International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 01 | Jan 2019 www.irjet.net p-ISSN: 2395-0072

Pay Later Loan Prediction for Online Shopping Customers

Dhiren Kumar Dalai[1], N. Venkatesh[2], Kalyani Nahak[3]

[1]*Data Analyst*
[2]*Assistant Professor*
[3]*System engineer*

Abstract – Now a day's online shopping requirement, popularization of consumer loans and the intense competition in financial market has increased the awareness of the critical delinquency issue for financial institutions in granting loans to potential applicants. In the past few decades, the scheme of artificial neural networks has been successfully applied to the financial field. Recently, the Support Vector Machine (SVM) has emerged as the better neural network in dealing with classification and forecasting problems due to its superior features of generalization performance and global optimum. This study develops a loan evaluation model using SVM to identify potential applicants for consumer loans. In addition to conducting experiments on performance comparison via cross-validation and paired *t* test, we analyze misclassification errors in terms of Type I and Type II and their effect on selecting network parameters of SVM. The analysis findings facilitate the development of a useful visual decision-support tool. The experimental results using a real-world data set reveal that SVM surpasses traditional neural network models in generalization performance and visualization via the visual tool, which helps decision makers, determine appropriate loan evaluation strategies.

Key Words: SVM, CNN, Type I, Type II, confusion matrix.

1.INTRODUCTION Pay Later is a proposition extended by select sellers to certain select customers, wherein the said sellers are extending the option to pay for their orders at a date later than purchase date, subject to a collective monthly purchase limit .but question is to whom the seller will provide the option. To identify the customer we used the loan prediction methods [1] by using machine learning algorithms like SVM, linear regression logistic regression we calculated the accuracy for prediction.

1.1 Attribute in the Data Set

Loan id, Gender, Married, Dependents, Education, Self Employed, Applicant income, Co applicant income, Loan Amount, Credit History, Locality Area, Bank Loan Status, frequently Orders.

These are the data set collected by primary and secondary data collection, for which we can predict the future by using machine learning algorithms. by training the the data set of historical data the model is ready to predict the new data point as our model had seen all the type of customer data.

1.2 Major observation from the Data

1. Applicants who are male and married tends to have more applicant income whereas applicant who are female and married have least applicant income
2. Applicants who are male and are graduated have more applicant income over the applicants who have not graduated.
3. Again the applicants who are married and graduated have the more applicant income.
4. Applicants who are not self employed have more applicant income than the applicants who are self employed.
5. Applicants who have more dependents have least applicant income whereas applicants which have no dependents have maximum applicant income.
6. Applicants who have property in urban and have credit history have maximum applicant income
7. Applicants who are graduate and have credit history have more applicant income.
8. Loan Amount is linearly dependent on Applicant income
9. From heatmaps, applicant income and loan amount are highly positively correlated.
10. Male applicants are more than female applicants.
11. No of applicants who are married are more than no of applicants who are not married.
12. Applicants with no dependents are maximum.
13. Applicants with graduation are more than applicants with no graduation.
14. Location of the customer.
15. How frequently used

2. Data Exploration and Preprocessing

The data set I use contains several tables with plenty of information about the accounts of the bank customers such as loans, transaction records and credit cards. Here, my main purpose is to predict customer behaviors about loan for each account. Thus, the most important table here is table "loan". And after checking the description of all the features, we think "order", "trans" and "card" contain useful info for our purpose. And I also need to use account and disposition to combine them together. Finally, the tables required are highlighted in the following figure.

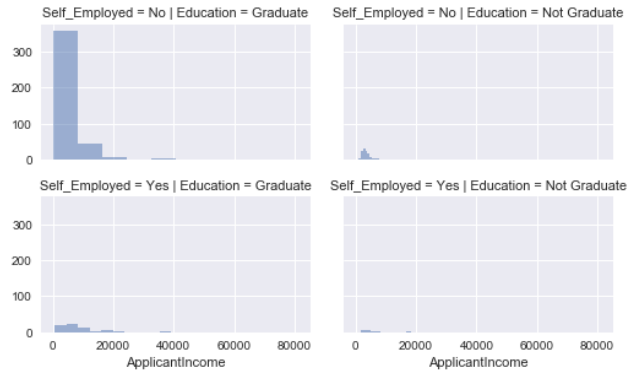
Table -1: data Set format

Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount
Male	No	0	Graduate	No	5849	0.0	NaN
Male	Yes	1	Graduate	No	4583	1508.0	128.0
Male	Yes	0	Graduate	Yes	3000	0.0	66.0
Male	Yes	0	Not Graduate	No	2583	2358.0	120.0
Male	No	0	Graduate	No	6000	0.0	141.0

#Predict test data using pruned decision tree computed.
Applying machine learning models

Using RandomForestClassifier with the parameters
tree_predp1 = predict(ptree1, test_data, type="class")
erp1<-mean(tree_predp1 != test_data$y) # misclassification error
Accupz<- 1-erp1; Accup1
we will improve the accuracy of this model further by using other params
for now we will go to next classification model i.e Random Forest

```
treerf <- randomForest(y ~ age
          +job+marital+education+default+frequently
order+loan,data=train_data,method="class")
treerf
plot(treerf)
legend("topright",
colnames(treerf$err.rate),col=1:4,cex=0.8,fill=1:4)
tree_predrf = predict(treerf, test_data, type="class")
errf <-mean(tree_predrf != test_data$y) # misclassification error
Accurf<- 1-errf; Accurf
```



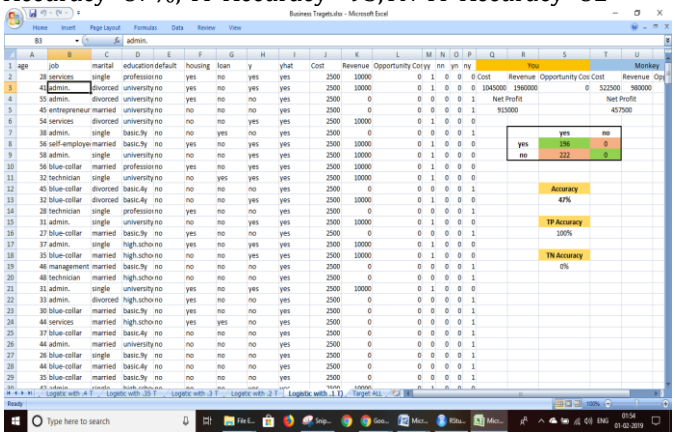
Tuned the values for the better prediction .confusion matrix calculated as

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

	yes	no
yes	137552	59452
no	95765	12752

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \qquad precision = \frac{true\ positives}{true\ positives + false\ positives}$$

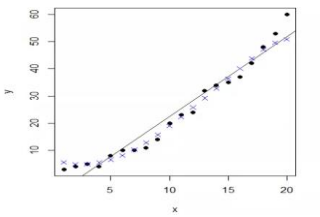
Accuracy=87%, TP Accuracy =93,TN TP Accuracy=82



Support vector machine:

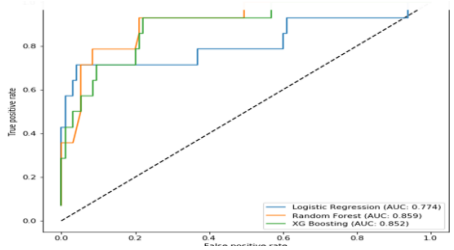
A support vector machine (SVM) is machine learning algorithm that analyzes data for classification and regression analysis. SVM is a supervised learning method that looks at data and sorts it into one of two categories. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible.

```
#SVM
library(e1071)
#Fit a model. The function syntax is very similar to lm function
model_svm <- svm(y ~ x , train)
#Use the predictions on the data
pred <- predict(model_svm, train)
#Plot the predictions and the plot to see our model fit
points(train$x, pred, col = "blue", pch=4)
```



Model Validation and Selection

Here I use K-Fold cross validation to split the data without holdout part into training data and validation data and then fit the model. Since the problem is a classification problem, I choose logistic regression, random forest and XG boosting. To compare the performance of these three models



3. CONCLUSION

Now a days, most of the online shopping sites, electronic commerce companies are competitive and trying for capitalize the insights product and trying to increase the large volume of the customers .so new things can apply by their payment method that is pay latter . To find out the customers to avail this option we need to use prediction methods by historical data set. To improve the recall of the model, we can use the probabilities predicted by the model and set threshold by ourselves. The threshold is set based on several factors such as business objectives. It is different case by case. In the pay later behavior prediction can be done, in this way, companies can detect the default behaviors in the earlier stage and conduct the corresponding actions to reduce the possible loss.

REFERENCES

[1]"The evaluation of consumer loans using support vector machines", panelSheng-TunLiaWeissorShiuebMeng-HuahHuangc, Volume 30, Issue 4, May 2006, Pages 772-782.

[2]" Bank Loan Default Prediction with Machine Learning" Hongri Jia, Apr 10, 2018.

[3] "loan predictor" Architectshwet,ML project-2018.

[4]" Predicting Loan Repayment", Imad Dabbura data scientist march 2018.

BIOGRAPHIES

Working As Data Analyst ,He studied M.E at Anna University, UGC NET Qualified and His research Interests are Network Security, and Data Science. He also Certified in Data science.

N.Venkatesh working as Assistant professor in MRIT, His research Interests are Network Security, and Data Mining.

She is working as System engineer, Her research Interests are It management, BI and Data Science.