Prediction of Approved Loan Duration of Loan Holders

Working For More Than 10 Years

John Arvee Flores

INTRODUCTION

Banks require an individual to meet several conditions for a loan request to be approved. Loan holders are qualified for their loan requests based on the information provided upon requesting a loan. These information are presented on this dataset which are essential part of this study to help on prediction of the loan duration of approved loan as this might help on future studies relating the processes involved in the field of debt studies.

Research Questions:

The dataset comprises of 1000 observations to help the researcher for this study. Some of the variables were chosen to predict the duration of the approved loan amount by the loan holders.

Here are the questions this research wants to answer:

1. How does approved loan amount affect the prediction of the duration of the loan of loan holders who are currently 10 years in their current job, considering their loan purpose is for debt consolidation?
2. How does loan holder's monthly debt affect the prediction of the duration of loan of loan holders who are currently 10 years in their current job with the loan purpose of debt consolidation?

This study would be beneficial for future studies as this would provide more understanding on how the variables utilized in this study affects the prediction of the duration of the approved loan.

This study is limited only on identifying the relationship of the variables mentioned in the research questions. Moreover, the subset of individuals to be utilized for this study are limited only to individuals who have 10 years on their current job history and have a loan purpose of debt consolidation as these groups are the majority among the dataset in terms of number of years in current job and loan purpose, respectively.

RELATED LITERATURE

Logistics Regression

This study aims to use of logistics regression as part of the methodology to be used in predicting the loan duration of the approved loan in relation to the variables available within the dataset. Logistics regression was chosen by the researcher since this has been one of the most used methods in earlier relevant failure-prediction studies, as well as one of the most practiced methods by banks in their corporate default-prediction modelling (Kohv et al, 2020). [1]

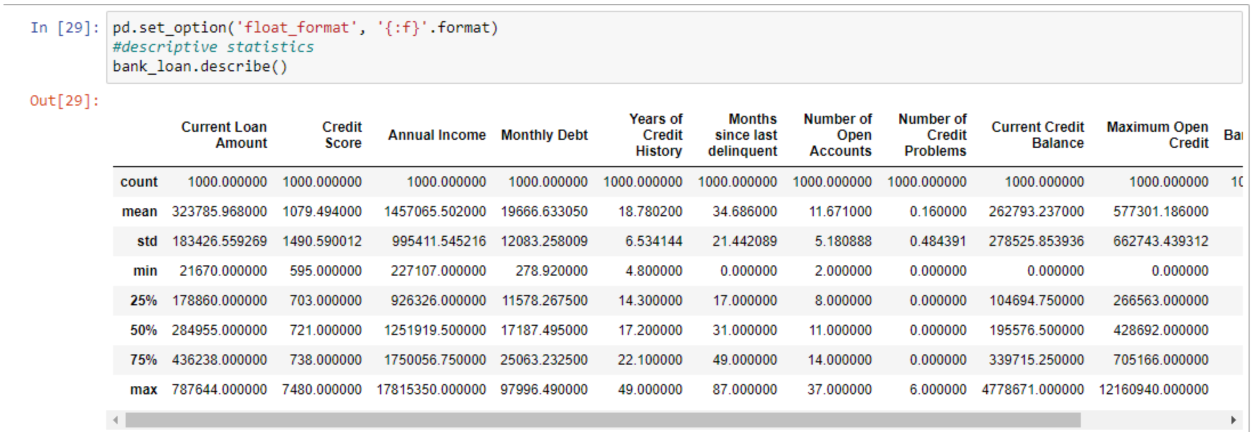
Another study was conducted that proves to have a significant improvement in providing better results when having an example-dependent cost-sensitive logistic regression for credit scoring in comparison to the typical cost-sensitive approach in which logistic regression was used as a main methodology for this study in which this study is related in terms of the application of logistics regression in doing prediction.[2]

EXPLORATORY DATA ANALYSIS

I. Background of Dataset

Before answer the research questions, let the researcher provide details on the dataset describing the descriptive information of the variables and how the researcher come up on the scope of this study covering only individuals with 10 years existence on their current job and have a loan purpose of debt consolidation.

The dataset has 1000 observations and is composed of 18 variables. Below is a descriptive summary statistic for this dataset.

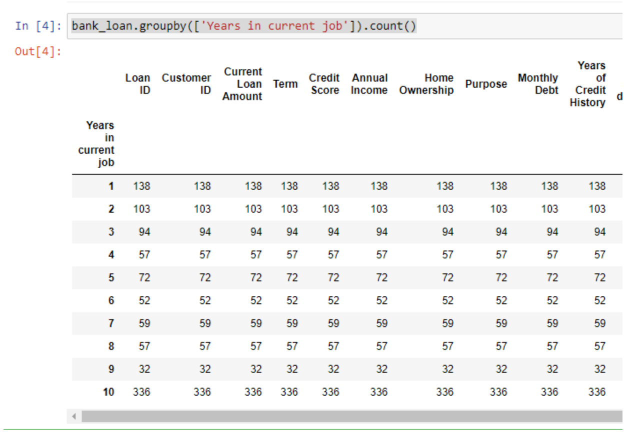


```
In [29]: pd.set_option('float_format', '{:f}'.format)
         #descriptive statistics
         bank_loan.describe()
```

Out[29]:

	Current Loan Amount	Credit Score	Annual Income	Monthly Debt	Years of Credit History	Months since last delinquent	Number of Open Accounts	Number of Credit Problems	Current Credit Balance	Maximum Open Credit	Ba
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1
mean	323785.968000	1079.494000	1457065.502000	19666.633050	18.780200	34.686000	11.671000	0.160000	262793.237000	577301.186000	
std	183426.559269	1490.590012	995411.545216	12083.258009	6.534144	21.442089	5.180888	0.484391	278525.853936	662743.439312	
min	21670.000000	595.000000	227107.000000	278.920000	4.800000	0.000000	2.000000	0.000000	0.000000	0.000000	
25%	178860.000000	703.000000	926326.000000	11578.267500	14.300000	17.000000	8.000000	0.000000	104694.750000	266563.000000	
50%	284955.000000	721.000000	1251919.500000	17187.495000	17.200000	31.000000	11.000000	0.000000	195576.500000	428692.000000	
75%	436238.000000	738.000000	1750056.750000	25063.232500	22.100000	49.000000	14.000000	0.000000	339715.250000	705166.000000	
max	787644.000000	7480.000000	17815350.000000	97996.490000	49.000000	87.000000	37.000000	6.000000	4778671.000000	12160940.000000	

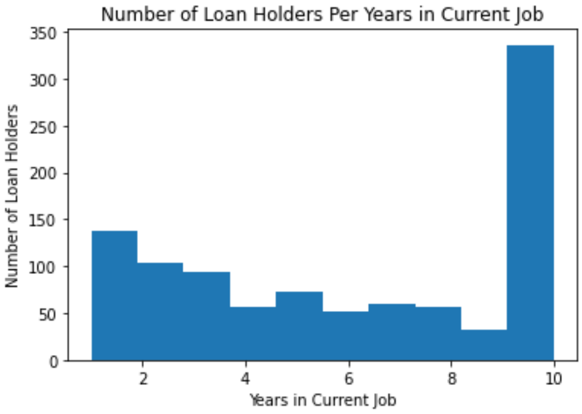
Out of 1000 records in dataset, the lowest loan amount an individual has is 21,670 while the highest loan amount is 787,644. On the other hand, the average annual income of the loan holders is 1,457,065.50 in which 25% of the loan holders have greater annual income than the average.



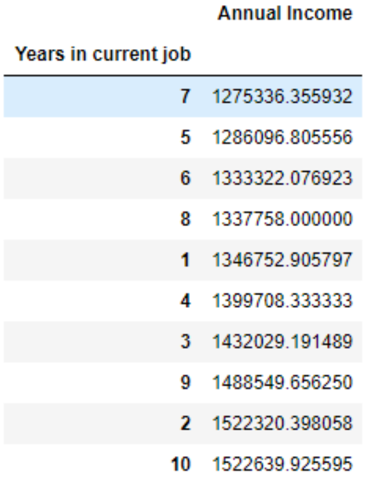
```
In [4]: bank_loan.groupby(['Years in current job']).count()
```

Out[4]:

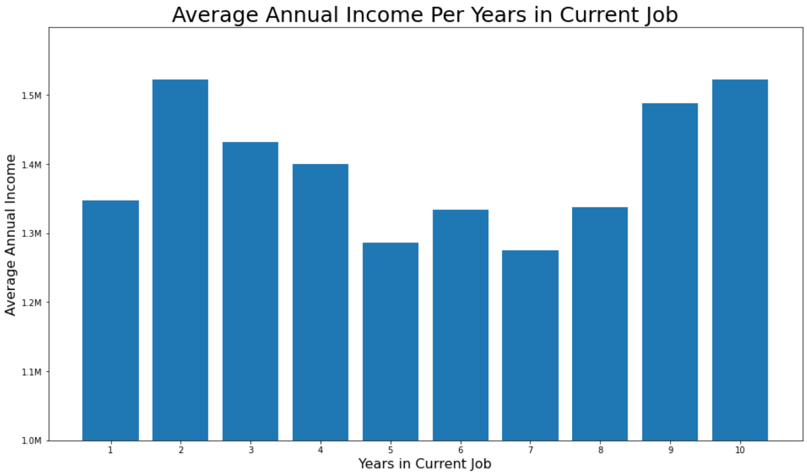
Years in current job	Loan ID	Customer ID	Current Loan Amount	Term	Credit Score	Annual Income	Home Ownership	Purpose	Monthly Debt	Years of Credit History	d
1	138	138	138	138	138	138	138	138	138	138	
2	103	103	103	103	103	103	103	103	103	103	
3	94	94	94	94	94	94	94	94	94	94	
4	57	57	57	57	57	57	57	57	57	57	
5	72	72	72	72	72	72	72	72	72	72	
6	52	52	52	52	52	52	52	52	52	52	
7	59	59	59	59	59	59	59	59	59	59	
8	57	57	57	57	57	57	57	57	57	57	
9	32	32	32	32	32	32	32	32	32	32	
10	336	336	336	336	336	336	336	336	336	336	



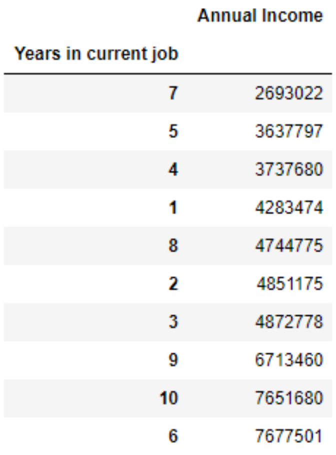
Looking at the number of years in current job, majority of the loan holders are working for at least 10 years in their current job followed by loan holders working for 1 year in their current job.



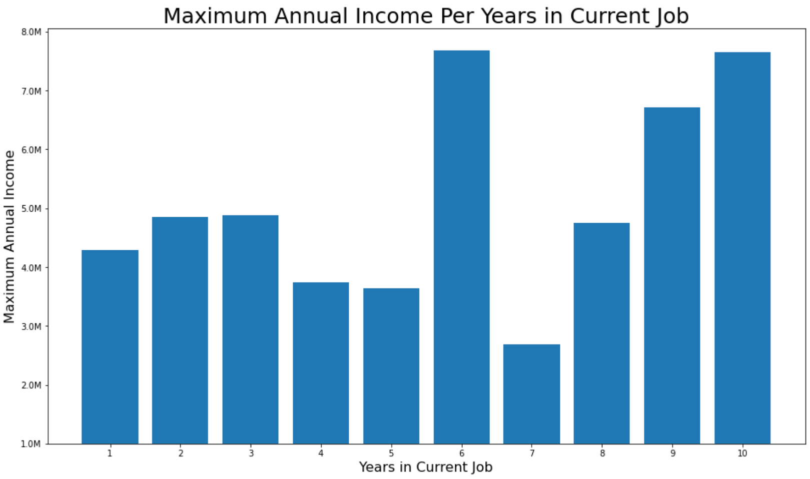
Years in current job	Annual Income
7	1275336.355932
5	1286096.805556
6	1333322.076923
8	1337758.000000
1	1346752.905797
4	1399708.333333
3	1432029.191489
9	1488549.656250
2	1522320.398058
10	1522639.925595



In terms of average annual income, loan holders who are in 10 years in their current job gets the highest average annual income followed by those who are in their 2nd year in their current job.

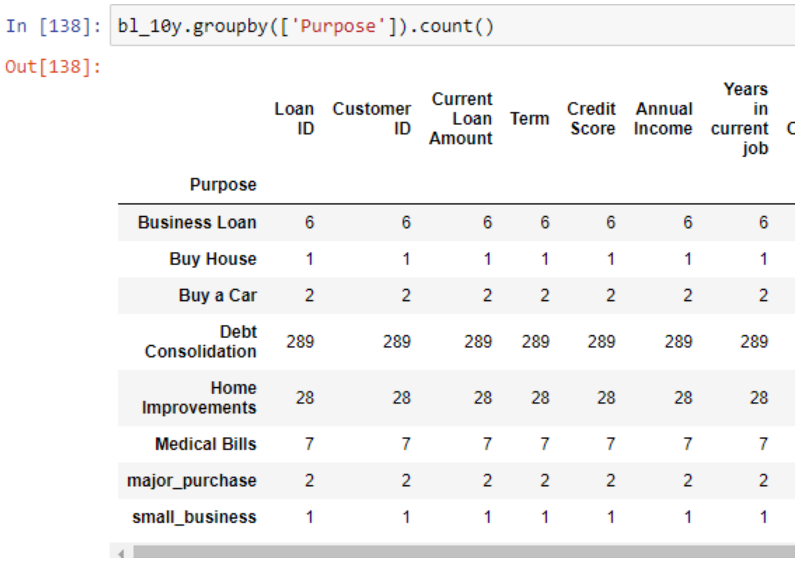


Years in current job	Annual Income
7	2693022
5	3637797
4	3737680
1	4283474
8	4744775
2	4851175
3	4872778
9	6713460
10	7651680
6	7677501



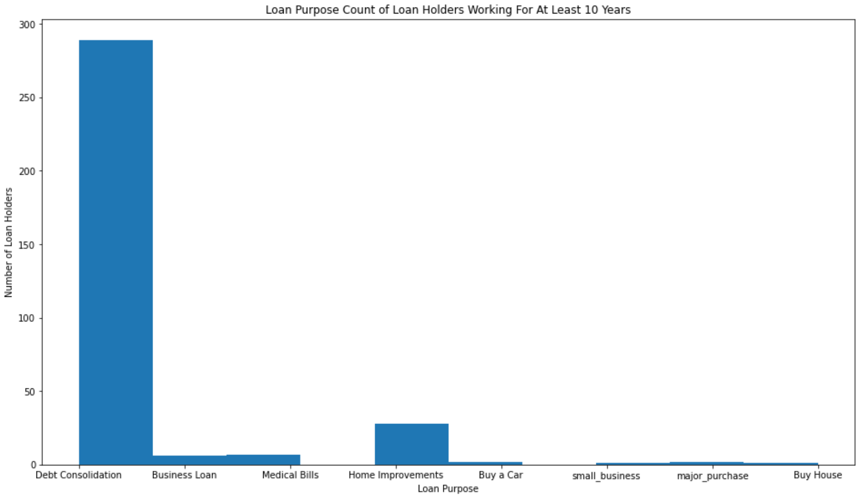
With regards to checking maximum annual income per year in current job, the highest annual income comes from the group who are working for 6 years in their current job followed by those working for 10 years and above.

Taking all these details into consideration, the researcher came up to a conclusion of using the group which majority of the loan holders are included as well as the group which has the highest average annual income and 2nd to the highest group with maximum annual income of loan holders. This group is composed of loan holders working for at least 10 years in their current job which includes 336 individuals.

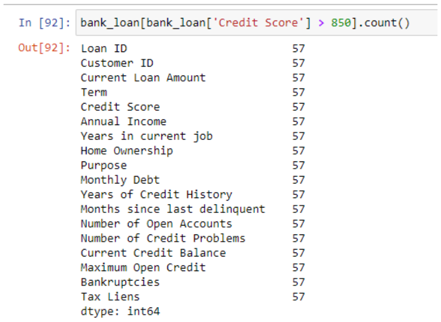


```
In [138]: bl_10y.groupby(['Purpose']).count()
Out[138]:
```

Purpose	Loan ID	Customer ID	Current Loan Amount	Term	Credit Score	Annual Income	Years in current job	C
Business Loan	6	6	6	6	6	6	6	
Buy House	1	1	1	1	1	1	1	
Buy a Car	2	2	2	2	2	2	2	
Debt Consolidation	289	289	289	289	289	289	289	
Home Improvements	28	28	28	28	28	28	28	
Medical Bills	7	7	7	7	7	7	7	
major_purchase	2	2	2	2	2	2	2	
small_business	1	1	1	1	1	1	1	

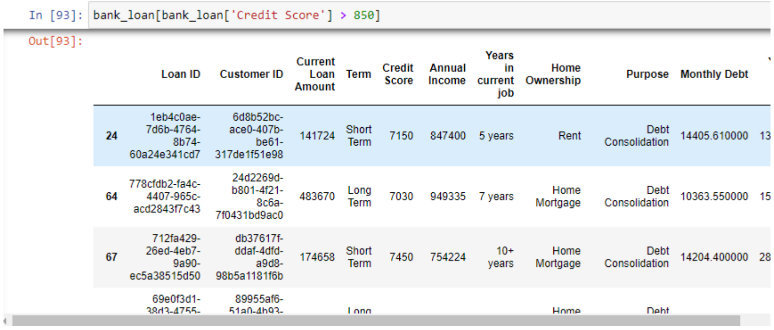


Among these 336 individuals who are working for at least 10 years in their current job, majority of their loan purpose is to have their debt consolidated in which 289 out of 336 have this purpose. As majority are of this purpose, this was the reason why the researcher chose this subgroup to focus on for this study.



```
In [92]: bank_loan[bank_loan['Credit Score'] > 850].count()

Out[92]: Loan ID                         57
         Customer ID                     57
         Current Loan Amount             57
         Term                            57
         Credit Score                    57
         Annual Income                   57
         Years in current job            57
         Home Ownership                  57
         Purpose                         57
         Monthly Debt                    57
         Years of Credit History         57
         Months since last delinquent    57
         Number of Open Accounts         57
         Number of Credit Problems       57
         Current Credit Balance          57
         Maximum Open Credit             57
         Bankruptcies                    57
         Tax Liens                       57
         dtype: int64
```



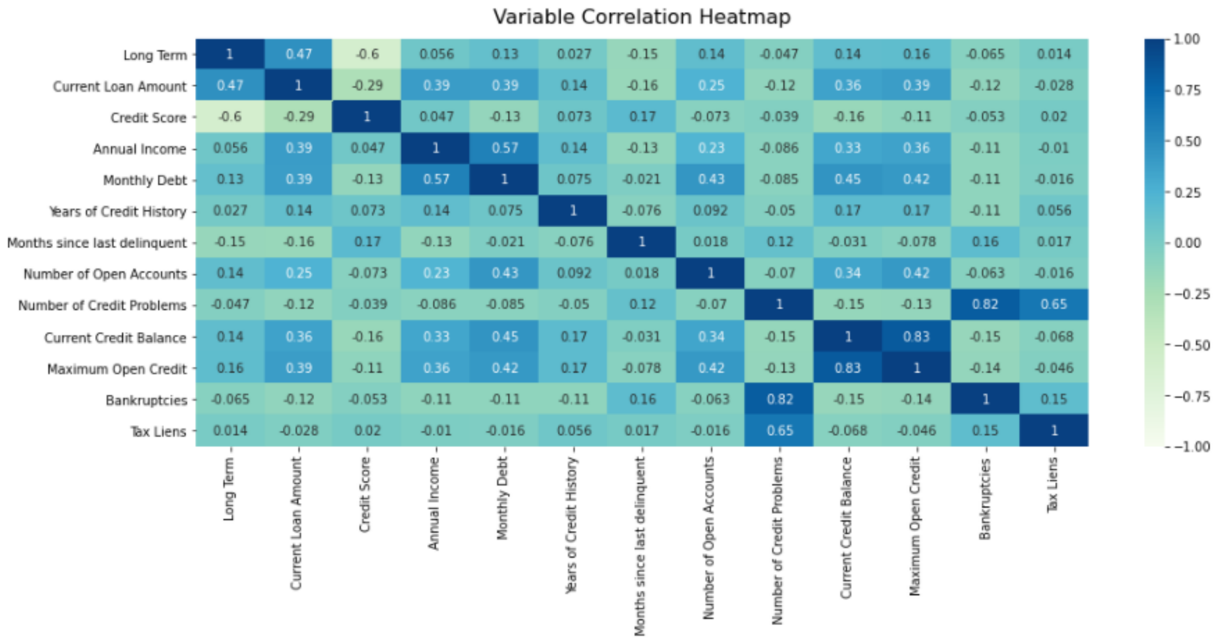
```
In [93]: bank_loan[bank_loan['Credit Score'] > 850]
```

Out[93]:

	Loan ID	Customer ID	Current Loan Amount	Term	Credit Score	Annual Income	Years in current job	Home Ownership	Purpose	Monthly Debt	
24	1eb4c0ae-7d6b-4764-8b74-60a24e341cd7	6d8b52bc-ace0-407b-be61-317de1f51e98	141724	Short Term	7150	847400	5 years	Rent	Debt Consolidation	14405.610000	13
64	778cfdb2-fa4c-4407-965c-acd2843f7c43	24d2269d-b801-4f21-8c6a-7f0431bd9ac0	483670	Long Term	7030	949335	7 years	Home Mortgage	Debt Consolidation	10363.550000	15
67	712fa429-26ed-4eb7-9a90-ec5a38515d50	db37617f-ddaf-4dfd-a9d8-98b5a1181f6b	174658	Short Term	7450	754224	10+ years	Home Mortgage	Debt Consolidation	14204.400000	28
	69e0f3d1-38d3-4755-	89955af6-51a0-4b93-		Long				Home	Debt		

For additional information, the above table shows the outliers in terms of credit score. These were easily identified as the maximum credit score which a person can reach is up to 850 only but 57 out of 1000 observations provided credit score higher than 850.

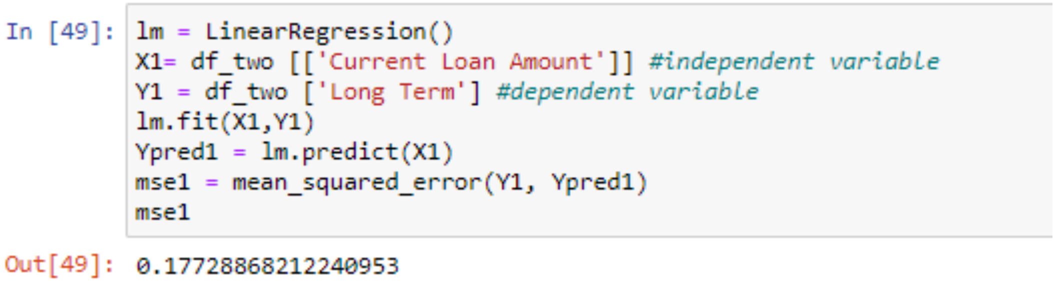
These outliers were not removed upon doing this study but instead, were fixed logically as the researcher only needs to omit the last digit on these observations for these outliers to become accurate.



A heatmap is presented above to easily identify which variables are correlated with each other. From here, the researcher verified that the closest possible variables which can be used in prediction of the duration of the approved load were the current loan amount as well as the monthly debt field. Further discussions will detail more understanding on the prediction outcome of this study.

The mean squared errors were identified for the possible independent variables to be used in order to understand which set of variables would be best when used for the regression analysis. We can see from the below results that using multi-independent variable provided lower MSE compared to using single independent variable. For this case, the author of this study used multi-independent variables since this resulted with a much lower mean squared error as compared to the result of using single independent variable. The independent variables chosen were loan amount and monthly debt while the dependent variable is the loan duration in which it is a categorical data and was translated to a binary data column called ‘Long Term’.

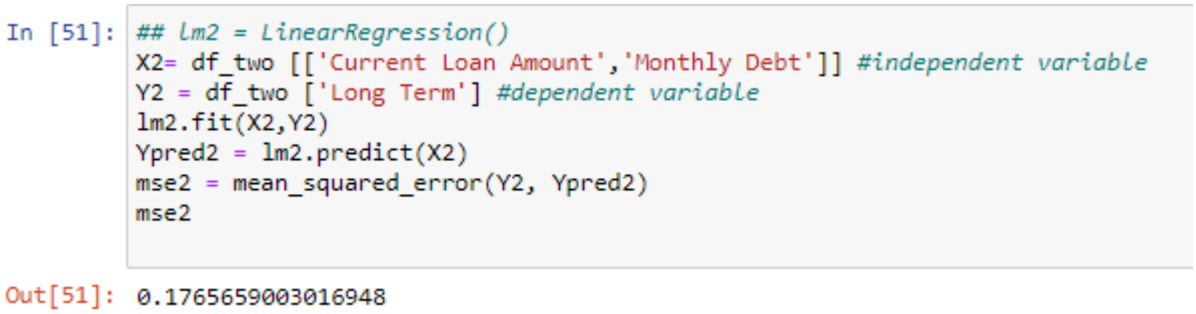
Single Independent Variable



```
In [49]: lm = LinearRegression()
         X1= df_two [['Current Loan Amount']] #independent variable
         Y1 = df_two ['Long Term'] #dependent variable
         lm.fit(X1,Y1)
         Ypred1 = lm.predict(X1)
         mse1 = mean_squared_error(Y1, Ypred1)
         mse1

Out[49]: 0.1772868212240953
```

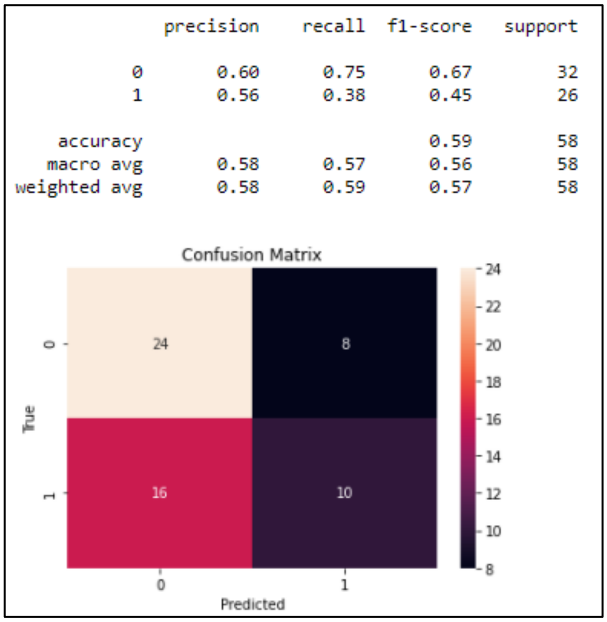
Multiple Independent Variable



```
In [51]: ## lm2 = LinearRegression()
         X2= df_two [['Current Loan Amount','Monthly Debt']] #independent variable
         Y2 = df_two ['Long Term'] #dependent variable
         lm2.fit(X2,Y2)
         Ypred2 = lm2.predict(X2)
         mse2 = mean_squared_error(Y2, Ypred2)
         mse2

Out[51]: 0.1765659003016948
```

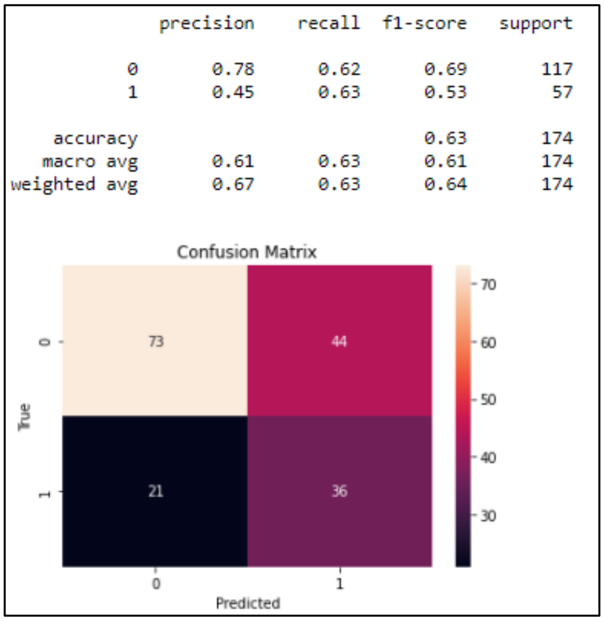
The independent and dependent variables used for linear regression were subjected for logistic regression analysis in answering the research question for this study. Using logistic regression with multi-independent variables, below are the results using different test sizes.



```
              precision    recall  f1-score   support

           0       0.60      0.75      0.67        32
           1       0.56      0.38      0.45        26

    accuracy                           0.59        58
   macro avg       0.58      0.57      0.56        58
weighted avg       0.58      0.59      0.57        58
```



```
              precision    recall  f1-score   support

           0       0.78      0.62      0.69       117
           1       0.45      0.63      0.53        57

    accuracy                           0.63       174
   macro avg       0.61      0.63      0.61       174
weighted avg       0.67      0.63      0.64       174
```

Presented above are tests results using two different test sizes by utilizing the independent variables ‘Current Loan Amount’ and ‘Monthly Debt’ and dependent variable ‘Long Term’ which is a binary data for the loan duration (Long Term or Short Term).

Looking on the left side, this used a test size of 0.2 in which it resulted to an overall accuracy of 59% while on the right side, it used 0.6 as test size having a greater result in overall accuracy of 63%. While testing, there were different sizes tested as well but presented here were the test sizes which garnered the lowest and highest overall accuracy.

For further discussion, precision is a measure of the accuracy provided that a class label has been predicted in which for 0 values, test size 0.6 had better result of 78% compared to 60% but for 1 values, 0.2 test size had more precision of 56% compared to 45%. In terms of recall, there is a better true positive rate on the 0.2 test size for 0 values of 75% matched with 62% of the 0.6 test size. In contrary, recall percentage in 1 values are greater for 0.6 test size with 63% compared to 38% of the 0.2 test size. For the f1 score, the test size 0.6 results had superior results for both 0 and 1 values having 69% and 53% respectively as opposed to the results of 0.2 test size of 67% and 45% for 0 and 1 values respectively.

This comparison of test size results showed clarity that providing greater test size enhances the overall accuracy of the prediction model for logistic regression. This also provides better understanding that using the provided model for prediction, the author of this study was able to prove that there is a 63% accuracy of using the variables ‘Current Loan Amount’ and ‘Monthly Debt’ as independent variables on predicting the duration of the approved loans.

For future related studies, provided below are some of the recommendations by the author upon completing this study:

1. Testing of different variable combinations as independent variables for better accuracy in predicting the dependent variables in your study.
2. Exploring other prediction models available for comparison in getting the best accurate model for your dataset.
3. Knowing your dataset by creating as much visualization as you can for having more knowledge on the dataset being used in the study.

References:

[1] What Best Predicts Corporate Bank Loan Defaults? An Analysis of Three Different Variable Domains. (January 2021). Kohv, K., and Lukason, O. Risks 9: 29. https://doi.org/10.3390/risks9020029

[2] Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring. (2015). Bahnsen, A., Aouada, D., and Ottersten, B.