Comparative Analysis Of Machine Learning Algorithm on Three Datasets

*Project Of Data Mining and Machine Learning

Prerna Vikas Mhatre
Msc. In Data Analytics
School of computing
National College of Ireland
Dublin, Ireland
x19235551@student.ncirl.ie

***Abstract*—Nowadays, an large amount of data is very frequently generated and available everywhere. It is important to analyze them and extract constructive information in order to promote better data-driven decision making and improve processes and workflow. Machine learning and data mining can be used to accomplish this. Machine learning is a component of artificial intelligence that allows a system to learn without being given explicit instructions. The aim of this paper is to provide a thorough analysis by analyzing various machine learning models and applying them to a variety of applications. There is lot of data generated in the banking sector which is use to analyse and take better decision in future. This research is related to banking sector on three different dataset which are finance banking prediction, bank marketing dataset, and the last is Default of credit card client dataset. On the basis of this research it will be very helpful to predict or take decision in some of the aspect of bank.**

Keywords—Regression algorithms, Decision Tree , Random forest, Multi-linear Regression.

I. Introduction

There is increasing interest in using machine learning (ML) methods to answer a variety of questions that are typically difficult to answer. For three separate datasets, this article aims to illustrate the benefits of using multiple linear regression, logistic regression, k nearest neighbors, decision trees, random forest, and k-means clustering.

Nowadays banking sector is growing tremendously, as everything is going online through internet which saves lots of money and even the man power. As everything is done online in the banking sector like from account opening to carrying any transaction from any place. So lots of efforts of going to bank is saved. But as everything is moving to online it is also generating data in large quantity. It's as if data is produced every hour.

Since this is where the data mining and machine learning techniques come in picture. Like others, in banking sector also the machine learning is the per-requisite feature of an intelligent system. It's the branch of AI that allows machines to start learning how to execute or complete tasks on their own. As bank are continuously producing large data, this data has large number of records and variables available, as well as its a challenging data. So, this is the reason I choose this domain for my project or research.

A) Finance Banking Prediction

This dataset contains almost 19 columns and 45000 rows. This dataset contains all the data related to the loan. All the fields which are considered or taken into consideration while the loan amount is finalized by the bank. There are some criteria's the which the customer should fulfill while taking the loan. Such as annual income, monthly debt, years of credit history, bankruptcies, credit score, current loan amount, etc. This dataset is in csv format and it is taken from the Kaggle website. Research Question for this dataset is as follows:

a) To predict the credit score of the customer using machine learning technique

The target variable or target value is the credit_score. Analysis is done like on which columns the credit score id dependent. The target variable is also know as dependent variable as it is dependent on predictive variables.

B) Bank Marketing Dataset

Bank Marketing dataset is collected from the UCI website. In this dataset there are 45000 records and 17 columns. Variable 'Y' would be the target variable or the dependent variable in this dataset. This dataset contains information or data related to marketing campaign of a financial institution. This data is used to improve future scope of marketing campaign for bank. As nowadays marketing is very much necessary in every field to grow. And banking sector is also growing rapidly. This dataset contains the variables such as job, marital status, age , education, balance, deposit, housing, etc. Research question for this dataset is as follow:

a) The classification goal of this dataset is to predict whether the client will subscribe term deposit or not?

C) Default of credit card client dataset

Nowadays everyone are using the credit cards and most of the clients don't pay the bill amount. This dataset is especially is for predicting which client will make the payment next month for their last six months data or record present in the dataset.This is the classification data set which contains 30000 records and 24 columns. This is all about the credit card and the prediction is also done related to credit card only . This dataset hold the variable like Education, marital status, age, balance limit, repayment status of six months, amount of bill statement of six months, amount of bill payment of six months, default payment, etc. All the payment values are in US_dollar. The default payment is the dependent variable which has the values in yes or no. Pay_1 variable contains the integer values such as -1 for pay duly, 1for payment delay for one month, 2 for payment delay for 2 months with maximum of value 9 for payment delay for 9 months.

a) To predict whether the client will do the payment of credit card or not?

II. Related Work

This section would go through any research articles, any journals, or other references that were used to help address the questions raised in sections.

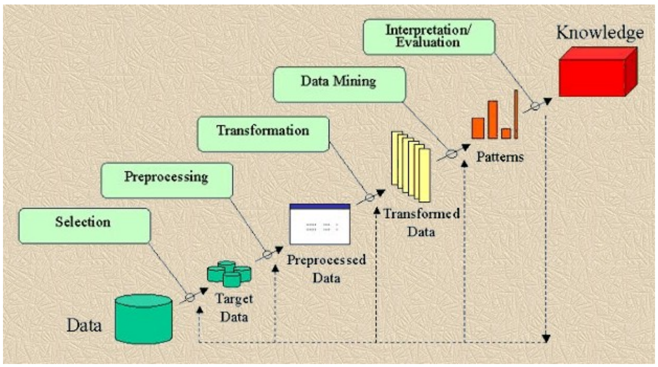
In the paper,[1] Predicting bank insolvencies using machine learning technique. In this paper it is analyzed that the foundation of supervisory authorities support for informed and timely decision making has always been proactive monitoring and assessment of financial institutions economic health. Supervisory authorities use a variety of statistical tools, as well as expert judgement, to assess the riskness of banks and forecast future bank insolvencies. In the research called “Predicting the success of bank telemarketing using various classification algorithm”[2], To minimize dimensionality, the author uses feature selection best subset Logistic Regression, Least Absolute Shrinkage and Selection Operator (LASSO), and Random Forest, in addition to the methods. The RF with the reduced subset of variables once again proved to be the most accurate. The paper "Handling Class Imbalance In Direct Marketing Dataset Us" took a similar approach. A reseach paper of “ A data-driven approach to predict the success of Bank Telemarketing” [3], Portuguese banks, in particular, were under pressure to raise capital requirements (e.g., by capturing more long term deposits). In this sense, a decision support system (DSS) based on a data-driven model to predict the outcome of a telemarketing phone call to sell long-term deposits is a useful tool to aid bank campaign managers' client selection decisions. The paper “Assessing bank efficiency and performance with operational research and artificial intelligence techniques”[4], studies attempt to create classification models that can forecast loss, bank credit scores, and under performers. They find studies and implementations of techniques from different fields that are both country-specific and cross-country. The majority of the studies depend heavily on financial data, though non-financial variables are also used in some cases. Comparisons of classification accuracies through studies should be approached with caution due to the variations in the methods used to validate the models. Only a few studies they found suggest combining individual model predictions into integrated meta-classifiers, and we believe this is a field of study that deserves more focus. In “Bank loan loss provisions research”[5], For four key factors, LLP (Loan Loss Provisions) research continues to be a productive field of banking research. One, LLP is a sizable discretionary accrual available to bank executives. Two, LLP has a direct effect on bank interest margins, which in turn has an impact on bank earnings overall. Three, LLP is related to bank regulators' micro-prudential oversight and the usefulness of accounting disclosures in financial report required by accounting standard-settlers.

In the paper[6], there have also been studies that use machine learning algorithms to forecast house and property prices. The authors of use decision trees to estimate the price of real estate. The Singapore public housing market was used as the dataset. Since they did not want to make certain assumptions in the model (e.g., assuming a linear model would be best), the authors chose to use a tree-based approach instead of a traditional regression approach. The training and validation sets had R2 values of 0.887 and 0.885, respectively. This paper demonstrated the importance of not making any underlying assumptions about the existence of the data (e.g., that the data is linear), as well as the importance of experimenting with various machine learning approaches to achieve optimal prediction success.[7] Another research project combines the study of work and wage forecasting. The author attempted to combine the effects of wage and job forecasting. For prediction, random forest, C4.5 Navie Bayes, and AODE algorithms were used, as well as bagging and boosting methods to improve model accuracy.[8]Multi-linear model-based controllers were studied and compared to normal and scheduled PI controllers. The multi-linear controllers have excellent closed-loop behavior for nonlinear processes like pH neutralization in a stirred tank. When a robust H controller is used, the best results are obtained. This is to be anticipated, given that the controller integrates not only noise but also vibration. This is to be anticipated, given that the controller includes noise filtering for smooth actuator response as well as robustness to account for inherent model-plant mismatch.

Tn this research[16], to search for gross outliers, initial PLS models were created for each analyte. Two T vs. U-score plots, both displaying the relationship between X and Y, both revealed one outlier, hence total no. of samples were reduced. Other traditional outlier diagnostics were also investigated, but no additional gross outliers were identified. By close inspection of the experimental setup and the two spectra in question, it was clear that the two samples were prepared wrongly. For the first outlier, the amount of catechol in the sample was less than it should be according to the experimental setup. The hydroquinone concentration was higher than expected in the second outlier. So, not only do these outliers make sense statistically, but it's also possible to go back and explore the root cause of the anomalous behavior.

III. Methodology

I have used the KDD methodology for data mining. Since this is an iterative multi-stage process, Knowledge Discovery in Database (KDD) is used for data analysis, database integration, and knowledge gathering.



A) Data Selection

I have collected all three dataset in Banking domain. Finance banking prediction, bank marketing dataset and the last is default of creditcard client dataset.

Finance banking dataset has downloaded from kaggle.It contains the information related to the loan in bank. The second dataset is marketing dataset carries all the bank details. Third dataset is default credit card client dataset which carries all the data related to the credit card payments and the client’s data.

Identify applicable funding agency here. If none, delete this text

B) *Data Pre-processing*

After data selection, immediately second step is data preprocessing. This pre-processing is nothing but cleaning our data.In preprocessing there are various steps like searching for null values and if there are null values then fixing that null values. Null values can be dropped so that our dataset will be clean and have no null values. Second option for null values is fixing with mean value of that specific column. In short we are assuming aprox value from the rest of the values. Then comes the finding the outliers and removing the outliers. Outliers means the values which are too big or too small and which is not fitting in the model . So we can remove that value from our data to fit.

C) *Transformation*

In transformation, Data is being transformed. After the dataset is pre-processed , I had splitted my dataset into two parts . Dependent variables and the independent variables. Moving the target variable to the dependent variable and keeping only independent variable in the dataset. And then splitting the dataset in training and test dataset.

D) *Data Mining*

After the dataset is being splitted into train and test dataset, the models are applied on the dataset for analyzing or predicting the result.

E) *Interpretation*

In this step, interpretation is done on the model. It is checked whether the model is best fit or not. This can be checked by various methods such as R squared methos, mean squared error,etc.

IV. Evaluation

A) *Bank Marketing Dataset*

a) *Logistic Regression*

This model can be used in machine learning to solve classification problems if the target variable is binary. This model provides probabilities, and new samples can be identified using discrete and continuous measurements. A simple linear regression model will not work for binary outcomes, so we will use the logistic function and feed our linear regression equation into logistic function. If we put some value into the sigmoid function now, it will produce output in the form of 0 and 1.

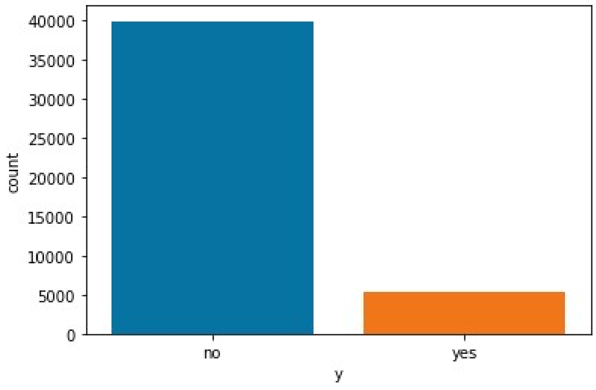


Fig. 1. Analysis of subscribed variable (target variable)

Initially after the dataset is splitted , I checked or analzed for freqency of subscribtion or target variable in the dataset. Fig.1. clearly shows there is the high frequency of ‘No’ for the subscribtion.

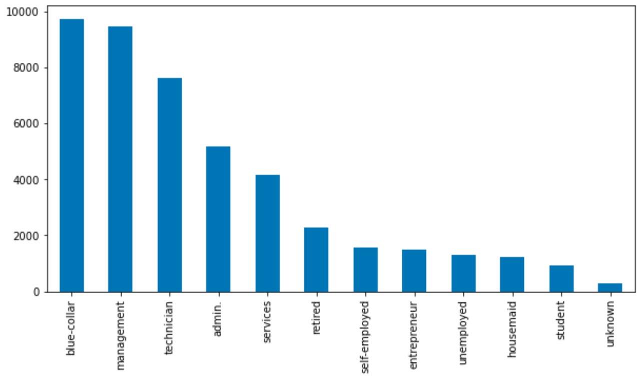


Fig. 2. Analysis of job variable

We can see in fig.2. most of the clients belong to blue-collar job and students are least, as they don't make term deposits or subscribtion.

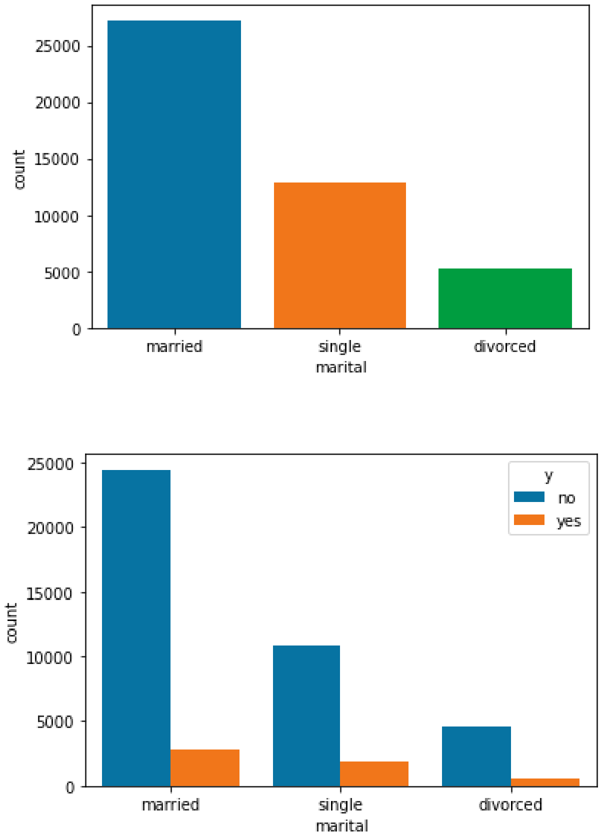


Fig. 3. Analysis of marital status vs subscribed variable (target variable)

In fig.3. we can see that clients who are married are the highlight number to make term deposit then the divorced are the least to make yes as compared to married.

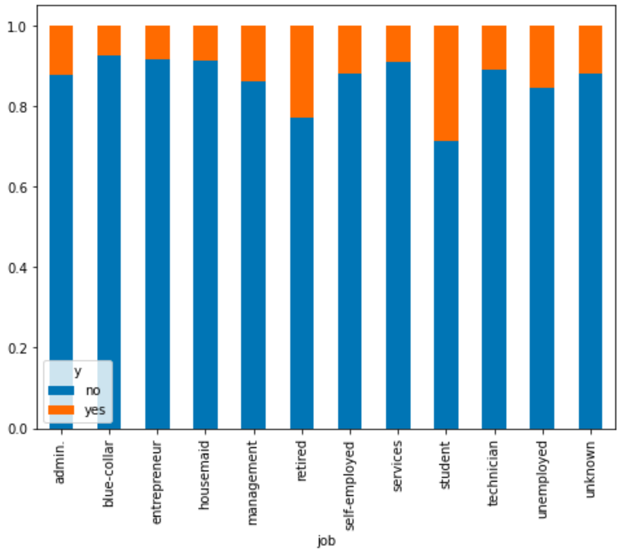
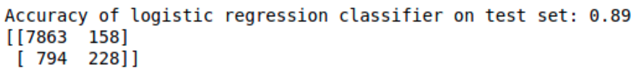


Fig. 4. Analysis of job against target variable



```
Accuracy of logistic regression classifier on test set: 0.89
[[7863  158]
 [ 794  228]]
```

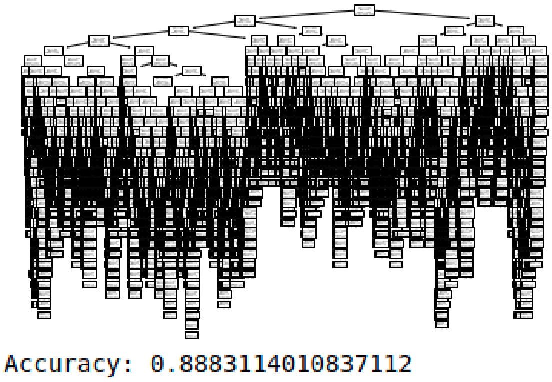
Fig. 5. Accuracy of logistic regression on Bank Marketing dataset

After doing the individual analysis on the columns, then I build the model. Applied the Logistic Regression model. In fig. 5 shows the accuracy of the logistic regression in Bank Marketing dataset. Achieved the accuracy of 0.89 .

b) Decision Tree

In general, we start at the root and work our way down the tree, dividing into branch nodes until we reach the leaf node. The same pattern is followed by the decision tree algorithm. This algorithm can be used to solve problems in both regression and classification. This algorithm can be used to solve problems in both regression and classification. Regression trees use repeated binary division to build a larger tree on the training data set, continuing until each terminal node has the fewest number of nodes than the least number of observations, and finally, it generates a numeric answer with the expected value. Decision tree is easy to interpret.

Applied decision tree on the Bank Marketing dataset using sklearn library in python. Got the accuracy of 0.84.



```
Accuracy: 0.8883114010837112
```

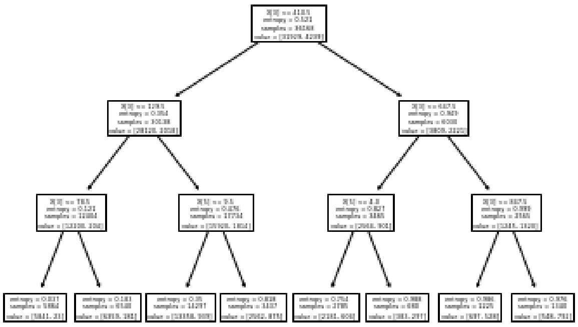


Fig. 6. visualization of decision with dept mentioned

In the above visualization, the depth of tree is not being defined. So it is too messy to interpret.

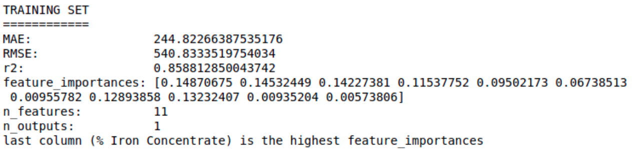
In the fig. 7 , now the dept is defined . So the visualization is able able to interpret and even the accuracy differs as we fluctuate our dept value. For fig.7. the dept of the tree is 3. Accuracy changed from 0.88 to 0.89.

B) Finance Banking Prediction

In this dataset , Initially preprocessing is done. All the null values from the dataset are dropped . Then the dataset is splitted into target variable and the predicted variables. And used train test split method from the sklearn library for splitting the dataset into training and test dataset in the ratio of 80% and 20%. This splitting is done so that we can apply the model separtely on train and test dataset and then copare the result at the end to find whether it is best fit or not.

a) Random Forest Regression

This algorithm is a step up from bagging, and the main difference between the two is the size of the predictor subset m. If the number of correlated predictors is high, using a small m value when constructing a random forest is advantageous. Using m = p, random forest produces unpruned trees and minimizes test and OOB errors. Random forest is used as a classifier as well as the regression methon.In python, I had applied RandomForestRegressor method from the sklearn library.



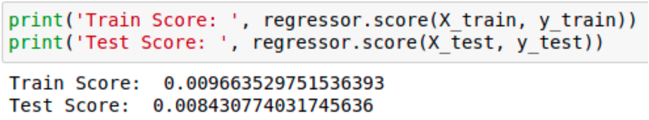
```
TRAINING SET
============
MAE:                     244.82266387535176
RMSE:                    540.8333519754034
r2:                      0.858812850043742
feature_importances: [0.14870675 0.14532449 0.14227381 0.11537752 0.09502173 0.06738513
 0.00955782 0.12893858 0.13232407 0.00935204 0.00573806]
n_features:              11
n_outputs:               1
last column (% Iron Concentrate) is the highest feature_importances
```

Fig. 7. Output of Random Forest

After applying the random forest I removed the r square to find what is the score of s square. And I achieved r square value of 0.85 which is pretty good value to tell that it fits the model.

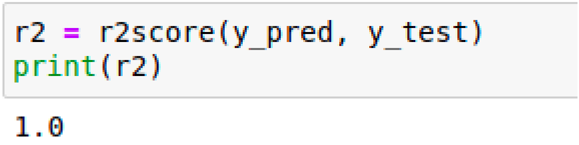
b) Multi-Linear Regression

Multi-Linear Regression is the simplest machine learning algorithm. It can only applied on the regression dataset. There are two or more independent variables in the multilinear regression used for pedicting the output. After preprocessing and the splitting of dataset into train I test dataset , I applied multi linear model on the dataset.



```
print('Train Score: ', regressor.score(X_train, y_train))
print('Test Score: ', regressor.score(X_test, y_test))

Train Score:  0.009663529751536393
Test Score:  0.008430774031745636
```

Fig. 8. Output of multi-linear regression



```
r2 = r2score(y_pred, y_test)
print(r2)

1.0
```

Fig. 9. R square value of multi-linear regression

As we saw both the random forest and the multi linear regression model on the Finance Banking Prediction. Random forecast fit best then the multi-linear model. R square value for multi-linear regression is 1.0 which shows its not fits best. From this analysis we can clearly understand the concept of best fitting.

C) *Default of Credit Card Client Dataset*

In this dataset, all the dataset is related to the credit card. It contains 24 columns and 30000 records.

a) Logistic Regression

After preprocessing and splitting , did analysis on target variable to see how many will do the next payment of the credit card. This analysis is shown in the below figure.

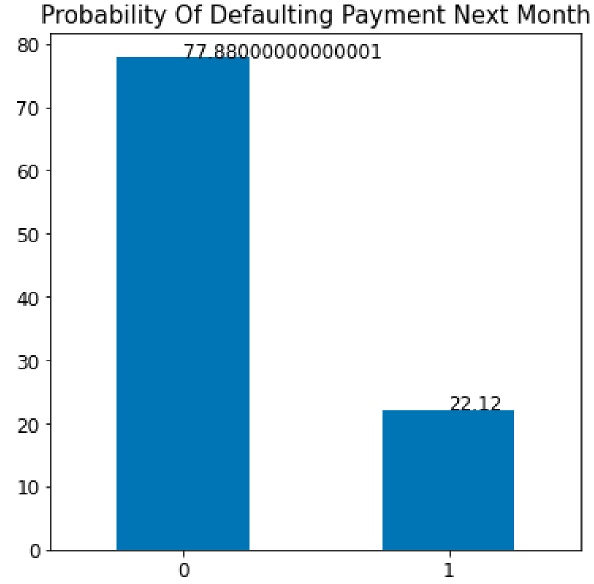
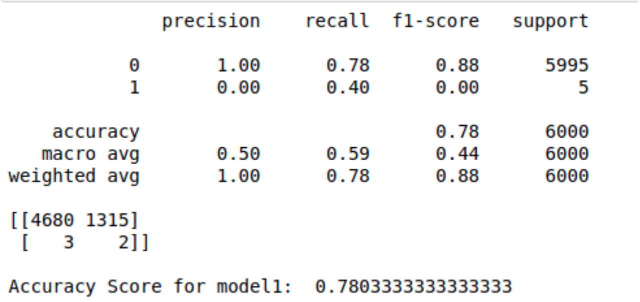


Fig. 10. Analysis of default payment of next month

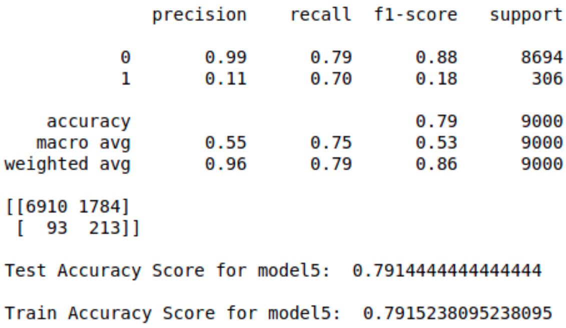
In the above figure, 0 is for No and 1 is for Yes. It clearly shows that almost 78% will not do the next month payment. And only 22 % of the clients will be doing next payment of the credit card.



```
              precision    recall  f1-score   support

           0       1.00      0.78      0.88      5995
           1       0.00      0.40      0.00         5

    accuracy                           0.78      6000
   macro avg       0.50      0.59      0.44      6000
weighted avg       1.00      0.78      0.88      6000

[[4680 1315]
 [   3    2]]

Accuracy Score for model1:  0.7803333333333333
```

Fig. 11. Accuracy score of logistic regression on credit card dataset

Achieved the accuracy score of 0.78 for the logistic regression model on Default credit card client dataset.



```
              precision    recall  f1-score   support

           0       0.99      0.79      0.88      8694
           1       0.11      0.70      0.18       306

    accuracy                           0.79      9000
   macro avg       0.55      0.75      0.53      9000
weighted avg       0.96      0.79      0.86      9000

[[6910 1784]
 [  93  213]]

Test Accuracy Score for model5:  0.7914444444444444

Train Accuracy Score for model5:  0.7915238095238095
```

Fig. 12. Limited independent variable

To improve the fit of model, selected only specific columns for predicting the target variable. After doing this we improved the accuracy score from0.78 to 0.79 which is good for our model.

V. Conclusion

In this paper, we attempted to cover a number of machine learning strategies in datasets that address reality with problems and potential solutions, as well as evaluate their efficacy using measurement criteria that are both practical and adaptable to future projects. We hope that this research will lead to new insights in the areas of consumer propensity and default analysis in the future.

The models would have included a variety of features. This study may have benefited from cross validation, ridge or lasso regression, and principal components analysis.

Different machine learning algorithms may improve the performance of regression models. Deep learning techniques were not used in this study, so there is still room for further research in this field.

VI. References

[1] *https://www.eba.europa.eu/sites/default/documents/files/documents/10180/1813140/aa08cc1c-4bc8-4fda-bae3-2c9214633f78/Session%202%20-%20Predicting%20bank%20insolvencies%20using%20machine%20learning%20techniques.pdf?retry=1*

[2] A. Muneeb, “Predicting the Success of Bank Telemarketing using various Classification Algorithms”, 2018.

[3] S. Moro, P. Cortez, and P. Rita, “A data-driven approach to predict the success of bank telemarketing,” Decision Support Systems, vol. 62, pp. 22–31, 2014.

[4] Marr, N.E. and Prendergast, G.P. (1993), "Consumer Adoption of Self service Technologies in Retail Banking: Is Expert Opinion Supported by Consumer Research?", *International Journal of Bank Marketing*, Vol. 11 No. 1, pp. 3-10.

[5] Borsa Istanbul Review Volume 17, Issue 3, September 2017.

[6] F.Gang-Zhi, S.E. Ong, H.Koh, (2006). "Determinants of House Price: A Decision Tree Approach", Urban Studies, vol 43, No. 12, pp. 2301-2316, Nov. 2006.doi:10.1080/00420980600990928.

[7] M. P. Wijayapala, L. Premaratne, and I. T. Jayamanne, “Employability and related context prediction framework for university graduands: a machine learning approach,” ICTer, vol. 9, no. 2, 2016.

[8] *https://www.eba.europa.eu/sites/default/documents/files/documents/10180/1813140/aa08cc1c-4bc8-4fda-bae3-2c9214633f78/Session%202%20-%20Predicting%20bank%20insolvencies%20using%20machine%20learning%20techniques.pdf?retry=1*

[9] A. Muneeb, “Predicting the Success of Bank Telemarketing using various Classification Algorithms”, 2018.

[10] S. Moro, P. Cortez, and P. Rita, “A data-driven approach to predict the success of bank telemarketing,” Decision Support Systems, vol. 62, pp. 22–31, 2014.

[11] Marr, N.E. and Prendergast, G.P. (1993), "Consumer Adoption of Self service Technologies in Retail Banking: Is Expert Opinion Supported by Consumer Research?", *International Journal of Bank Marketing*, Vol. 11 No. 1, pp. 3-10.

[12] Borsa Istanbul Review Volume 17, Issue 3, September 2017.

[13] F.Gang-Zhi, S.E. Ong, H.Koh, (2006). "Determinants of House Price: A Decision Tree Approach", Urban Studies, vol 43, No. 12, pp. 2301-2316, Nov. 2006.doi:10.1080/00420980600990928.

[14] M. P. Wijayapala, L. Premaratne, and I. T. Jayamanne, “Employability and related context prediction framework for university graduands: a machine learning approach,” ICTer, vol. 9, no. 2, 2016.

[15] *Real-Time Implementation of multi-linear model basel control stratergies—An application toa bench-scalepH neutralizationreactor.*

[16] *Standard error of prediction for multilinear PLS . Practical implementation in fluuorescence spectroscopy.*

Colors and Lines to choose No Fill and No Line.