks. VIII. References 1) https

> ://www.iosrjournals.org/iosr-jce/papers/conf.15013/ V olume 3 /4.% 2018-21 .pdf? 3

id=7557 2) https://medium.com/@vishnumbaprof/case-study-loan-pre diction-ac035f3ec9e4 3) http://cloudstechnologies.in/cloudtech-admin/basepaperfile s/1593149297loan approval prediction using%20decision tree in python.pdf Figure 17: Confusion matrix 3 Test Accuracy: 0.8536585365853658 Test F1 Score: 0.903225806451613 Here, we almost got the same accuracy as that of the decision tree classifier. Test Accuracy: 0.8536585365853658 Test F1 Score: 0.903225806451613 Confusion Matrix on Test Data VI. Experimental Results Looking at the accuracy values and the least number of misclassifications, we have used logistic regression as our proposed model for prediction. But we must make not of some of the assumptions made by the model : 1) The third category of gender is not included. 2)

The probability of clearing the loan amount would be higher if the applicant is a 1

graduate. 3)

If co applicant income is higher , the probability of being eligible would be higher. 4) If Loan amount is higher , the 1

probability of repaying would be lesser and vice 1

versa. VII. Conclusion Hence, we conclude that

the loan eligibility process is done based on 1

the accuracy level of different models. As doing this manually