AMITY UNIVERCITY

MACHINE LEARNING & AI

LENDING CLUB DATA ANALYSIS AND DEFAULT LOAN/RATING PREDICTION

PROJECT REPORT

**SUBMITTED BY**

**KIRAN KALICHERLA**

**1.Introduction to the project:**

**This is ML and AI course project. Under the scope of the course project we are required to solve an analysis/learning problem using the techniques taught in the course.** **We will use “Lending Club historical dataset” for this project. This is an open source dataset which contains complete loan data for all the loan issued through 2007-2015. The data is available to download on the following site**- <https://www.kaggle.com/wendykan/lending-club-loan-data>

**Lending Club is an online peer to peer credit marketplace which matches borrowers and investors. For evaluating the credit-worthiness of their borrowers, Lending Club relies on many factors related to borrowers such as credit history, employment, income, ratings etc. Lending club then assign’s rating/sub-rating to their borrowers based on their credit-history.**

**Our project scope is to run the exploratory data analysis using Python/Scikit-Learn to find the business insights from our loan data, and to build a learning model using machine leaning algorithms that will use the historic loan data to learn and helps to identify loans/borrowers which are likely to default. As per the recent studies, 4-5% of the total loans defaults every year. This is the huge risk for the investors who is funding the loans. Investors require more comprehensive assessment of these borrowers than what is presented by Lending Club in order to make a smart business decision. Machine Learning model/analysis could help predicting the loan default likelihood which may allow investors to avoid loan defaults thus limiting the risk of their investments.**

**2.Projcet Problem:**

**Run the exploratory data analysis using Python/Scikit-Learn to find the business insights from our loan data, and to build a learning model using machine leaning algorithms that will use the historic loan data to learn and helps to identify loans/borrowers which are likely to default.**

**As per the recent studies, 4-5% of the total loans defaults every year. This is the huge risk for the investors who is funding the loans. Investors require more comprehensive assessment of these borrowers than what is presented by Lending Club in order to make a smart business decision. Machine Learning model/analysis could help predicting the loan default likelihood which may allow investors to avoid loan defaults thus limiting the risk of their investments.**

**3.Key Contributions:**

**For this project, I performed explanatory data analysis using python, Pandas and scikit frame work to gather business insights from the historical loan data. I built 4 machine learning classifiers to identify the borrowers who are more likely to default on their loan. I developed Logistic Regression, Random Forest, XGBOOST and Regularized Linear Regression(Lasso) models. I have used Weigh of Evidence (WOE) approach for encoding the categorical variables.**

**I have used confusion matrix and Classification report to evaluate each model. Among these models Logistic Regression predicting the nearly perfect defaulters.**

**Technology Specs: Python, Pandas, Scikit Learn ,Matplotlib ,Jupyter, Seaborn.**

**4.Machine Learning:**

**4.1 Data set recourse:**

The data used for this project is the structured data with few missing/null values. This data consists of 80+ features of three distinct types: continuous, categorical, ordinal. We gathered business domain knowledge about the data to deal with data cleaning and missing data imputation. During our preliminary exploration of data, we noticed some of the features with more than 50% of the missing data, we planned to drop these features while data processing for learning model.

For data analysis phase, we converted some of the feature type to relevant primitive types and moved on. We handled data cleaning and imputation part while preparing data for learning model. For our preliminary data cleanup, we used pandas library to convert the feature data types.

As part of our exploratory data analysis we tried to find the interesting fact and findings about the loans from the historical loan data. This analysis helped us to develop our understanding about data and its distribution patterns. In addition, this assisted us to select the most effecting features and develop business rules for missing data imputation.

* **Analyzing loan amount and interest rates:**

By using Seaborn , we aggregated our data by loan amount and loan interest rate and plotted the distribution plot and box plot.

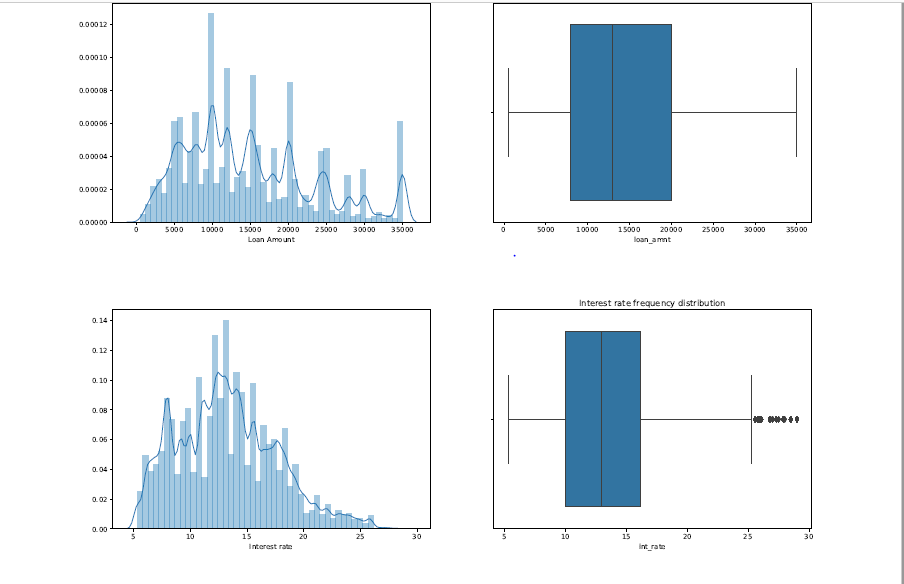


Figure1: Distribution of loan and Interest.

From above Figure 1, we conclude that both Interest rate & loan amount is normally distributed, slightly right skewed. i.e. most of the customers are seeking loan amount ranging 5-15K and their interest rate is ranging from 10-15%. Our loan amount shows no outliers point, since it fits perfectly within the five-point summary. However, Interest rates shows few outliers falling outside the 1.5IQR range. Those customers may have **poor credit ratings**, therefore **high interest rate.**

We analyzed the distribution for each of our features included in our final model. From our analysis, we determine the appropriate distribution function to use in estimating Maximum Likelihood Estimator. The more detailed description is added in the later section of the report.

* **Analyzing Loans Interest rates over time:**

To analyze how are loan book is changing over time in terms of loan amount and interest rate, we first created a new column for quarterly date range. I used pandas dt.year and dt.quarter to extract this information from the loan issue date column. After which I aggregated the data over quarterly date range and plotted in the point plot. To get the total loan amounts over the time we grouped by on the date variable and aggregate all the loans amount to get the total loans.

I calculated customers loan requirements over the time (Median loan amount and median interest) by using *median()*  function.

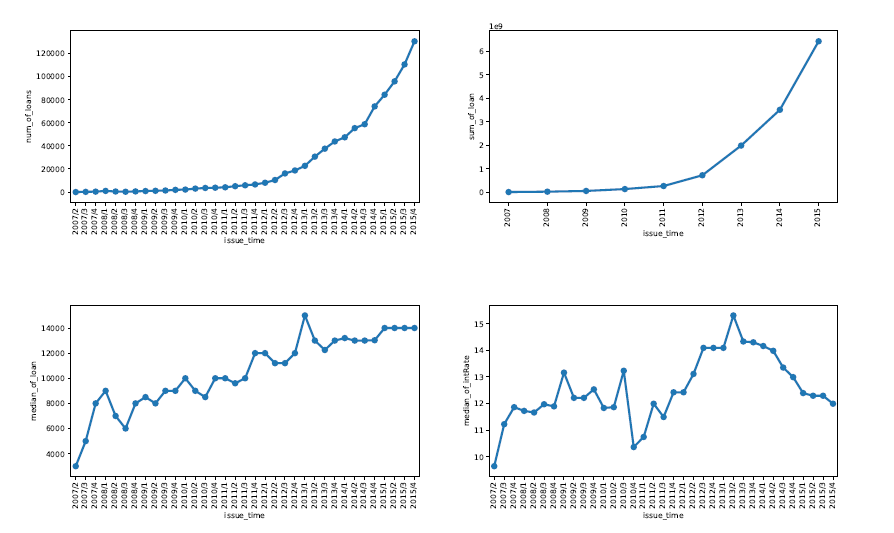


Figure2: Analyzing Loans Interest rates over time.

The first two plots show how number of loans and total loan amount is growing over time, this gives an indication of growing borrowers at lending club platform. The next two plots are between median loan amount & interest over time. This indicates that the borrower’s loan requirements are increasing over time so as the median interest rates. However, there is a sharp decline in the interest rate during 2010, which might be due to 2009 fiscal crisis.

* **Analyzing Loans over loan Status:**

We analyze the loans according to their current status. I used Seaborn library to plot the Loans distribution over loan status.

To show the distribution of loan amount, we have used the violin plot (figure 5). A Violin Plot is used to visualize the distribution of the data and its probability density. This chart is a combination of a Box Plot and a Density Plot. The density plot is rotated and placed on both sides of the box plot, to show the distribution shape of the data.

We also analyzed about the verification status of the loan which were defaulted. We achieve this result by running a group by aggregations on verification\_status and default loan status.

|  |  |
| --- | --- |
| verification\_status | Count |
| Verified | 479 |
| Source Verified | 462 |
| Not Verified | 278 |

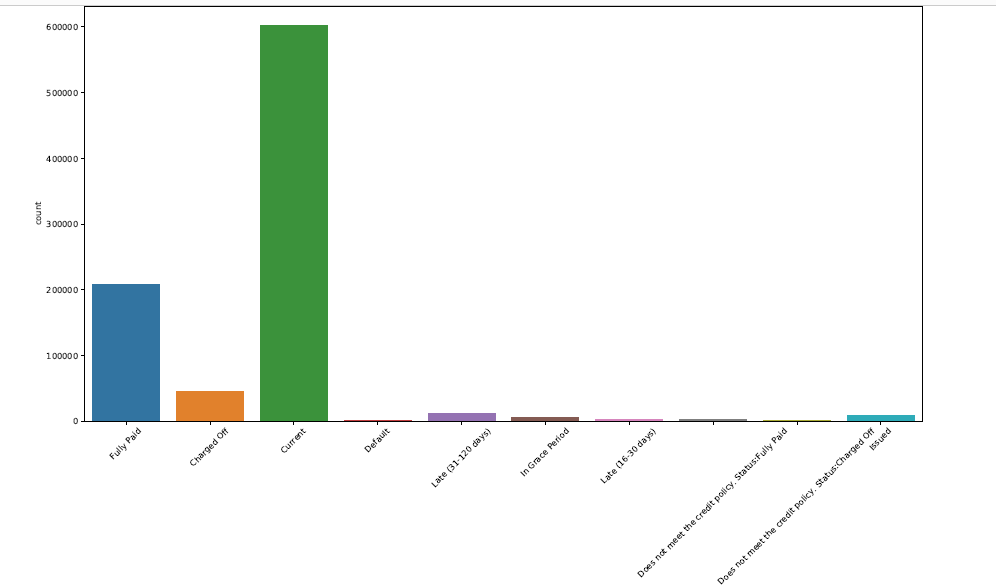


Figure3: Bar Plot for analyzing loans over loan status.

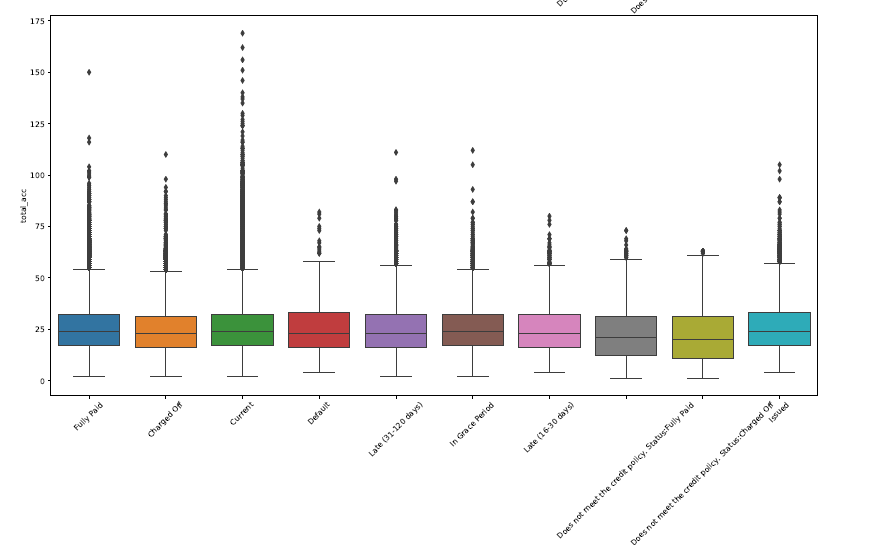


Figure 4: Box plot for analyzing loans over loan status.

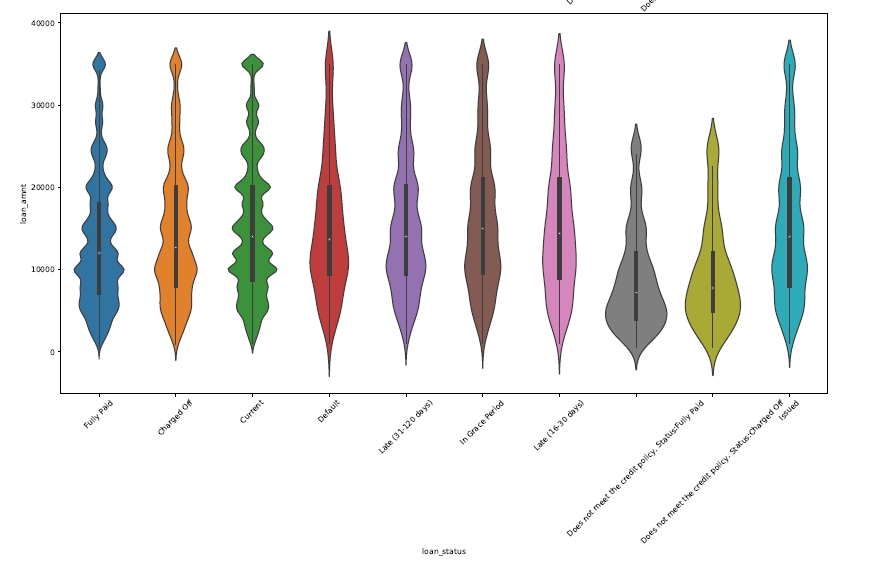
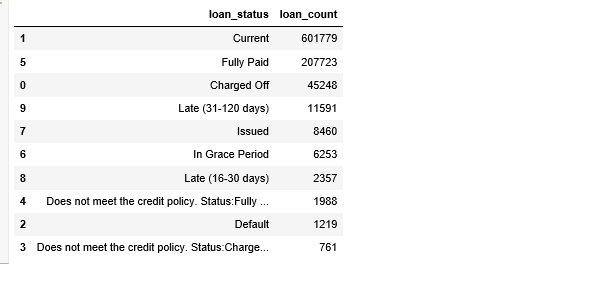


Figure 5: Violin plot for analyzing loans over loan status.

By running a group by query, we found out the number of loans for each loan status. The results are listed in following table.

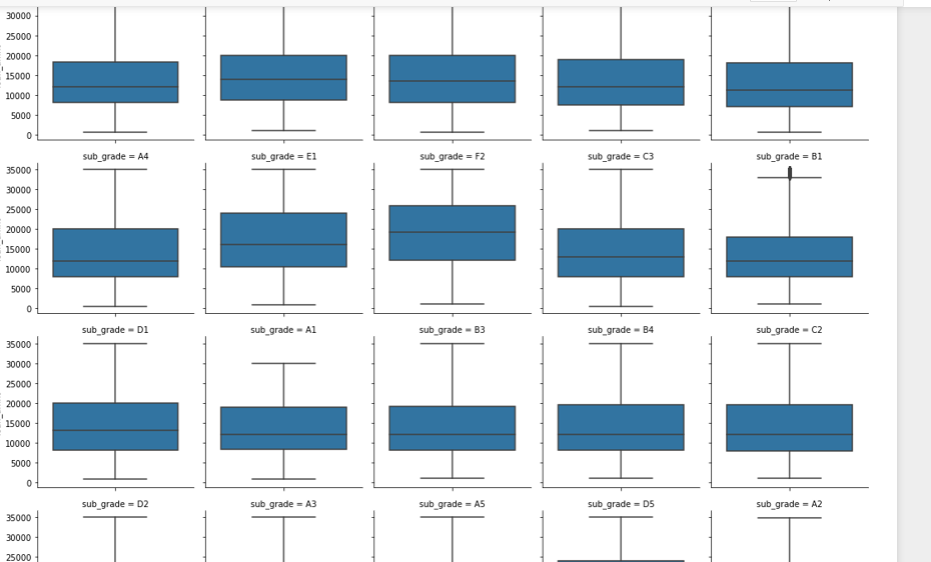


* **Analyzing loan amount and interest rate quantile summary for each grade, factored over sub grade.**

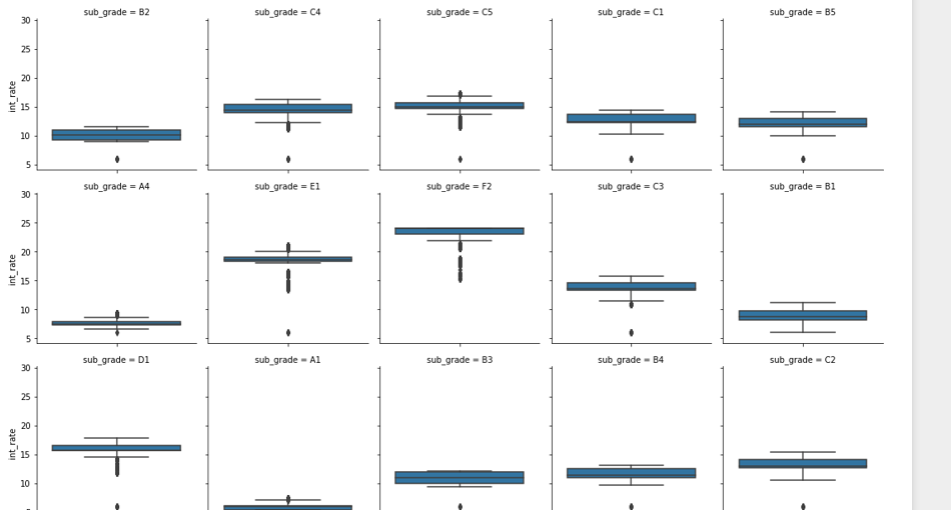
Based on borrower’s credit score, credit history, desired loan amount and the borrower’s debt-to-income ratio, Lending Club determines whether the borrower is credit worthy. After that they assign a credit grade that determines payable interest rate and fees to their approved loans. These grades are assigned within an alphabetical range from A to G. Each of these letter grades has five finer-grain sub-grades, numbered 1 to 5, with 1 being the highest category with-in the grade. Loan interest rates is inverse proportionate to the credit grade. ‘A’ being the highest grade, therefore low interest rate and vice-a-versa.

In figure 6, we can see a linear relationship between loan amount and customer credit ratings, notice here that requested loan amount is slightly higher for the low rating customers. This indicates that higher rating borrowers are financially stronger and require lesser loan amount in general.

In figure 7, we can see that the median interest rates increase for low credit rating customers.



**Figure 6:Analyzing loan amount distribution for each grade , factored over sub grade.**



**Figure 7: Analyzing interest rate distribution for each grade, factored over sub grade.**

**Data set Pre-processing:**

Features present within the dataset provided an ample amount of information which we could use to identify relationships and gauge their effect upon the success or failure of a borrower fulfilling the terms of their loan agreement. We required only the variables that had a direct or indirect response to a borrower’s potential to default. To achieve this, we have prepared the data by choosing select variables that would best fit these criteria

* **Data Cleaning:**

As a first step, we removed all unique id fields which is to represent a loan request, since it does not contribute in data analysis or model building. Next, we removed all the features which has more than 50% missing data. After which removed few categorical features which had only one category in the data.

There were few features in our dataset which were specific to the defaulted loans, we removed all the features since they possibly can create leakage in our model.

* **Missing Data Imputation**:

For Categorical variables, I have imputed missing data with “Missing”. For numerical variables missing data is imputed with median value. I have used Pandas to impute the missing data.

* **Label Creation:**

Our data doesn’t have the label column by itself but there is combination of the features which represent the default loan. For our classification model we needed to add a class label variable to our data set. We added a new column “*isDefault*” and populate the values based in the combinations of features such as *loan\_status, tot\_curr\_bal, delig\_month* etc. The class count for each class is shown in following table.

|  |  |
| --- | --- |
| Class Label | Count |
| 1(Defaulted) | 61176 |
| 0(Not Defaulted) | 826203 |

* **Encoding:**

After data cleaning and missing data imputation, I have encoded all categorical and discrete variables using method **categories\_to\_woe** (Used Weight of evidence concept). Categorical features are [ 'grade', 'sub\_grade', 'emp\_title', 'emp\_length', 'home\_ownership', 'verification\_status', 'purpose', 'title', 'addr\_state'] and discrete Features are [collections\_12\_mths\_ex\_med', 'policy\_code', 'acc\_now\_delinq'].

**4). Machine Learning Model Implementation:**

For this project, we built and tested four different classifier learning models. Each of our learning models was comprised by a different combination of the hyper parameters relevant to the individual learning models we learnt over the span of our degree courses. Each of the classifiers are implemented using Python and Pandas,scikit learn,Jupyter notebook,Seaborn and Matplotlib libraries.

These models are evaluated using Confusion matrix and classification report.

**1.xgboost**

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the [Gradient Boosting](https://en.wikipedia.org/wiki/Gradient_boosting) framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

I used XGBRegressor from xgboost.

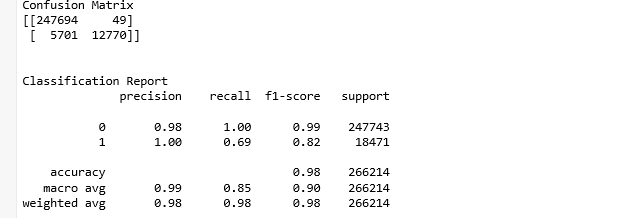


Table1: Result Summary of XGBRegressor.

**Accuracy – 98%**

**2.Random Forest:**

The next algorithm we tested was Random Forest (the ensembles of decision trees) to further deal with the data imbalance issue. Random forest is the aggregation of multiple decision trees which uses entropy/Gini to find the impurities, which makes it less sensitive to the class imbalance. I used RandomforestRegressor from sklearn.ensemble.

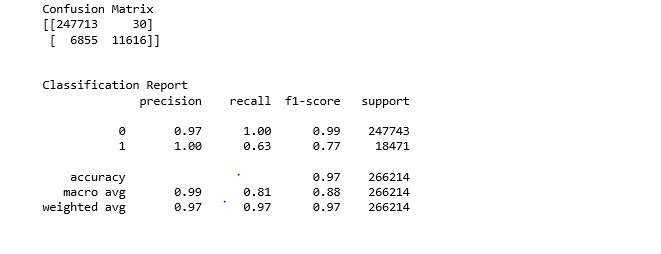


Table2: Result Summary of Random Forest.

**Accuracy – 97%**

**3.Logistic Regression:**

Logistic regression is a popular method to predict a categorical/binary response. It is a specialized case of “Generalized Linear Models” that predicts the probability of the outcomes. The main intention behind using this learning algorithm is to handle the class imbalance.

I used LogisticRegression from sklearn.linear\_model.

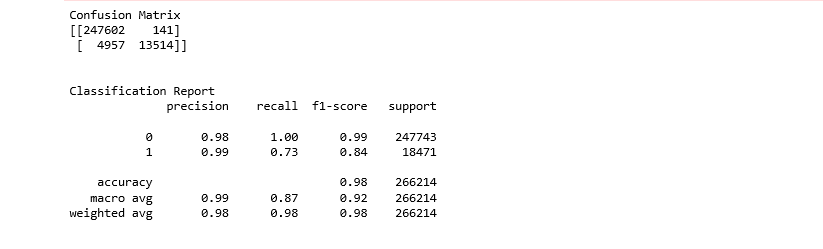


Table3: Result summary of Logistic Regression.

**Accuracy – 98%**

**4.Regularized Linear Regression(Lasso):**

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The **lasso** procedure encourages simple, sparse models. I used Lasso from sklearn.linear\_model.

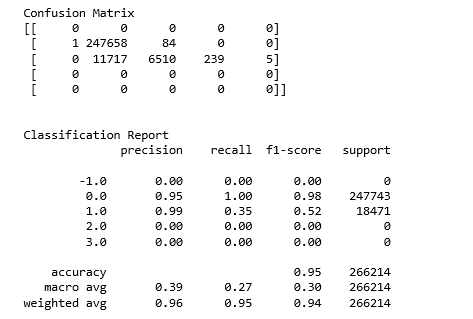


Table 4: Result Summary of Lasso Regression.

**Accuracy – 95%**

**5.Analysis of Results:**

After seeing the classification report and confusion matrix of the all 4 machine learning models. We can see higher accuracy to predict defaults for **Logistic Regression** model. From this I can conclude that Logistic regression with **98%** Accuracy is the best model among these 4 models.

**References**:

<https://www.lendingclub.com/>

<https://www.kaggle.com/wendykan>

<https://www.wikipedia.org/>

<https://stackoverflow.com>