Predicting Loan Repayment Using [Logistic Regression Algorithm],[Decision Tree Algorithm],[KNN Algorithm],[K means clustering Algorithm]

CSE-0420 Spring 2022

Name:Abu Noman Farhad[ID:UG02-48-18-009] Name: Rukaiya Khan Mithila[ID:UG02-43-16-019] Foysal Khan
Department of Computer Science and Engineering
State University of Bangladesh (SUB)
Dhaka, Bangladesh
email address:ahamedfarhad1@gmail.com, khanrukaiya78@gmail.com

***Abstract*—Python Using Decision Tree Algorithm ,KNN Algorithm,K means clustering Algorithm ,Logistic Regression Algorithm,Random Forest Algorithm Here We will try to find out the Accuracy For All Algorithm**

***Index Terms*—Importing Libraries– Numpy as np, Pandas as pd,Matplotlib.pyplot as plt,Seaborn as sns,LabelEncoder,sklearn,**

	Loan_ID	Gender	Married	Dependen	Education	Self_Empl	Applicant	Coapplica	LoanAmo	Loan_Am	Credit_His	Property_Area
2	LP001015	Male	Yes	0	Graduate	No	5720	0	110	360	1	Urban
3	LP001022	Male	Yes	1	Graduate	No	3076	1500	126	360	1	Urban
4	LP001031	Male	Yes	2	Graduate	No	5000	1800	208	360	1	Urban
5	LP001035	Male	Yes	2	Graduate	No	2340	2546	100	360		Urban
6	LP001051	Male	No	0	Not Gradu	No	3276	0	78	360	1	Urban
7	LP001054	Male	Yes	0	Not Gradu	Yes	2165	3422	152	360	1	Urban
8	LP001055	Female	No	1	Not Gradu	No	2226	0	59	360	1	Semiurban
9	LP001056	Male	Yes	2	Not Gradu	No	3881	0	147	360	0	Rural
10	LP001059	Male	Yes	2	Graduate		13633	0	280	240	1	Urban
11	LP001067	Male	No	0	Not Gradu	No	2400	2400	123	360	1	Semiurban
12	LP001078	Male	No	0	Not Gradu	No	3091	0	90	360	1	Urban

I. INTRODUCTION

The loan is one of the most important products of the financial institutes. All the institutes are trying to figure out effective business strategies to persuade more customers to apply their loans. However, there are some customers are not able to pay off the loan after their application are approved. Therefore, many Financial institutions take several variables into account when approving a loan[Hon18]. Determining whether a given borrower will fully pay off the loan or cause it to be charged off (not fully pay off the loan) is difficult. If the lender is too strict, fewer loans get approved, which means there's less interest to collect.But if they're too lax, they end up approving loans that default . In this study, loan behaviors are analyzed with several machine learning models

II. PROBLEM STATEMENT

Our topic is "Predicting Loan Repayment" if any person take loan from any bank ,she/he will payback the loan or not .Suppose "x" bank gave loan to 12 person ,which is given below in the data-set. here we can apply those Algorithm for this data set [Logistic Regression Algorithm],[Decision Tree Algorithm],[KNN Algorithm],[K means clustering Algorithm]

III. LOGISTIC REGRESSION ALGORITHM

Logistic Regression Algorithm: Here we can using Logistic Regression Algorithm find out the Logistic Regression Accuracy.

What is Logistic Regression? The logistic regression statistic modeling technique is used when we have a binary outcome variable. For example: given the parameters, will the student pass or fail? Will it rain or not? etc.

So, though we may have continuous or categorical independent variables, we can use the logistic regression modeling technique to predict the outcome when the outcome variable is binary.

above is the function for logistic regression:

$$Sig(x) = \frac{1}{1 + e^{-x}}$$

Here we can find out the Logistic Regression accuracy using Logistic Regression algorithm.

```
In [57]: print(confusion_matrix(y_test,lr_prediction))
         print('\n')
         print(classification_report(y_test,lr_prediction))
         print('\n')
         print('Logistic Regression accuracy: ', accuracy_score(y_test,lr_prediction))

                      precision    recall  f1-score   support

                   0       0.60      0.15      0.24        20
                   1       0.00      0.00      0.00        19
                   2       0.38      1.00      0.55        21

            accuracy                           0.40        60
           macro avg       0.33      0.38      0.26        60
        weighted avg       0.33      0.40      0.27        60


        Logistic Regression accuracy:  0.4
        C:\Users\WALTON\anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1248: UndefinedMetricWarning: Precision and F-
```

Logistic Regression accuracy: 0.4

IV. Decision Tree Algorithm

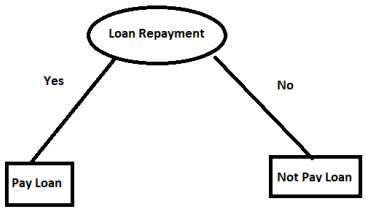
The decision tree Algorithm belongs to the family of supervised machine learning algorithms. It can be used for both a classification problem as well as for regression problem.

The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

Let's take a sample data set to move further ...

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
2	LP001015	Male	Yes	0	Graduate	No	5720	0	110	360	1	Urban
3	LP001022	Male	Yes	1	Graduate	No	3076	1500	126	360	1	Urban
4	LP001031	Male	Yes	2	Graduate	No	5000	1800	208	360	1	Urban
5	LP001035	Male	Yes	2	Graduate	No	2340	2546	100	360		Urban
6	LP001051	Male	No	0	Not Graduate	No	3276	0	78	360	1	Urban
7	LP001054	Male	Yes	0	Not Graduate	Yes	2165	3422	152	360	1	Urban
8	LP001055	Female	No	1	Not Graduate	No	2226	0	59	360	1	Semiurban
9	LP001056	Male	Yes	2	Not Graduate	No	3881	0	147	360	0	Rural
10	LP001059	Male	Yes	2	Graduate		13633	0	280	240	1	Urban
11	LP001067	Male	No	0	Not Graduate	No	2400	2400	123	360	1	Semiurban
12	LP001078	Male	No	0	Not Graduate	No	3091	0	90	360	1	Urban

Suppose we have a sample of 12 person data set and we have to predict which Predicting Loan Repayment to suggest




```
In [61]: print(confusion_matrix(y_test,dt_prediction))
         print('\n')
         print(classification_report(y_test,dt_prediction))
         print('\n')
         print('Decision Tree Accuracy: ', accuracy_score(y_test,dt_prediction))

         [[7 6 7]
          [4 8 7]
          [7 7 7]]


                       precision    recall  f1-score   support

                    0       0.39      0.35      0.37        20
                    1       0.38      0.42      0.40        19
                    2       0.33      0.33      0.33        21

             accuracy                           0.37        60
            macro avg       0.37      0.37      0.37        60
         weighted avg       0.37      0.37      0.37        60


         Decision Tree Accuracy:  0.36666666666666664
```

V. KNN Algorithm

The abbreviation KNN stands for "K-Nearest Neighbour". It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.

Let's take a sample data set to move further ...

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
2	LP001015	Male	Yes	0	Graduate	No	5720	0	110	360	1	Urban
3	LP001022	Male	Yes	1	Graduate	No	3076	1500	126	360	1	Urban
4	LP001031	Male	Yes	2	Graduate	No	5000	1800	208	360	1	Urban
5	LP001035	Male	Yes	2	Graduate	No	2340	2546	100	360		Urban
6	LP001051	Male	No	0	Not Graduate	No	3276	0	78	360	1	Urban
7	LP001054	Male	Yes	0	Not Graduate	Yes	2165	3422	152	360	1	Urban
8	LP001055	Female	No	1	Not Graduate	No	2226	0	59	360	1	Semiurban
9	LP001056	Male	Yes	2	Not Graduate	No	3881	0	147	360	0	Rural
10	LP001059	Male	Yes	2	Graduate		13633	0	280	240	1	Urban
11	LP001067	Male	No	0	Not Graduate	No	2400	2400	123	360	1	Semiurban
12	LP001078	Male	No	0	Not Graduate	No	3091	0	90	360	1	Urban

Suppose we have a sample of 12 person data set and we have to predict which Predicting Loan Repayment to suggest.Here This data set we annalist and find out the

```
In [73]: knn_prediction=knn.predict(X_test)

In [74]: print(confusion_matrix(y_test,knn_prediction))
         print('\n')
         print(classification_report(y_test,knn_prediction))
         print('\n')
         print('KNN accuracy Accuracy: ', accuracy_score(y_test,knn_prediction))

         [[ 1  7 12]
          [ 2  0 17]
          [ 2  2 17]]


                       precision    recall  f1-score   support

                    0       0.20      0.05      0.08        20
                    1       0.00      0.00      0.00        19
                    2       0.37      0.81      0.51        21

             accuracy                           0.30        60
            macro avg       0.19      0.29      0.20        60
         weighted avg       0.20      0.30      0.20        60


         KNN accuracy Accuracy:  0.3
```

.KNN Algorithm Accuracy: 0.3

VI. K means clustering Algorithm

The K-means clustering algorithm computes centrists and repeats until the optimal centred is found. It is preemptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter 'K' in K-means.

In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centred is as small as possible. It is essential to note that reduced diversity within clusters leads to more identical data points within the same cluster.

Let's take a sample data set to move further ...

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
2	LP001015	Male	Yes	0	Graduate	No	5720	0	110	360	1	Urban
3	LP001022	Male	Yes	1	Graduate	No	3076	1500	126	360	1	Urban
4	LP001031	Male	Yes	2	Graduate	No	5000	1800	208	360	1	Urban
5	LP001035	Male	Yes	2	Graduate	No	2340	2546	100	360		Urban
6	LP001051	Male	No	0	Not Graduate	No	3276	0	78	360	1	Urban
7	LP001054	Male	Yes	0	Not Graduate	Yes	2165	3422	152	360	1	Urban
8	LP001055	Female	No	1	Not Graduate	No	2226	0	59	360	1	Semiurban
9	LP001056	Male	Yes	2	Not Graduate	No	3881	0	147	360	0	Rural
10	LP001059	Male	Yes	2	Graduate		13633	0	280	240	1	Urban
11	LP001067	Male	No	0	Not Graduate	No	2400	2400	123	360	1	Semiurban
12	LP001078	Male	No	0	Not Graduate	No	3091	0	90	360	1	Urban

Suppose we have a sample of 12 person data set and we have to predict which Predicting Loan Repayment to suggest.Here This data set we annalist and find out the K means clustering Algorithm

```
In [81]: print(confusion_matrix(y_test,svc_prediction))
         print('\n')
         print(classification_report(y_test,svc_prediction))
         print('\n')
         print('K means clustering File: ', accuracy_score(y_test,svc_prediction))

         [[ 0  0 20]
          [ 0  0 19]
          [ 1  0 20]]


                       precision    recall  f1-score   support

                    0       0.00      0.00      0.00        20
                    1       0.00      0.00      0.00        19
                    2       0.34      0.95      0.50        21

             accuracy                           0.33        60
            macro avg       0.11      0.32      0.17        60
         weighted avg       0.12      0.33      0.17        60


         K means clustering File:  0.3333333333333333
```

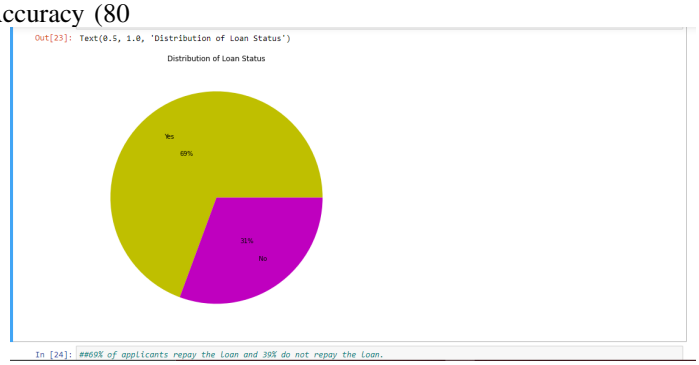
K means clustering File: 0.3333333333333333

VII. Conclusion

By doing all the algorithms we have finally come out with a decisions. Logistic Regression accuracy: 0.4 Decision Tree Algorithm Accuracy : 0.36666666666666664 KNN Algorithm Accuracy: 0.3 K means clustering Algorithm Accuracy : 0.3333333333333333

The Loan Status is heavily dependent on the Credit History for Predictions.

The Logistic Regression algorithm gives us the maximum Accuracy (80



References

Acknowledgment

I would like to thank my honourable **Sabrina Jesmin** for his time, generosity and critical insights into this project.

References

[1] [Alt92]Naomi S Altman. “An introduction to kernel and nearest-neighbor nonparametricregression.

[2] ”The American Statistician,46(3):175–185, 1992.[Asa18] (Sudharsan Asaithambi. “Why, How and When to apply Feature Selection.” Jan2018.[Bad19] Will Badr.

[3] “Why Feature Correlation Matters . . . . A Lot!”Towards Data Science,Jan 18, 2019.[Bha18] Abhishek Bhagat et al.Predicting Loan Defaults using Machine Learning Tech-niques. PhD thesis, California State University, Northridge, 2018.[Bro18]

[4] Jason Brownlee. “A gentle introduction to k-fold cross-validation.”Accessed Oc-tober,7:2018, 2018.[Cha17] Ofir Chakon. “PRACTICAL MACHINE LEARNING: RIDGE REGRESSIONVS.LASSO.” August 2017.[Don18] Niklas Donges.

[5] “The Random Forest Algorithm.”Statistical Methods, Feb 2018.[ELL11] Brian S Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. “Miscellaneousclustering methods.

[6] ”Cluster Analysis, 5th Edition, John Wiley Sons, Ltd,Chichester, UK, 2011.[HCL03] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. “A practical guide tosupport vector classification.” 2003.[Hon18] Hongri. “Jia Bank Loan Default Prediction with Machine Learning.” pp. 137–163,Apr 10, 2018.[KZ01]Gary King and Langche Zeng.

[7] “Logistic regression in rare events data.”Politicalanalysis,9(2):137–163, 2001.[LWZ06] Xu Ying Liu, Jianxin Wu, and Zhi Hua Zhou.

[8] “Exploratory Under-Sampling forClass-Imbalance Learning.” InInternational Conference on Data Mining, 2006.[PLI02] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll.

[9] “An introductionto logistic regression analysis and reporting.”The journal of educational research,96(1):3–14, 2002.[Ras14] Sebastian Raschka.

[10] “About feature scaling and normalization.”Sebastian Racha.Disques, nd Web. Dec, 2014.[Swa18] Alvira Swalin. “How to handle missing value.”Towards Data Science, 2018.