Lending loan club

Group -16

**Group Members:**

**Kartik Chavan**

**Roma Patil**

**VidyalakshmiKrishnakumar**

|  |  |
| --- | --- |
| **Table Of Contents** |  |
| **1.0 Problem Statement** | **2** |
| **2.0 Introduction** | **2** |
| **3.0 Dataset** | **2** |
| **3.1 Attributes in Dataset** | **2** |
| **3.2 Grades and Interest Rates** | **3** |
| **4.0 Work-Flow of the Project** | **4** |
| **4.1 Data Cleaning** | **4** |
| **4.2 Classification** | **5** |
| **4.3 Validation** | **5** |
| **4.3.1 Confusion Matrix & Heat Map** | **6** |
| **4.4Interest Rate v/s Number of Approved Cases** | **9** |
| **5.0 Business Implications** | **10** |
| **6.0 Lessons Learned** | **10** |
| **7.0 Future Scope** | **10** |
| **8.0 References** | **11** |

**1.0 Problem Statement:**

All the banks attach a grade to a loan. There arise problems in the deciding the category of rate of interest when separated for different grades while lending loans. Even after the approval of the loan, not all of the approved loans are fully paid. Some tends to be default, late charges,etc. It becomes difficult to analyze whether the borrower is creditworthy or not. This project is a way to help the loan lenders to predict the value of the interest rate and the details about the borrower for the bank to not face any loss in the future.

**2.0 Introduction**

Lending Club is the world’s largest online credit marketplace, facilitating [personal loans](https://www.lendingclub.com/public/personal-loans.action), [business loans](https://www.lendingclub.com/business/?utm_source=LC&utm_medium=link&utm_campaign=pl_about_us&u=5), and financing for [elective medical procedures](https://www.lendingclub.com/patientsolutions). Borrowers access lower interest rate loans through a fast and easy online or mobile interface. Investors provide the capital to enable many of the loans in exchange for earning interest. They operate fully online with no branch infrastructure, and use technology to lower cost and deliver an amazing experience. They pass the cost savings to borrowers in the form of lower rates and investors in the form of attractive returns. They’re transforming the banking system into a frictionless, transparent and highly efficient online marketplace, helping people achieve their financial goals everyday.

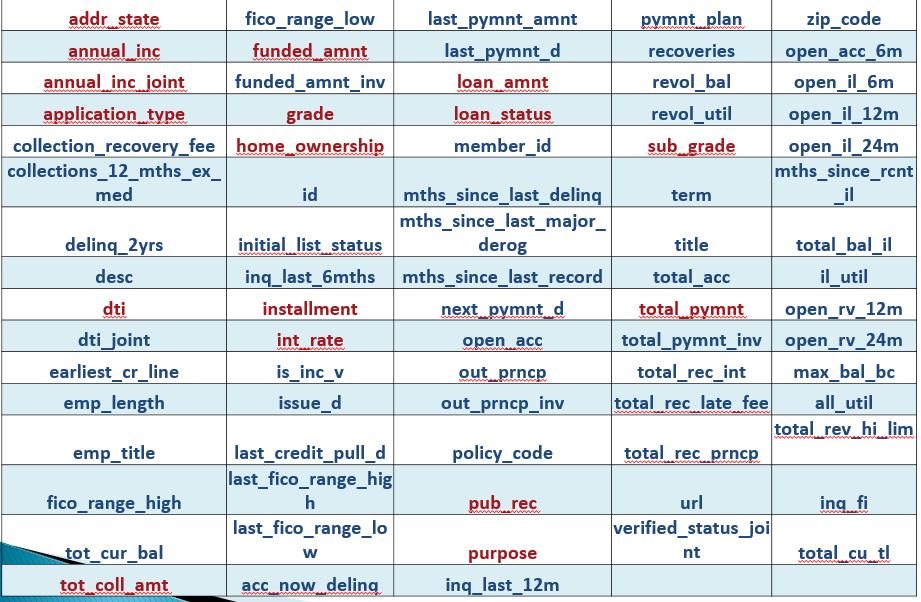
**3.0 Dataset**

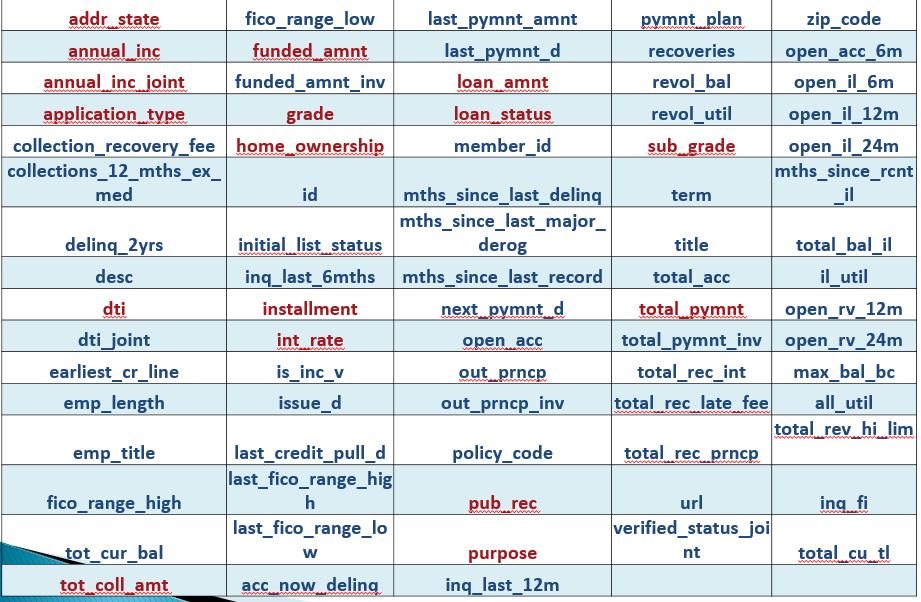
Our dataset has been obtained from <https://www.kaggle.com/>

This dataset contains complete loan data for all loans issued through the 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others. The file is a matrix of about 890 thousand observations and 75 variables. A data dictionary is provided in a separate file.

**3.1 Attributes in Dataset:**

We have the following attributes in dataset:

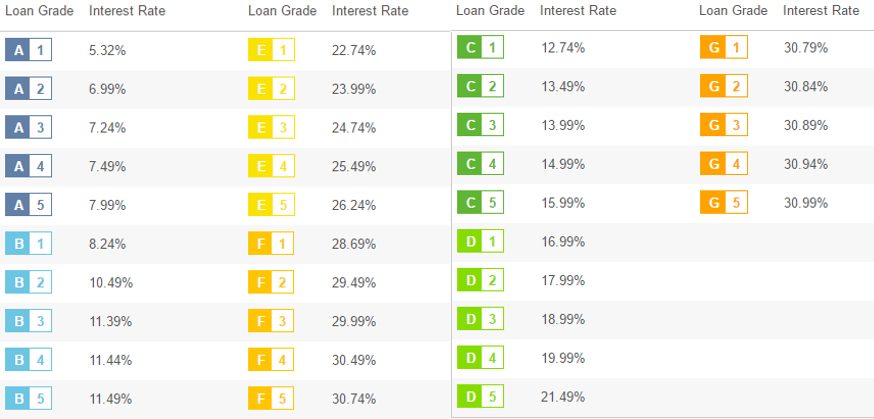




All the attributes have been explained in the LCDataDictionary file submitted in Blackboard.

**3.2 Grades and Interest Rates:**

Based on each loan application and credit report, every loan is assigned a grade ranging from A1 to G5 with a corresponding interest rate. Each loan grade and its corresponding interest rate is displayed below.

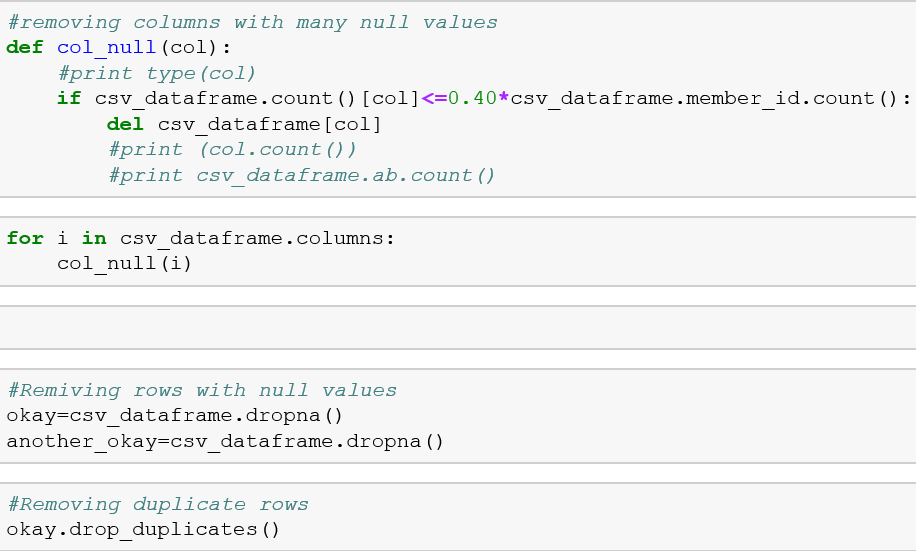


**4.0 Work-Flow of the Project:**

**4.1 Data Cleaning:**

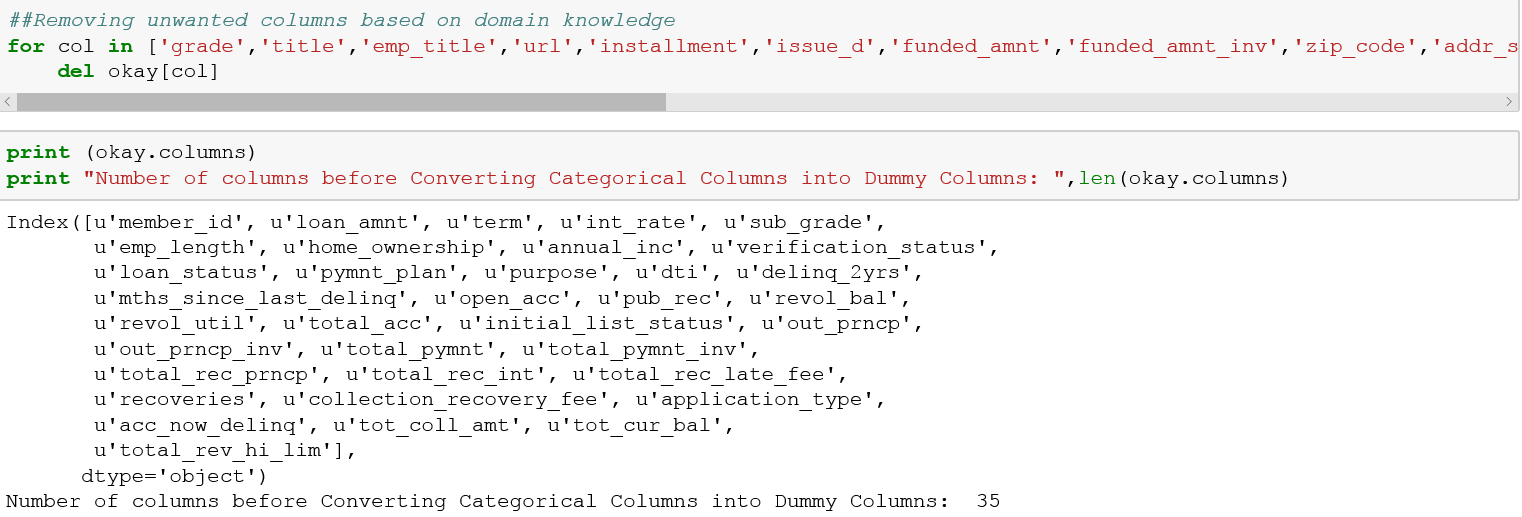
Our dataset had many columns with even more than half the null value. Hence, we defined a function called col\_null, to count the null values in each column, and remove if they are more than 40% of the total records.

Since we had a huge dataset, to remove other Null values, we removed those records. Also, we checked for duplicate records, but we didn't have any. Fig .a shows the removal of null values.



**Fig .a**

Our dataset had many categorical columns. We firstly used domain knowledge to remove all columns which were not needed for our problem statement. We also deleted few columns which had huge number of categories. Fig.b illustrates the categorical columns.



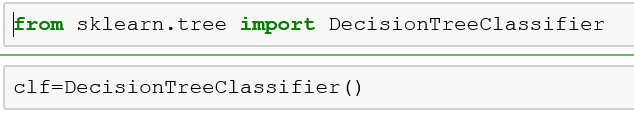
**Fig.b**

Then we used pandas get\_dummies function to convert remaining categorical columns into dummy columns. Converting into dummy variables means that if a column has 4 categorical values, then it is converted into 4 columns with binary values, 1 or 0.

After data cleaning procedures like removing columns with high null frequencies, and duplicate records, we had 32 columns, which upon converting categorical columns into dummy variables, we got 57 columns, and 291427 records which we used for classification.

**4.2 Classification:**

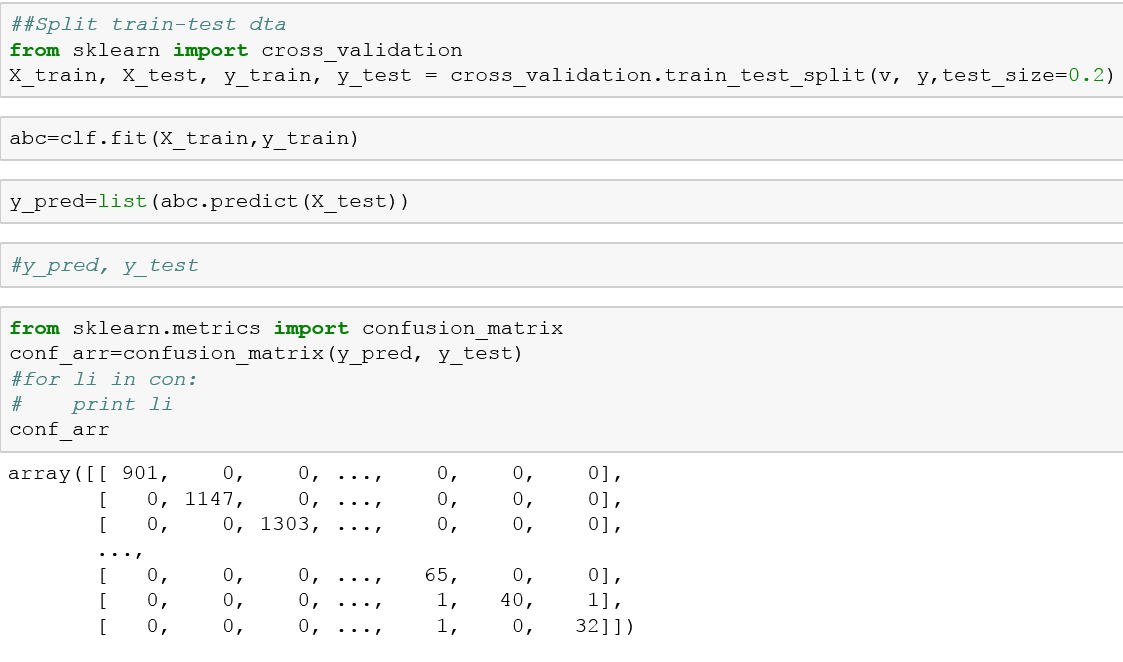
From sklearn we imported DecisionTree classifier and cross\_validation as shown in Figure c. Cross validation was to split the data into training and testing datasets. We used training dataset to train the data, while testing to test the model. We tried several, and found out that 80/20 split gave better accuracy.



**Fig.c**

**4.3 Validation:**

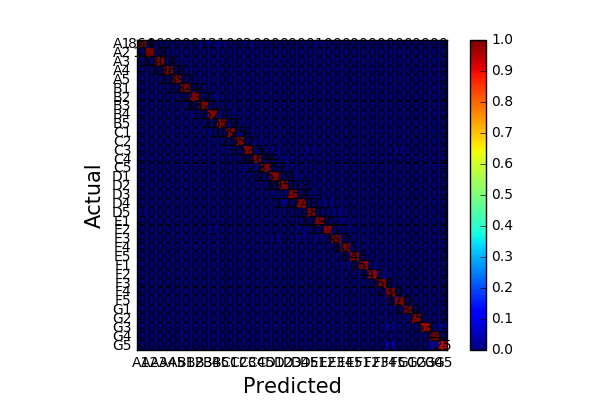
No model is considered true unless we validate it. But accuracy is not the best validator. Hence, we went for plotting confusion matrix. Also, we plot the seaborn heatmap.



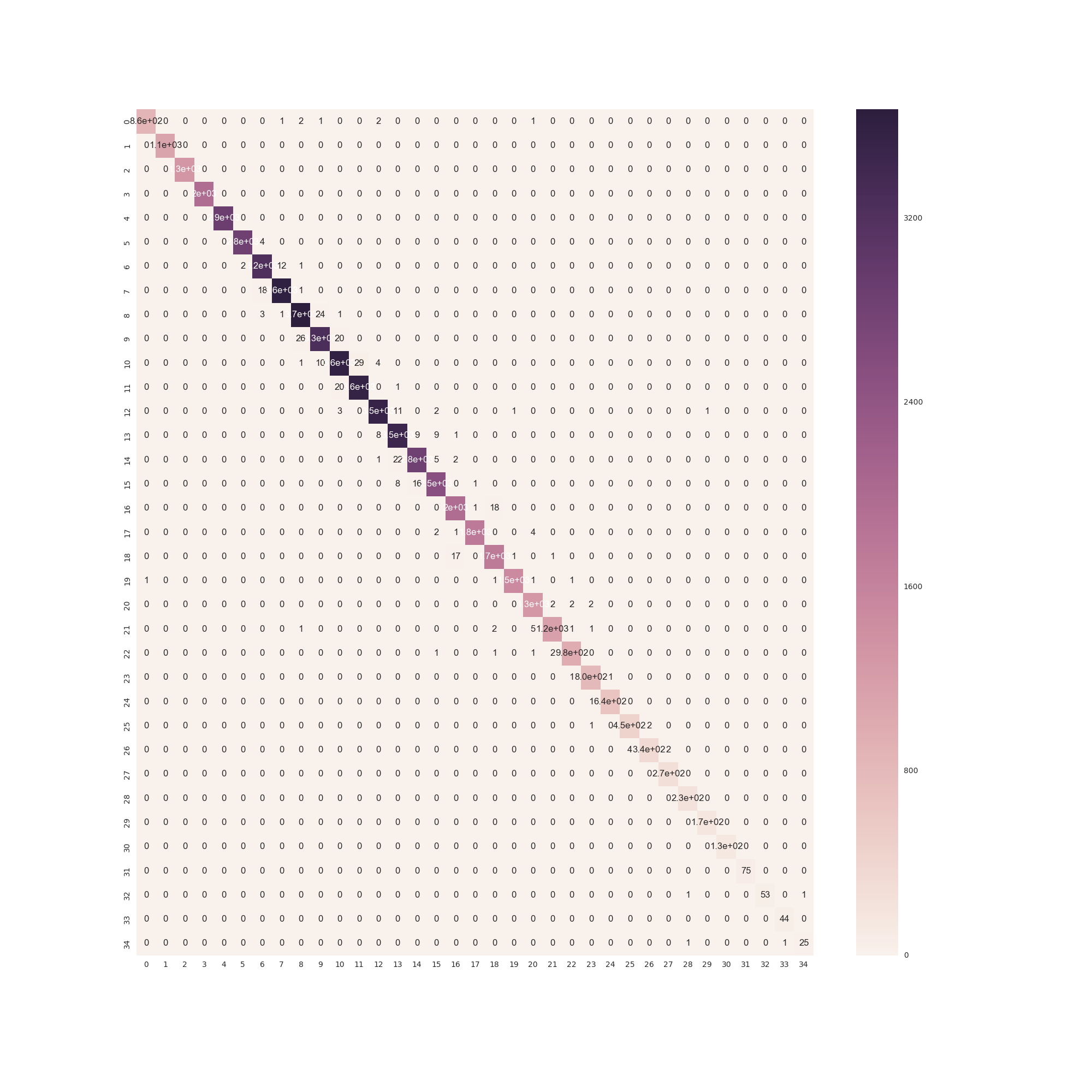
**Fig.d**

**4.3.1Confusion Matrix & Heat Map:**

As our dataset was huge, even after removing many null valued rows or columns, we got pretty high accuracy. In the confusion matrix below, the dark red diagonal and rest being blue is pretty close to a desired confusion matrix.

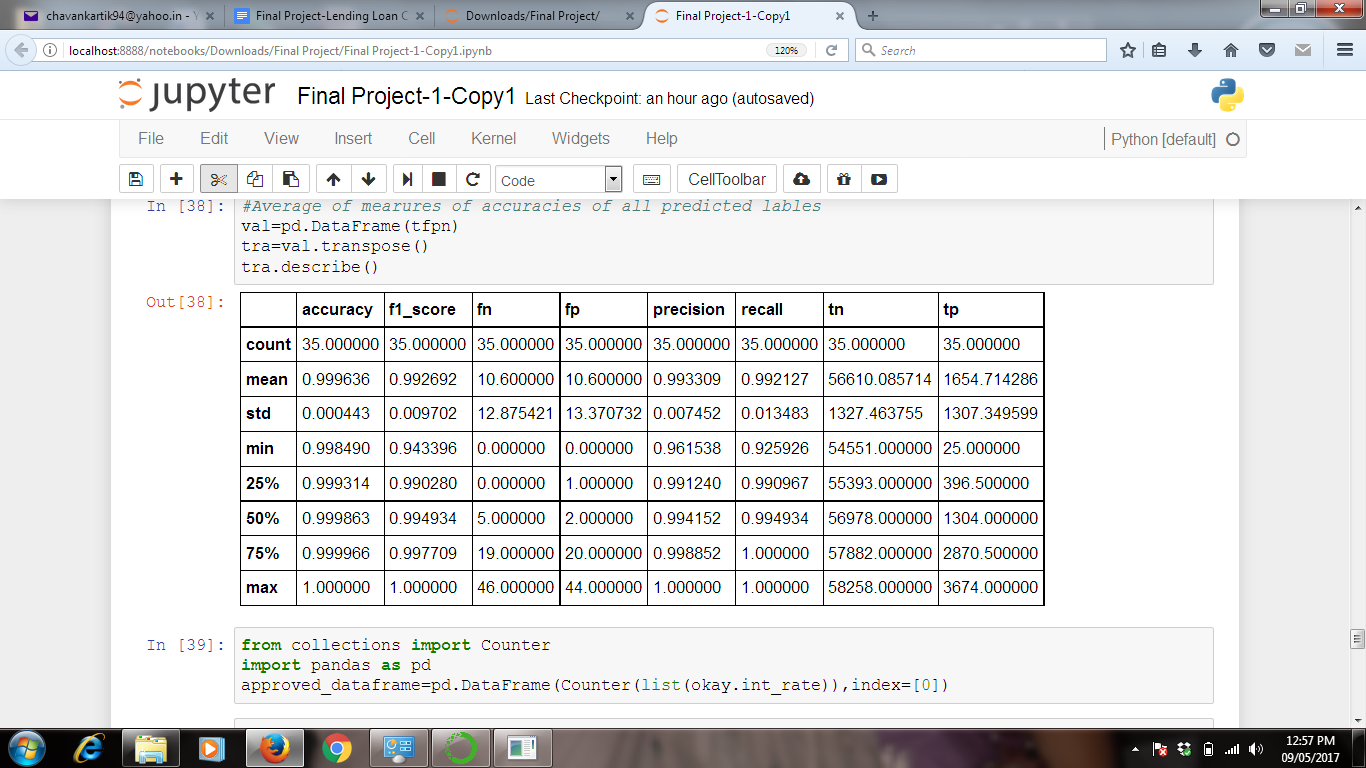


**Fig. e :Confusion Matrix**



**Fig. f : Heat Map**

We defined a class to find the True Positive, True Negative, False Positive, False Negative for each label in target variable, ie; sub\_grade. Our class also has functions to return the accuracy, precision, recall and f1\_score for each target value. We stored these values into a dictionary and stored as a data frame. Then we used describe function to get mean of all these variables. We found that DecisionTree with 80/20 split with our narrowed down columns gave the average accuracy about 99.96%, the average precision about 99.33%, the average recall about 99.21%, and the average f1\_score about 99.26%, as shown below-



**Fig .g**

We observed that, as the dataset increases, there is improvement in the model. Hence, for our huge dataset we found a pretty good model. Hence, we worked removing null values very carefully, by trial and error method.

Looking at these values, it proves the efficiency of the model. Hence, we went for final model building. Now, we used the whole dataset to build the model.

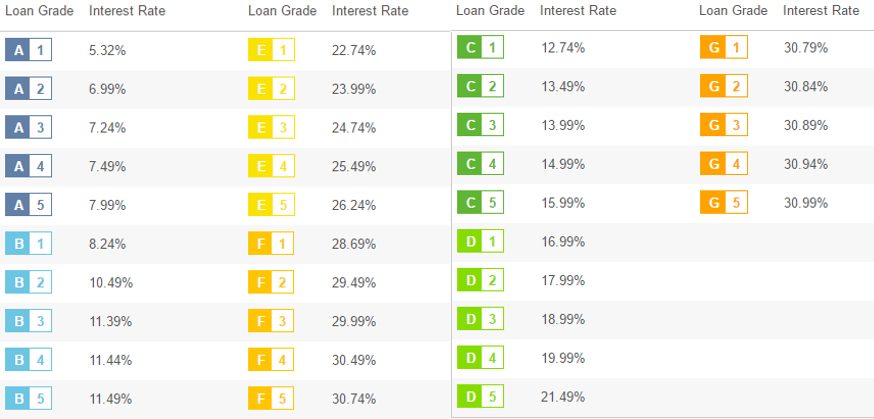
**4.4Interest Rate v/s Number of Approved Cases:**



**Fig. h : Interest Rate v/s Number Of approved Cases.**

**5.0 Business Implications:**

* Through this project it becomes easier to categorize the interest rate of all the sub grades. Once we classify the customer into sub\_gradesie; Loan Grades depending on the requirement of customer or his current historical data, using the table given below, we can say the interest rate the person would have to give.



**6.0 Lessons Learned:**

* Working on categorical data needs patience. Exploring such categorical data after converting into dummy variables can be challenge.
* Larger the data, better the accuracy.
* Handling null value is work of expertise. Sometimes, removing null values reduces the error drastically. Hence domain knowledge is important factor in handling null values.
* Always start early, be it selecting the project topic or working on the project.

**7.0 Future Scope:**

* We learnt that, even though these are approved loans, there 1.5% cases who have paid late fees or other charges. Which means, there is still scope of improvement in Selecting Creditworthy Borrowers. We have all plan, self-derived formulae on this, which we will implement after exams.
* Depending on current borrowers, and their late payment details, we plan to build a state wise loss prediction model.
* We intend to determining the attributes which affect the percentage of loan sanctioned, installment amount of loan sanctioned.

**8.0 References:**

1)[https://www.lendingclub.com](https://www.lendingclub.com/public/rates-and-fees.action)

2)<https://www.kaggle.com/>

3)<https://www.lendingclub.com/public/rates-and-fees.action>

4)[www.stackoverflow.com](http://www.stackoverflow.com)

5)[www.youtube.com](http://www.youtube.com)