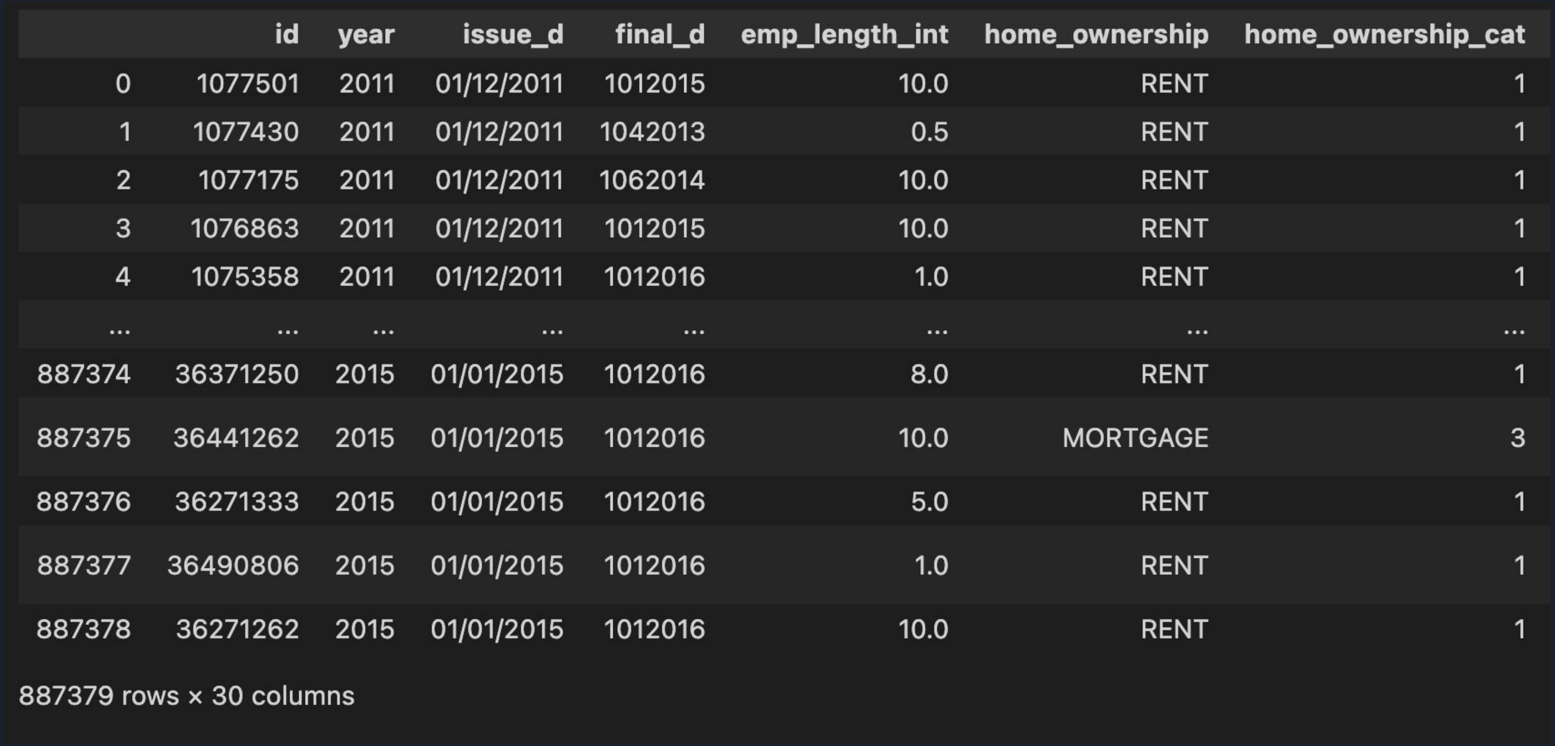
Implementing Logistic Regression For Binary Classification in Bank-loan Dataset

Azis Muslim
Data Scientist

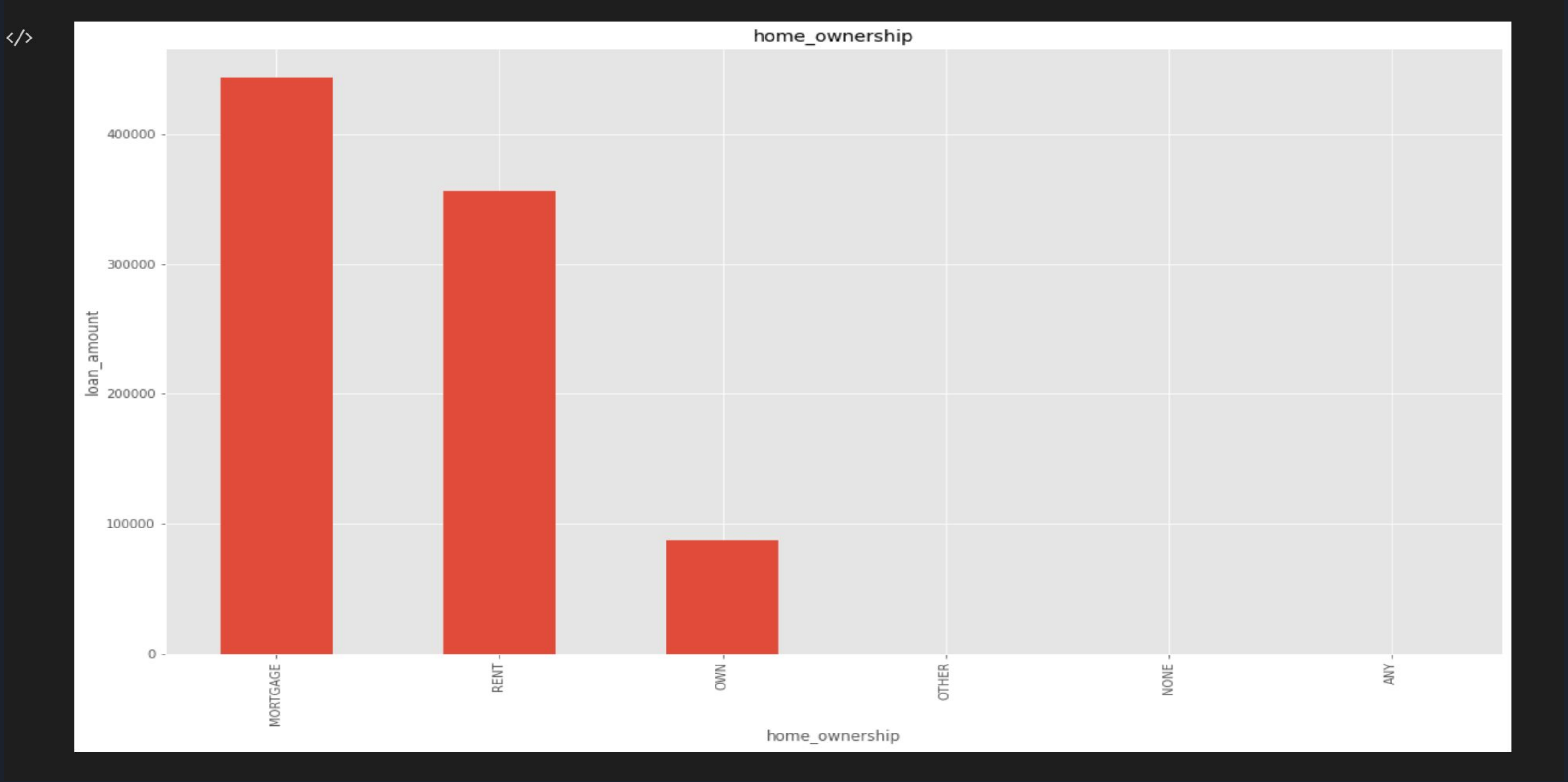
Dataset Overview



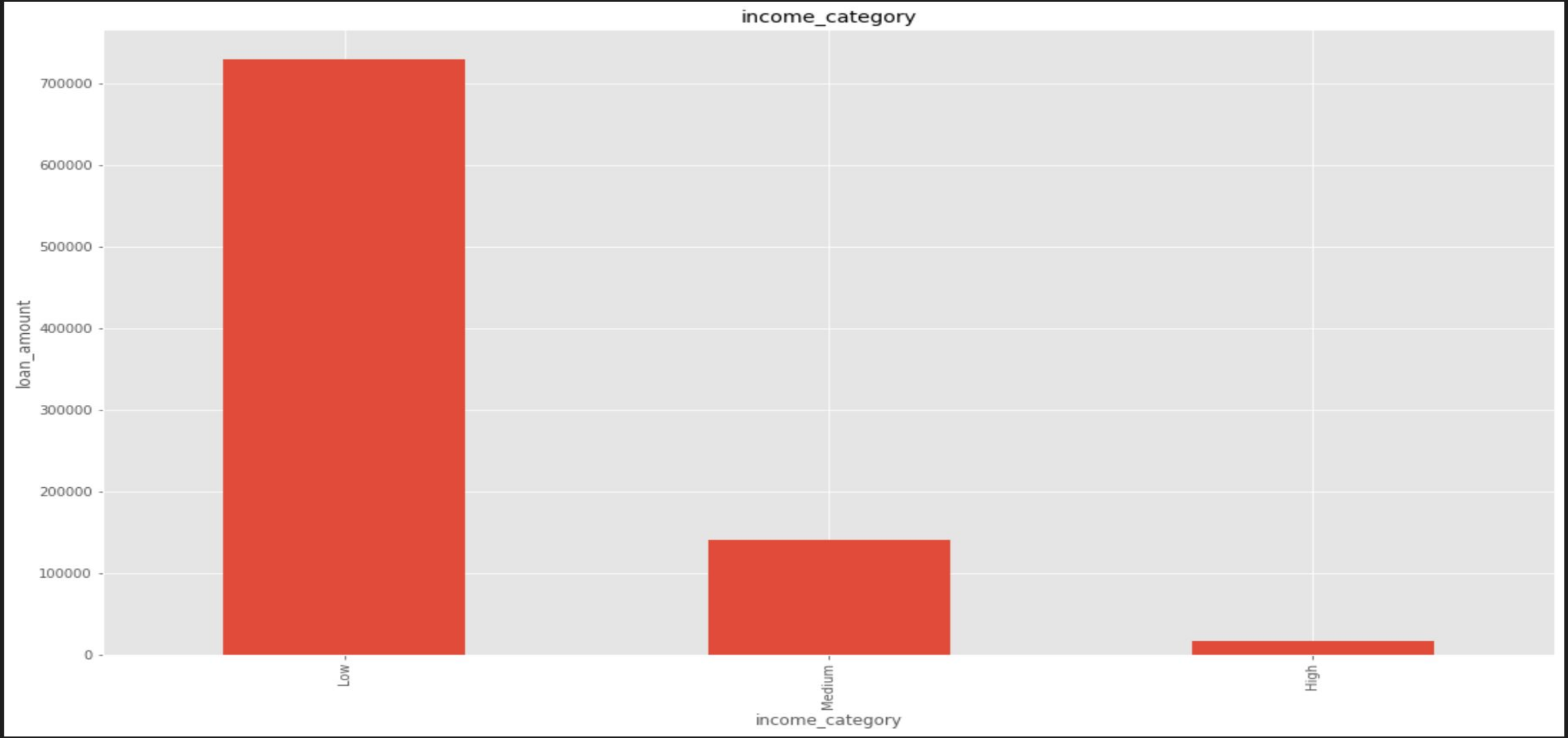
	id	year	issue_d	final_d	emp_length_int	home_ownership	home_ownership_cat
0	1077501	2011	01/12/2011	1012015	10.0	RENT	1
1	1077430	2011	01/12/2011	1042013	0.5	RENT	1
2	1077175	2011	01/12/2011	1062014	10.0	RENT	1
3	1076863	2011	01/12/2011	1012015	10.0	RENT	1
4	1075358	2011	01/12/2011	1012016	1.0	RENT	1
...	...	...	...	...	...	...	...
887374	36371250	2015	01/01/2015	1012016	8.0	RENT	1
887375	36441262	2015	01/01/2015	1012016	10.0	MORTGAGE	3
887376	36271333	2015	01/01/2015	1012016	5.0	RENT	1
887377	36490806	2015	01/01/2015	1012016	1.0	RENT	1
887378	36271262	2015	01/01/2015	1012016	10.0	RENT	1

887379 rows × 30 columns

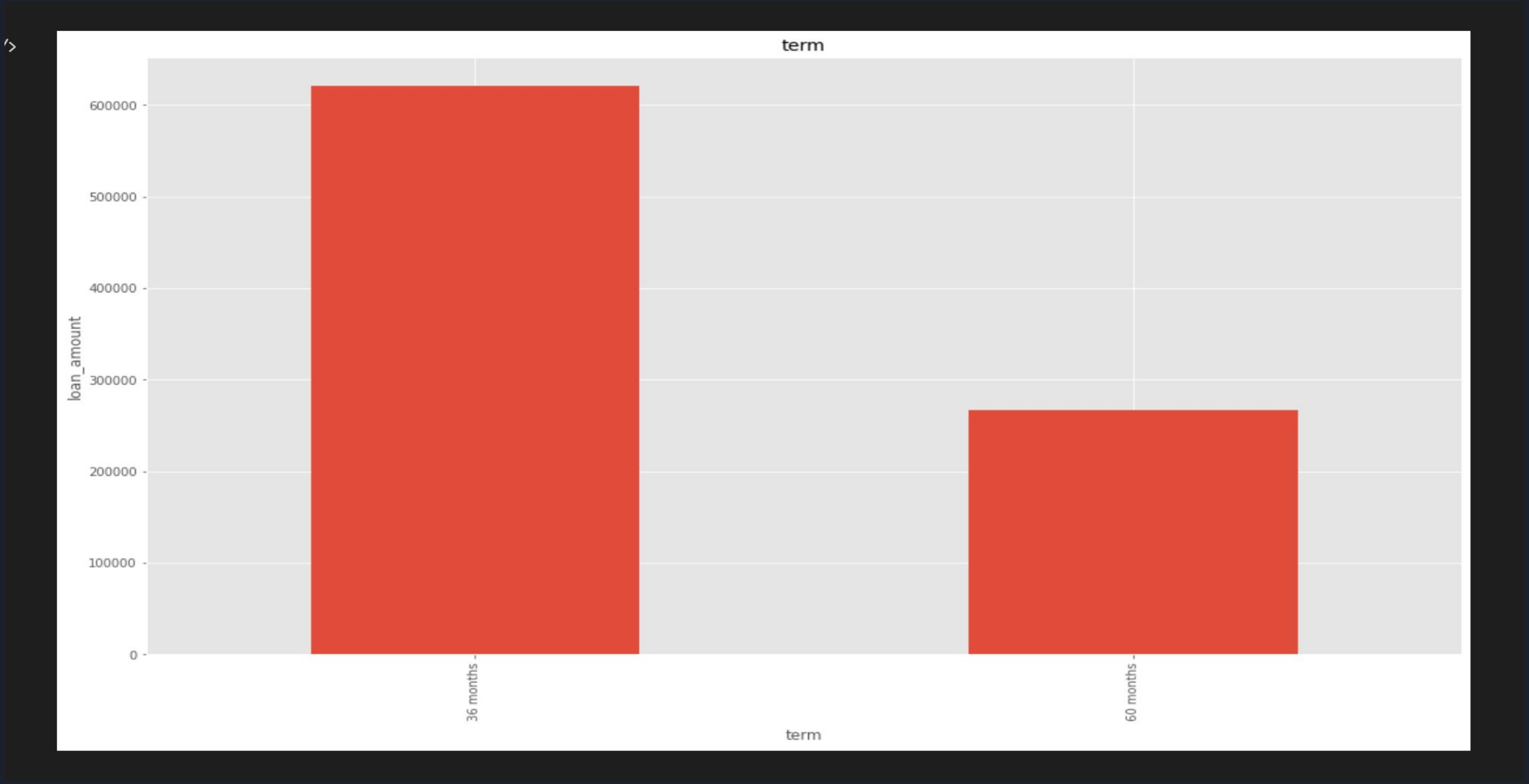
Exploratory Data Analysis



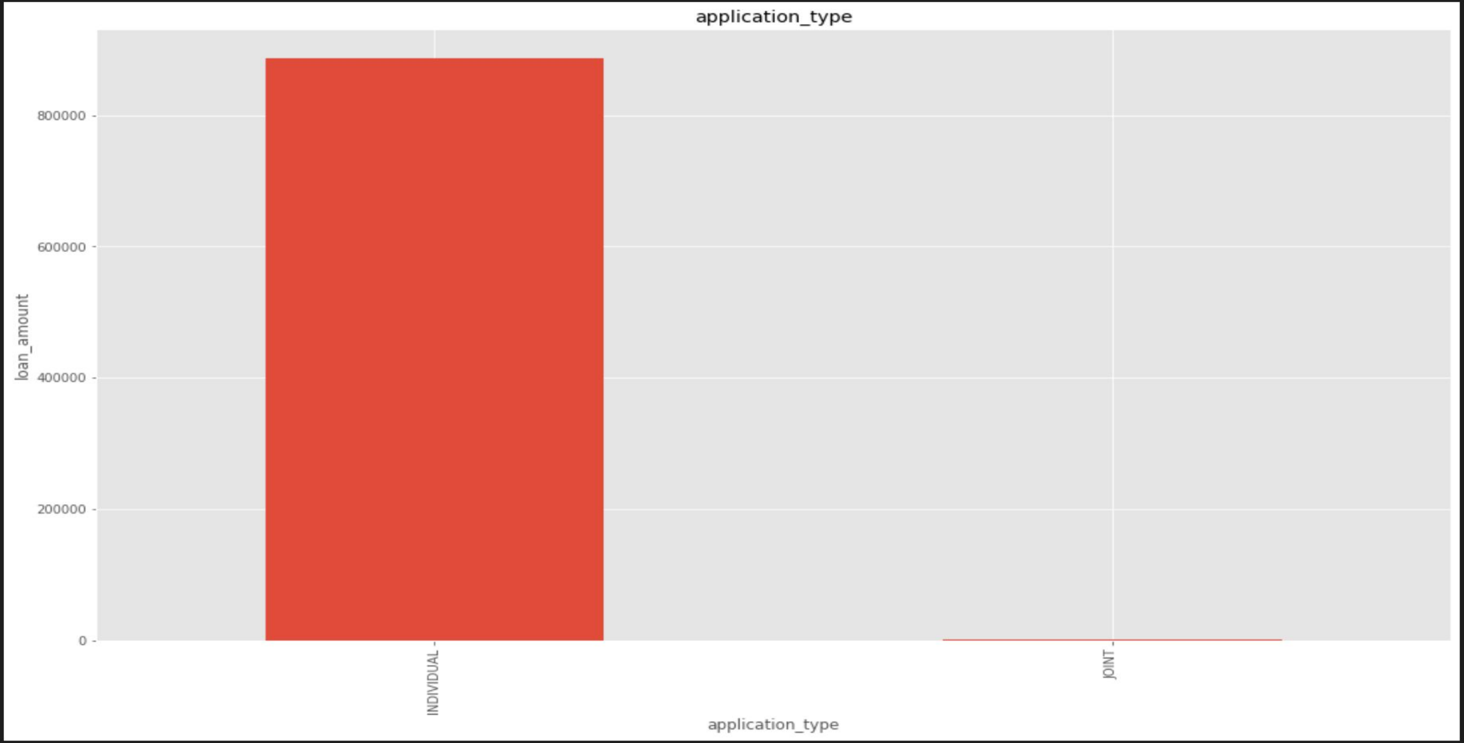
Income Category



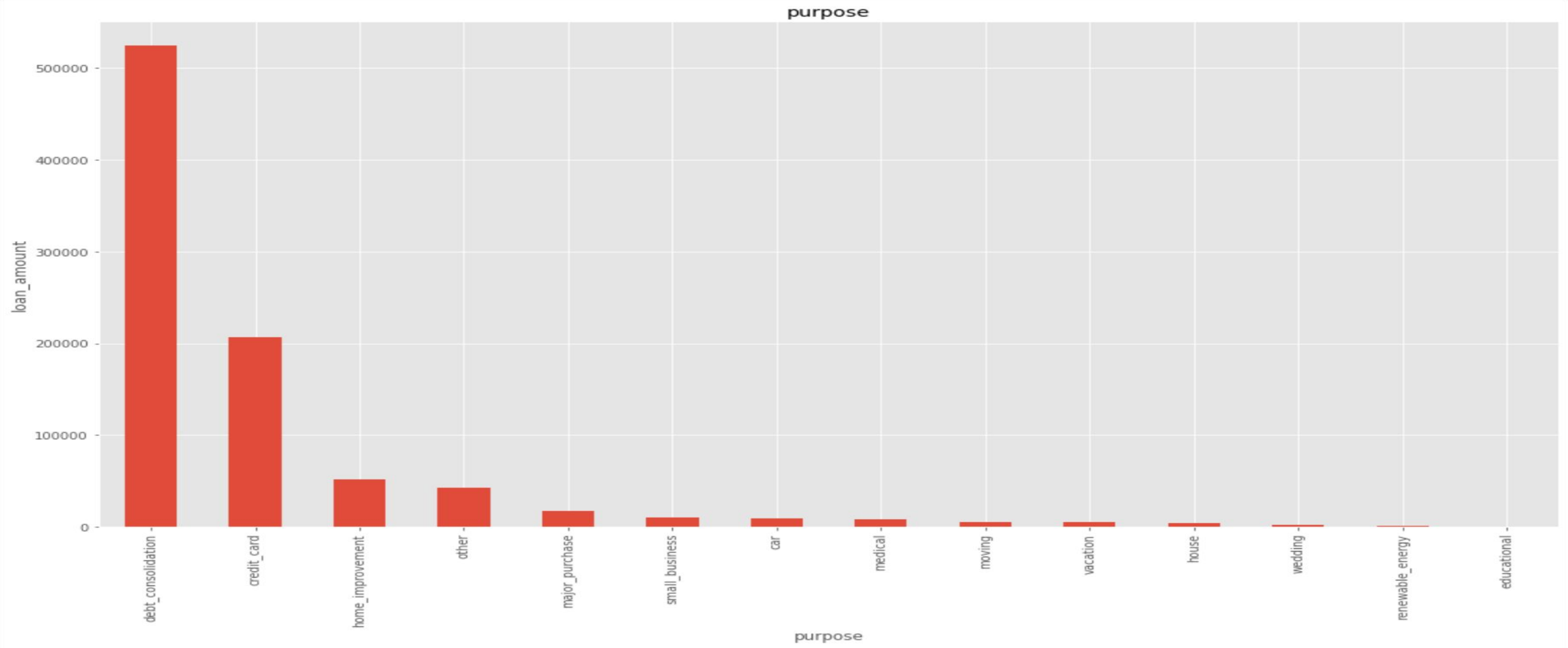
Loan Maturity Dates



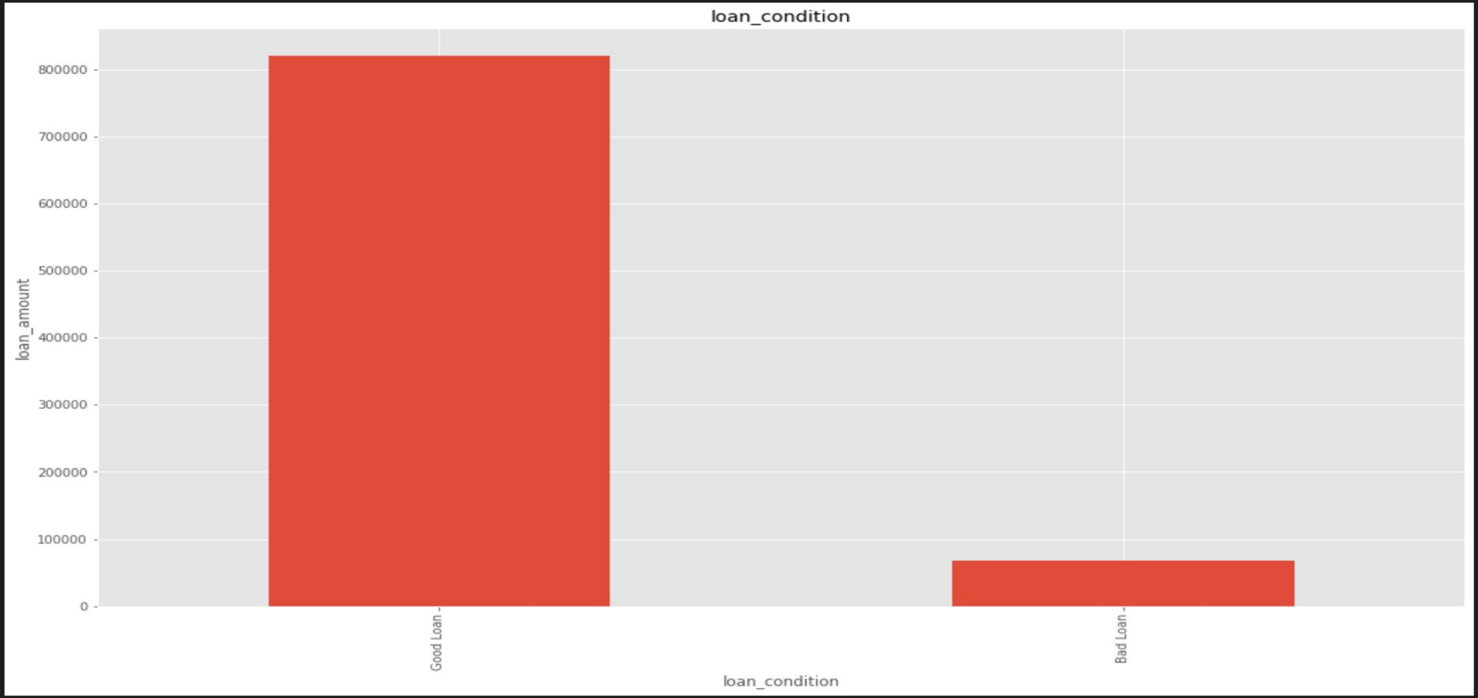
Application Type



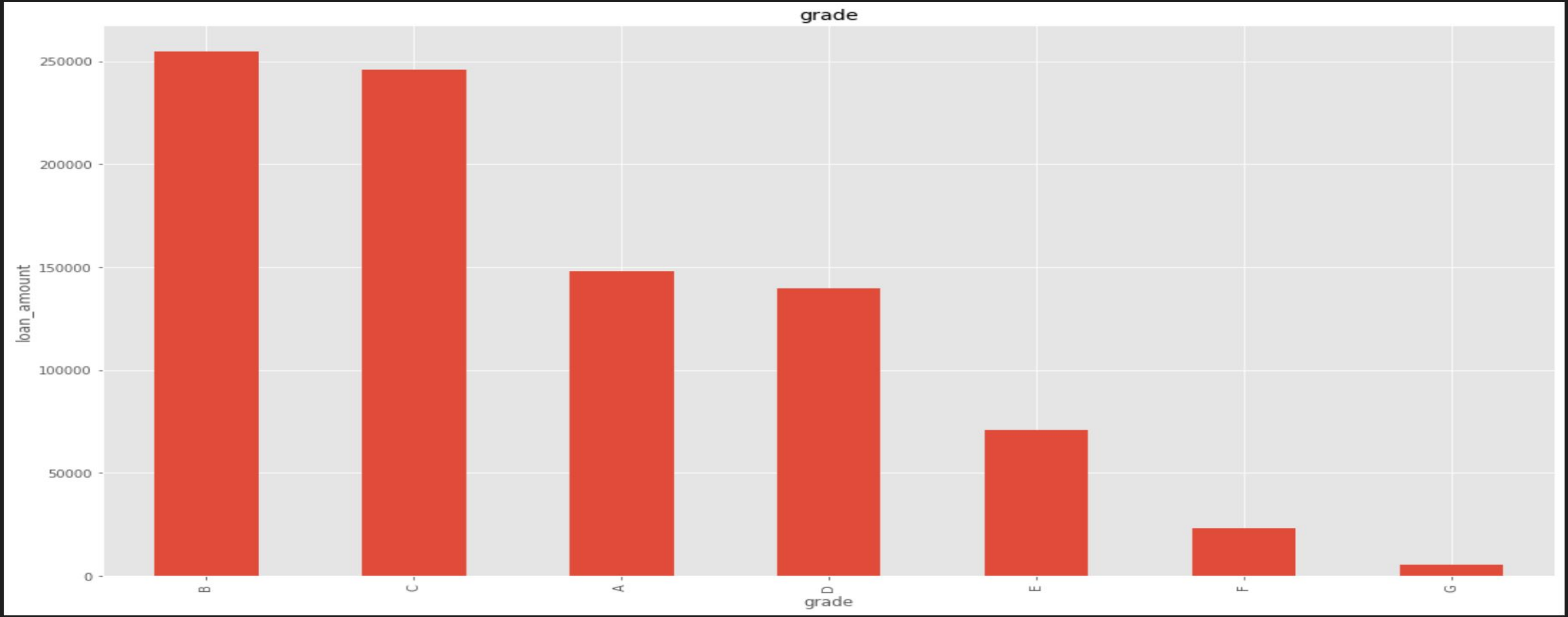
Loan Purpose



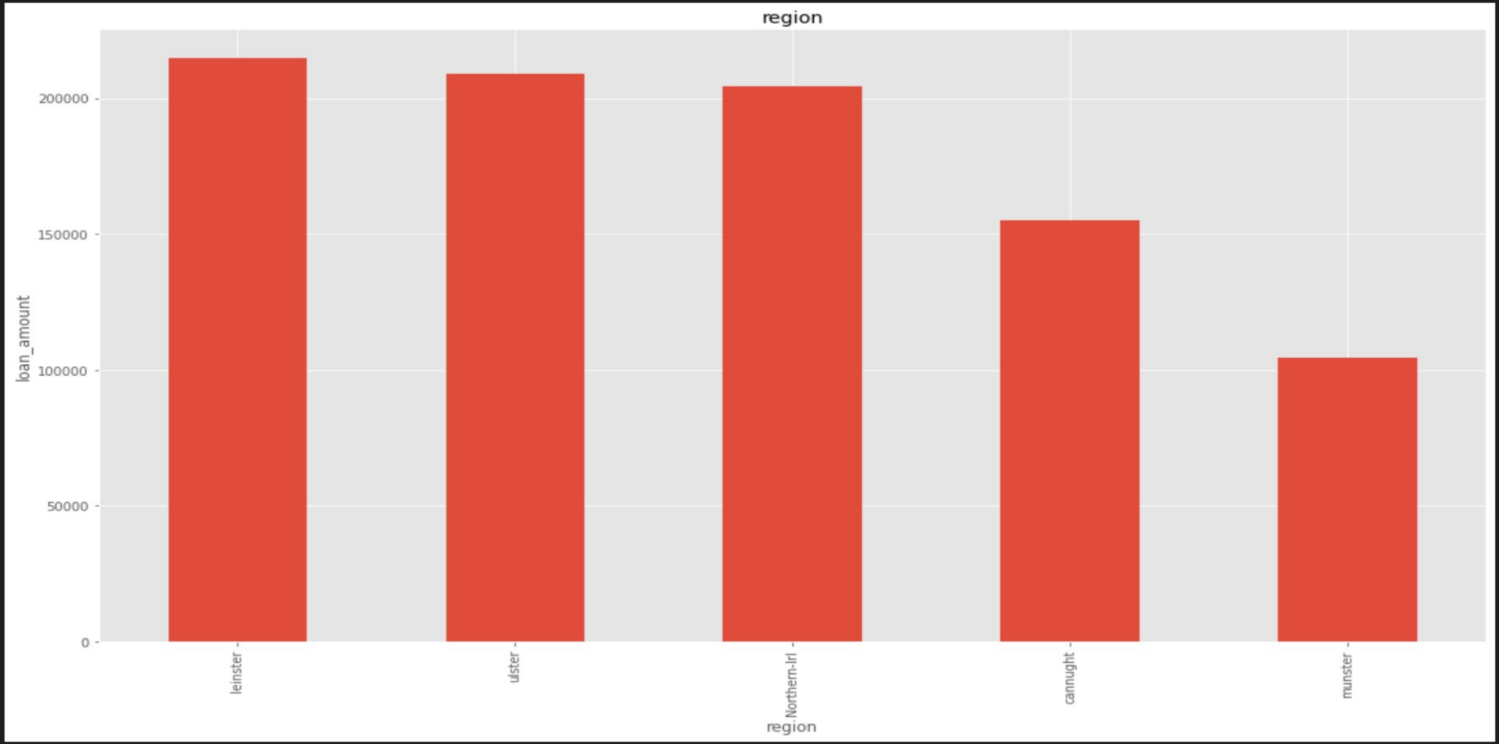
Loan Conditions



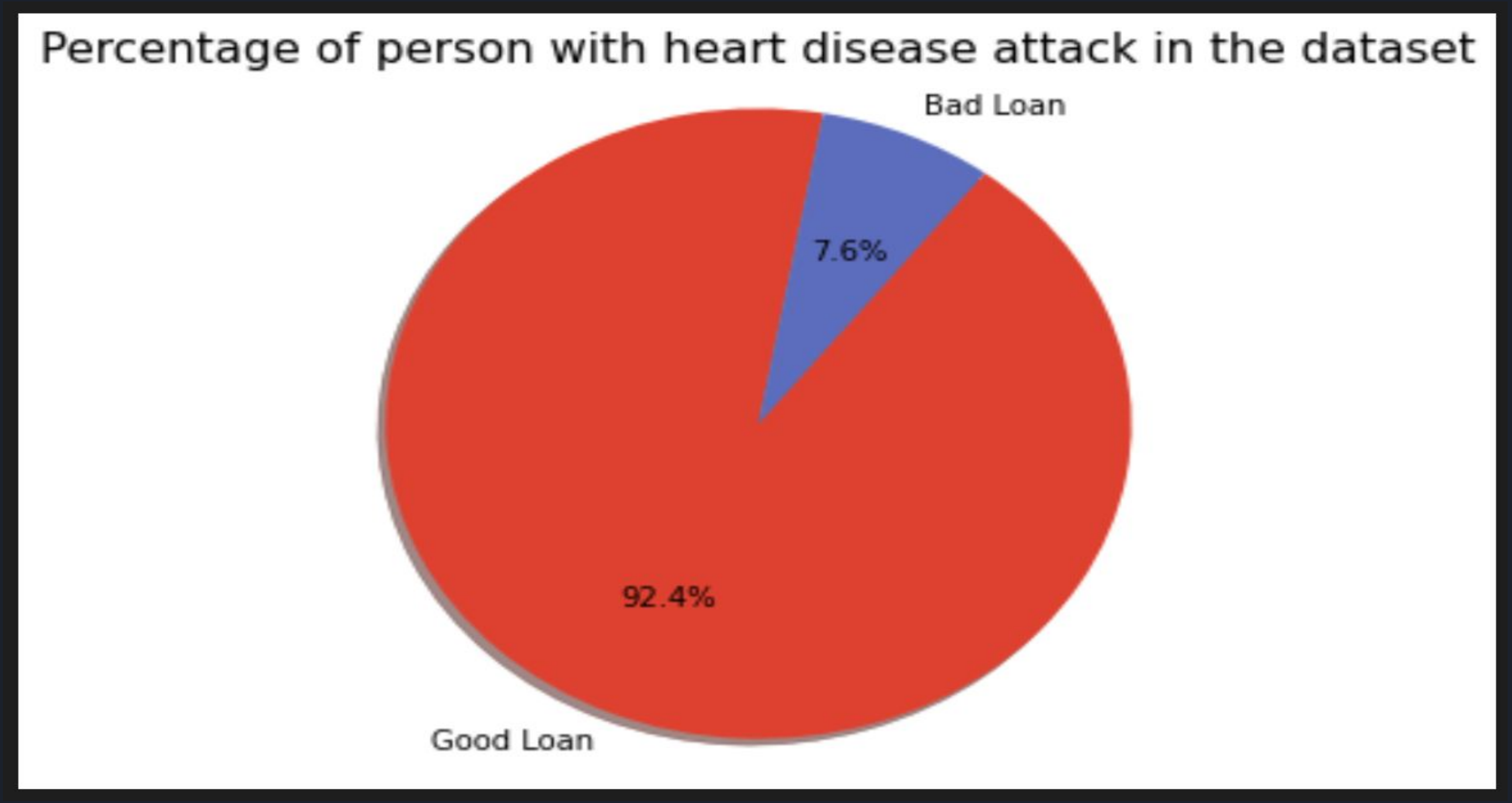
Loan Grades


grade
250000
200000
150000
100000
50000
0
loan_amount
B
C
A
D
E
F
G
grade

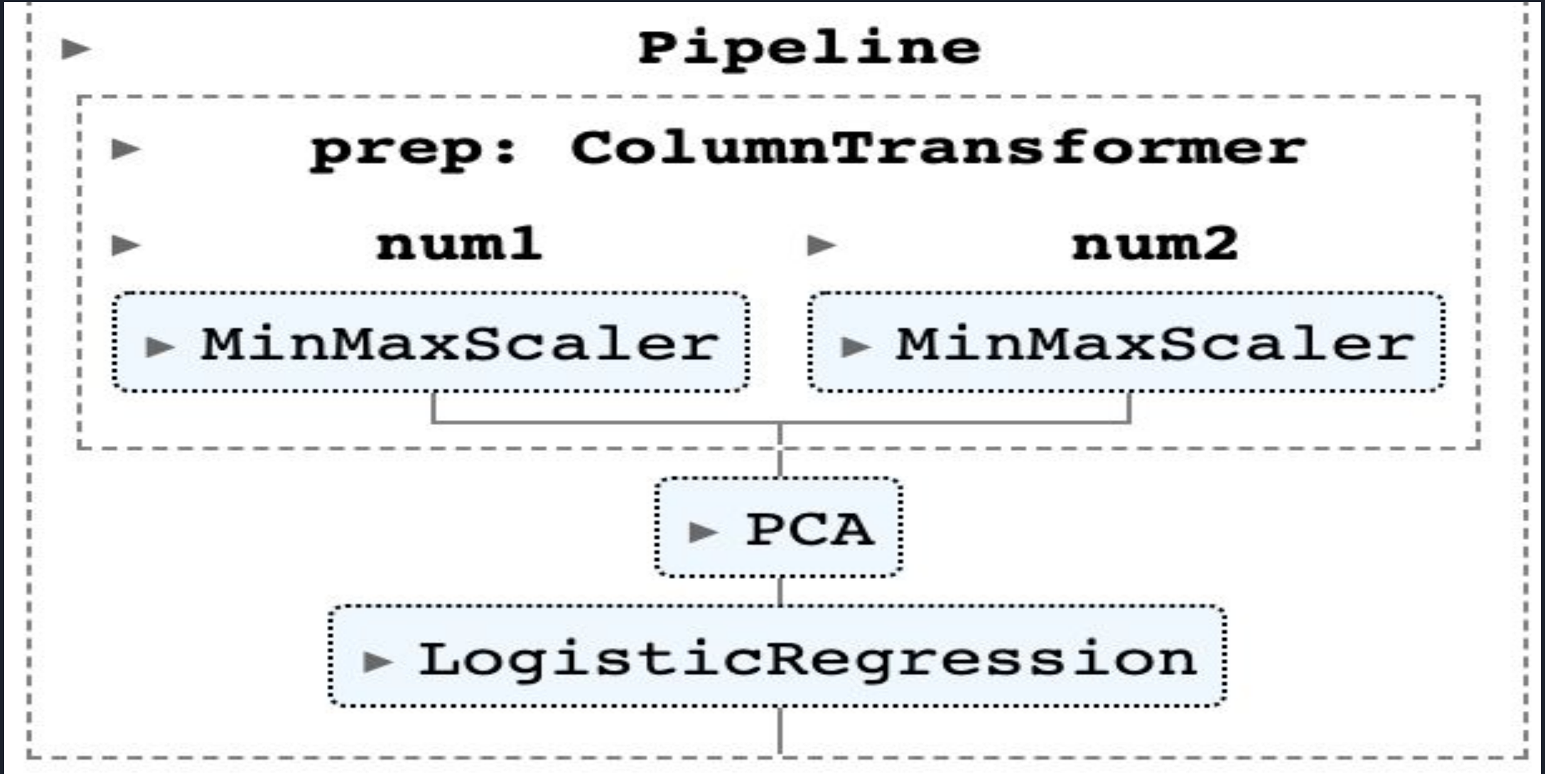
Loan based on region



Data Target Proportion



Building Data Pipeline



Pipeline

prep: ColumnTransformer

num1

MinMaxScaler

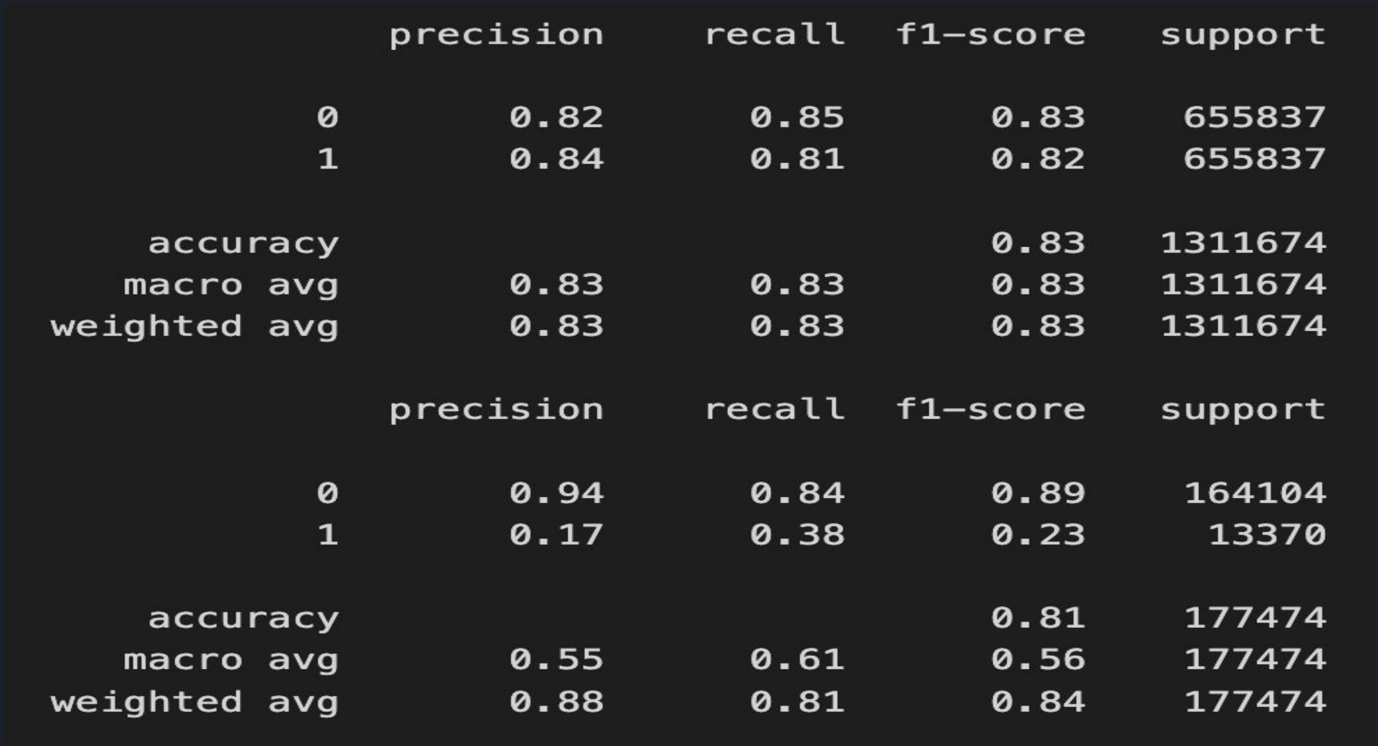
num2

MinMaxScaler

PCA

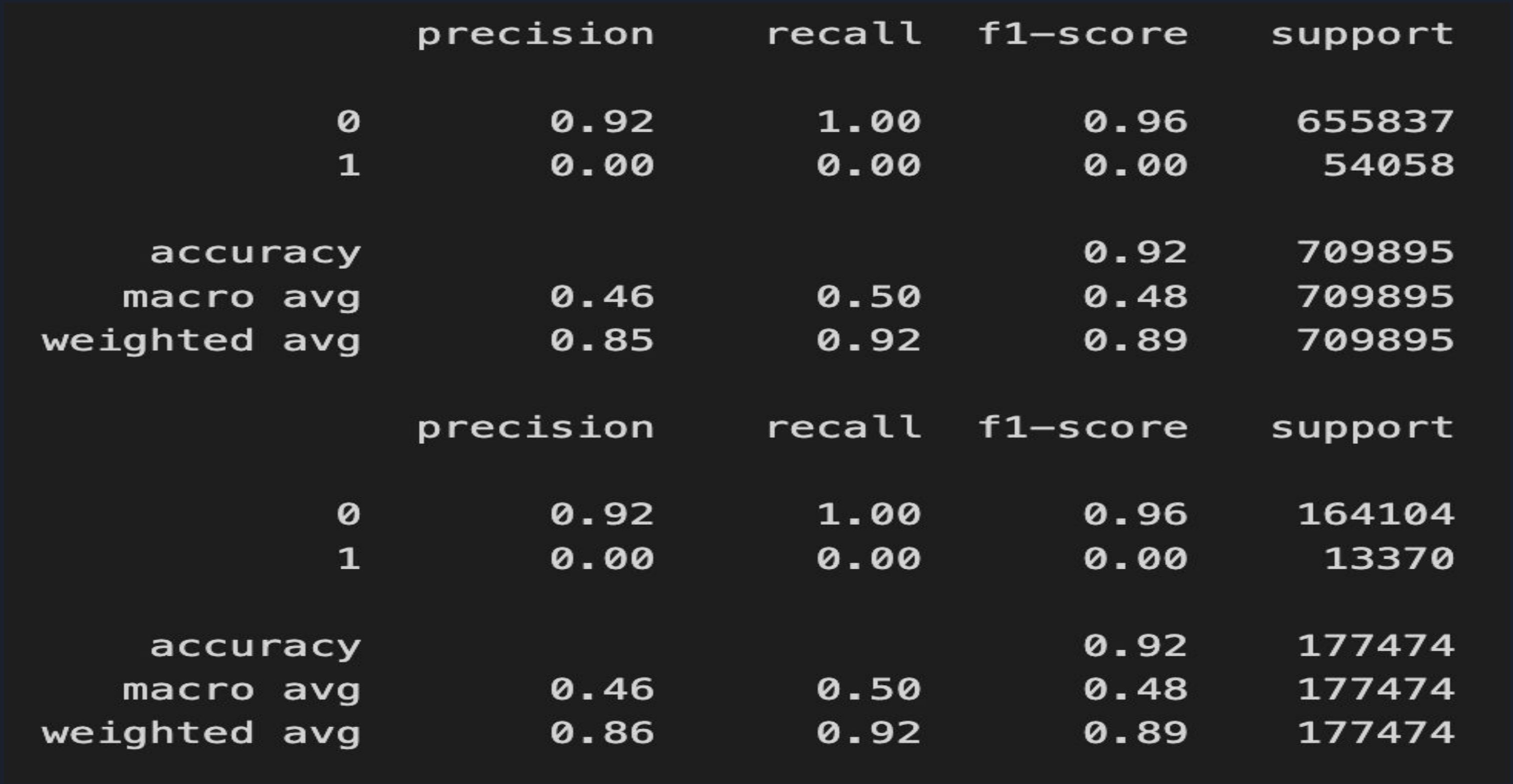
LogisticRegression

Model Evaluation by Implementing SMOTE



```
              precision    recall  f1-score   support

           0       0.82      0.85      0.83    655837
           1       0.84      0.81      0.82    655837

    accuracy                           0.83   1311674
   macro avg       0.83      0.83      0.83   1311674
weighted avg       0.83      0.83      0.83   1311674

              precision    recall  f1-score   support

           0       0.94      0.84      0.89    164104
           1       0.17      0.38      0.23     13370

    accuracy                           0.81    177474
   macro avg       0.55      0.61      0.56    177474
weighted avg       0.88      0.81      0.84    177474
```

Model Evaluation without SMOTE



```
              precision    recall  f1-score   support

           0       0.92      1.00      0.96    655837
           1       0.00      0.00      0.00     54058

    accuracy                           0.92    709895
   macro avg       0.46      0.50      0.48    709895
weighted avg       0.85      0.92      0.89    709895

              precision    recall  f1-score   support

           0       0.92      1.00      0.96    164104
           1       0.00      0.00      0.00     13370

    accuracy                           0.92    177474
   macro avg       0.46      0.50      0.48    177474
weighted avg       0.86      0.92      0.89    177474
```

Conclusion

The result of this logistic regression research showed that the accuracy without handling imbalance dataset was higher than the accuracy by handling imbalanced dataset. Otherwise the technique of handling the imbalanced dataset was working to teach the lowest proportion of target variable to learn from the training. It was shown in the metric valuation with handling imbalance data which shoId that the label one in data training have 80% of accuracy. Even-though it was still overfitting because the test accuracy was at 79 %. But it was better because in the specific metric valuation such as precision and recall that have score at 0.82 and 0.80 respectively and still overfitting because the precision and recall in the test Ire at 0.15 and 0.49.