****

**CAPSTONE PROJECT**

**INTEREST RATE PREDICTION**



**MENTOR -: Anjana Agrawal**

**SUBMITTED BY-:**

**Prem Chavan,Rupal Nikum,Atharva Shastri, Prakhar Choudhary,Ashish Bikkad,Chaitanya Thipse**

**PROBLEM DEFINITION**

Banks and lending firms offer various kinds of accounts and provide loans based on the requirements. Apart from it, there are other various activities like investments in market and different funds. Overall, the banking sector has a wide impact on the economy directly and indirectly.

There are many banks across the globe that are leveraging machine learning and AI in their daily routine and getting benefits out of it.

For example, top banks in the US like JPMorgan, Wells Fargo, Bank of America, City Bank and US banks are already using machine learning to provide various facilities to customers as well as for risk prevention and detection.

We know that banks have massive overheads, with thousands of employees to pay and hundreds of branches to maintain. To maintain profitability, banks must take large margins on the money that passes through them. Earning out of the difference in interests (what it pays to depositors and what it charges from borrowers) is the main source of revenue for any bank and has been the key element in the functioning of all traditional financial institutions.

We can use strategies which will permit these clubs maximise profits and minimize the risks involved in the sector.

**TECHNOLOGY USED**

Financial firms like banks and lending groups are generally free to determine the interest rate they will pay for deposits and charge for loans, but they must take the competition into account, as well as the market levels for numerous interest rates and policies.

There are many types of interest rates and loan products. When it comes to setting rates, certain loans, such as residential home mortgage loans, may not be based on the prime rate fixed by the firm but rather according to the policies set by the government.

These firms use an array of factors to set interest rates mainly to maximize profits for their shareholders. On the flip side, consumers and businesses seek the lowest rate possible. They start from client inputs, such as credit score, collateral provided, down payment and duration for the loan, employment status, assets owned and so on to calculate the optimum interest rates.**PROBLEM IMPORTANCE**



The data for the problem is an example of Peer to peer lending (or P2P lending) Club which is one of the most innovative financial products of recent times. It enables creditworthy borrowers lower their cost of loans and individual lenders/investors to lend directly to their peers and community thereby earning higher returns.

Lending clubs provide a virtual market place where borrowers and lenders can interact directly, without having to go through the traditional financial intermediaries like banks, who have become such behemoths in today’s time that they dictate all terms and conditions for both borrowers and lenders.

The project will use machine learning algorithms that leverage different determining factors of a loan applicant. Selection of significant factors will help develop a prediction algorithm which can estimate loan interest rates based on clients’ information. On one hand, knowing the factors will help consumers and borrowers to increase their credit worthiness and place themselves in a better position to negotiate for getting a lower interest rate. On the other hand, this will help lending companies to get an immediate fixed interest rate estimation based on client’s information. Here, our goal is to use a training dataset to predict the loan rate category (1 / 2 / 3) that will be assigned to each loan in our test set. We will use combination of the features in the dataset to make our loan rate category predictions.

**SUGGESTED SOLUTION**

So our problem is dealing with the identifying the customer according to their loan dispenising category in other words we can say that we have a to assign a label to the customer given various attributes. To overcome this problem we will be applying Supervise learning classification ML models on the given dataset. We'll be going through various steps for that as in the data set there our missing values so we'll be imputing them , we have to develop new features as well , compare the accuracy among different models and should come to conclusive model which will be predicting the best and giving a good accuracy.

STEP 5

Hyper Parameter Tuning of the best Model to further increase the accuracy

STEP 4

Base model and evaluation.

Testing various models and evaluation.

STEP 2

Feature Engineering:

Statistical analysis

STEP 1

EDA, Visualisations, missing value imputation, outlier treatment

STEP 3

Feature selection

Checking Balance Data

**DATASET INFORMATION**

The dataset had 164309 Rows and 14 columns in excel format. The data comprises of different features pertaining to various factors of every customer applying for loan.

**Data Dictionary**

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| Loan\_ID | A unique id for the loan. |
| Loan\_Amount\_Requested | The listed amount of the loan applied for by the borrower. |
| Length\_Employed | Employment length in years |
| Home\_Owner | The home ownership status provided by the borrower during registration. Values are: Rent, Own, Mortgage, Other. |
| Annual\_Income | The annual income provided by the borrower during registration. |
| Income\_Verified | Indicates if income was verified, not verified, or if the income source was verified |
| Purpose\_Of\_Loan | A category provided by the borrower for the loan request. |
| Debt\_To\_Income | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower’s self-reported monthly income. |
| Inquiries\_Last\_6Mo | The number of inquiries by creditors during the past 6 months. |
| Months\_Since\_Deliquency | The number of months since the borrower's last delinquency. |
| Number\_Open\_Accounts | The number of open credit lines in the borrower's credit file. |
| Total\_Accounts | The total number of credit lines currently in the borrower's credit file |
| Gender | Gender |
| Interest\_Rate | Target Variable: Interest Rate category (1/2/3) of the loan application |

- Variable categorization (count of numeric and categorical)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Numerical** | **Categorical** |  | **Missing Values** | |
|  |  |  |  | |
| Loan\_Amount\_Requested | Home\_Owner |  | Length\_Employed | 7371 |
| Length\_Employed | Income\_Verified |  | Home\_Owner | 25349 |
| Annual\_Income | Purpose\_Of\_Loan |  | Annual\_Income | 25102 |
| Debt\_To\_Income | Gender |  | Months\_Since\_Deliquency | 88379 |
| Number\_Open\_Accounts | Interest\_Rate |  |  |  |
| Months\_Since\_Deliquency | Inquiries\_Last\_6Mo |  |  | |
| Total\_Accounts |  |  |  | |
| **Total 7 Features** | **Total 6 Features** |  | **6.36% of total Data values** | |

**DATA CLEANING**

While conducting EDA the following discrepancies were found in the dataset:

1. Loan\_ID feature Is a unique id column that does don’t provide any type of insight not would help the machine learning model in prediction.
2. Loan\_Amount\_Requested is a feature that should be numerical but is object because of the commas present between the digits.
3. Length\_Employed also has special characters and strings instead of numerical values.

**FEATURE ENGINEERING**

Creating new features that might help our model predict more accurately

1. Accounts closed to Open accounts Ratio: Combining Total accounts and no. of open accounts to form a new feature which tells the no of accounts that have been closed by a client.
2. Assets or Liability: Categorise loans according to their purpose by segregating them according to whether the purpose of loan will help earn money in the future or not.
3. Financial Growth score: A new category that combines annual income and employment length of client, giving an idea of his/her financial growth over the years. (we will abstain from introducing this column as our annual income feature has lot of missing values, which will be imputed by us and this new feature might be biased towards our imputed value)

**MISSING VALUE IMPUTATION**

The dataset has a total of 6.36 % missing values. Which is not a suitable number to drop them, which will result in data loss. Hence we use standard null value imputation technique of fillna.

1. Home ownership null values replaced with new value created ‘NoHouse’.

2. Employment length null values impute with median.

### 3. Annual income null values filled we built the LinearRegression model and predict the Null value and imputed them

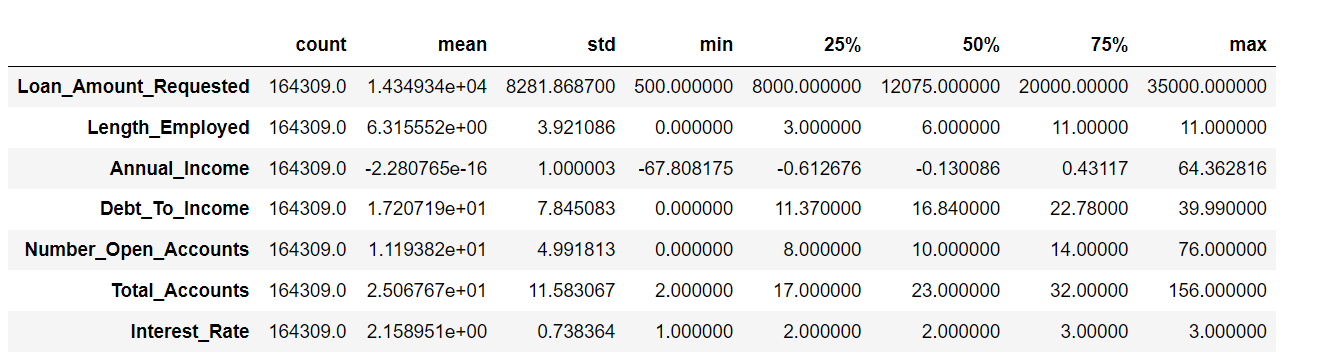
4. Months since Delinquency: as we have 53.55% Null value in this so we drop this variable

**ENCODING**

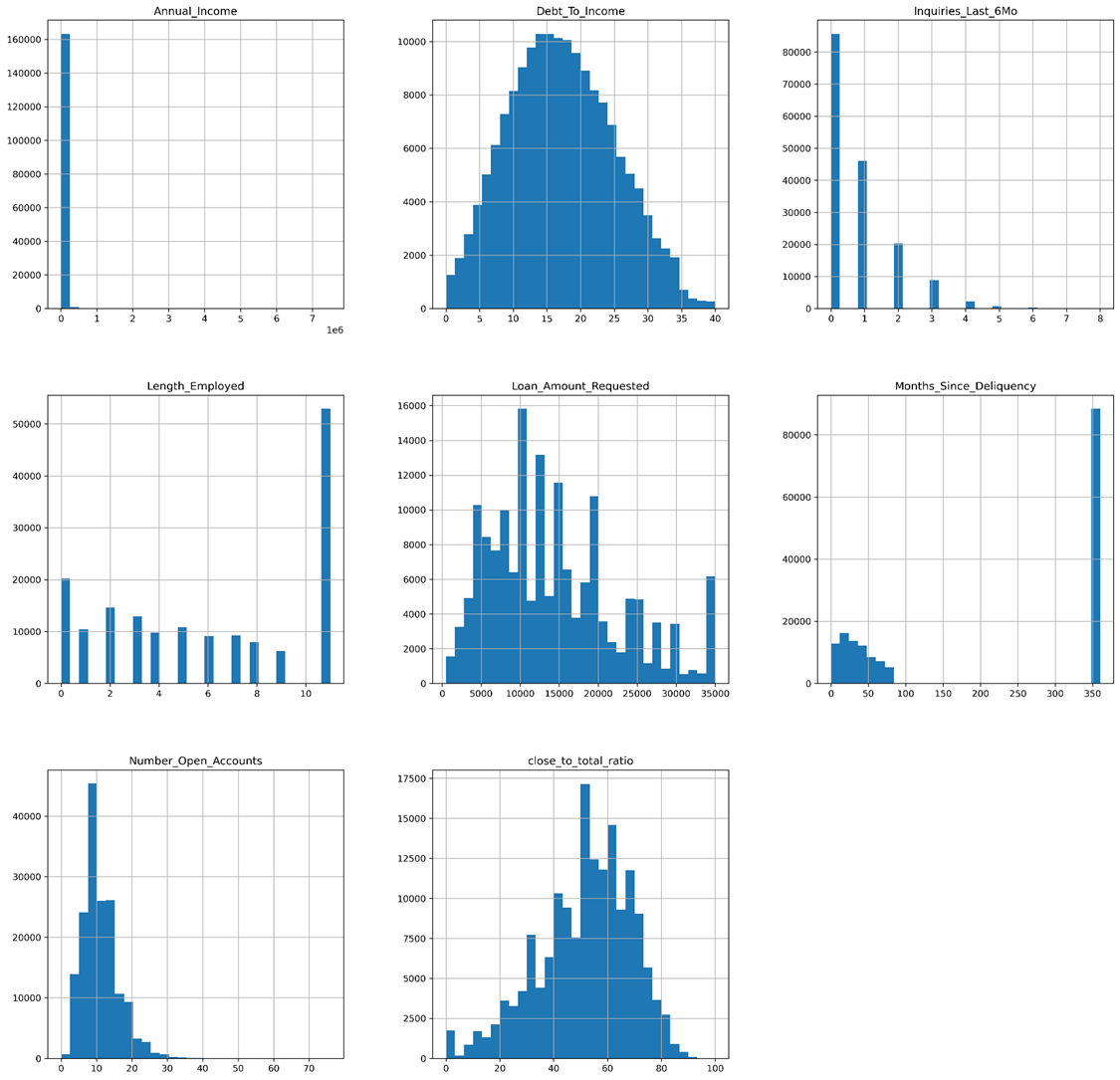
**One Hot Encoding:** The columns Inquiries\_Last\_6Mo ,Assets\_Liability, Home\_Owner, Purpose\_Of\_Loan and Gender will be one hot encoded using the pd.get\_dummies.

**DATA DISTRIBUTION**

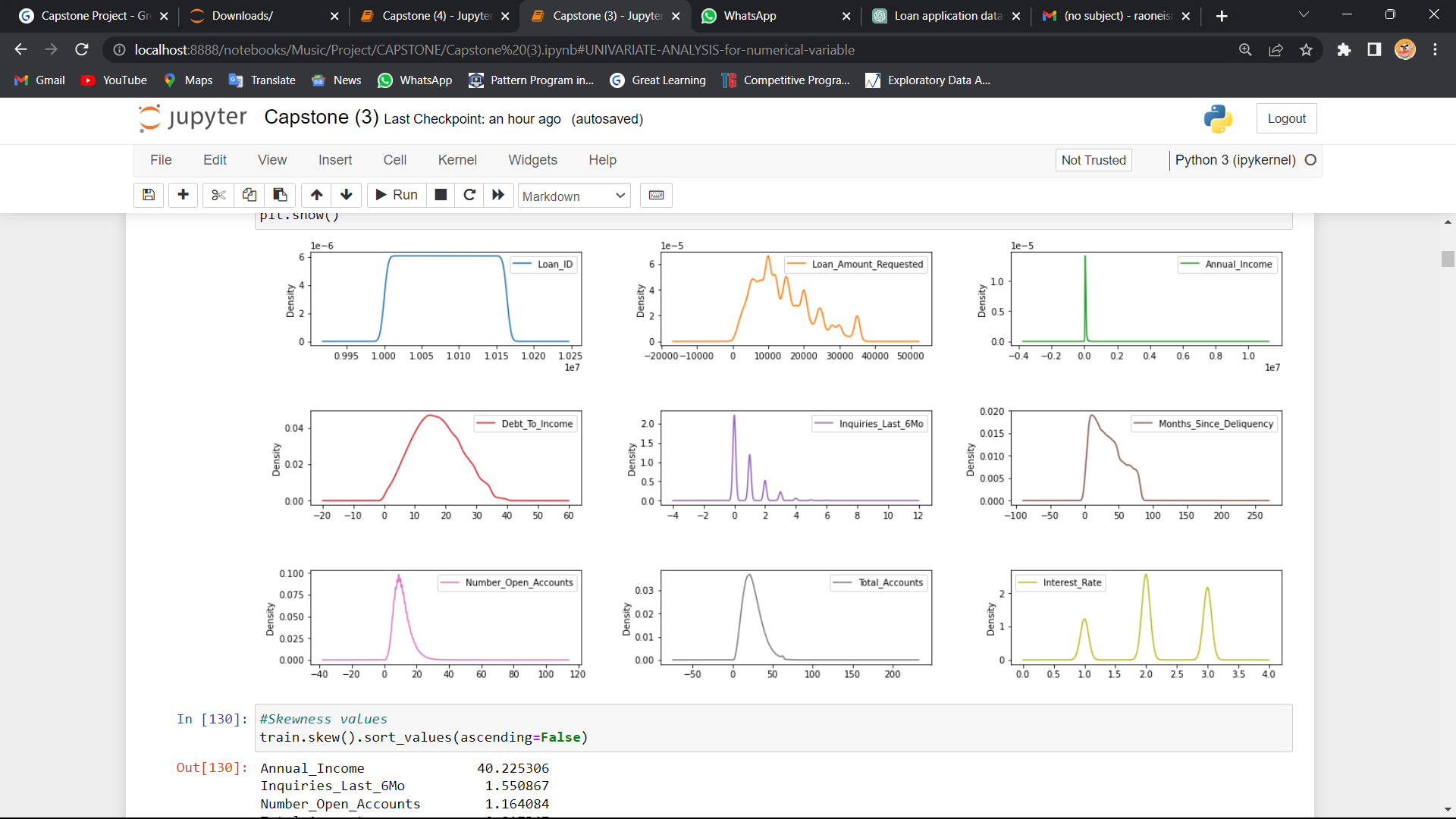
Using Data.describe() to create a 5 point summary of the data to get a better understanding of the numerical features in the dataset.



Using histograms plot in python to visualise data distribution.

****

# UNIVARIATE ANALYSIS

****

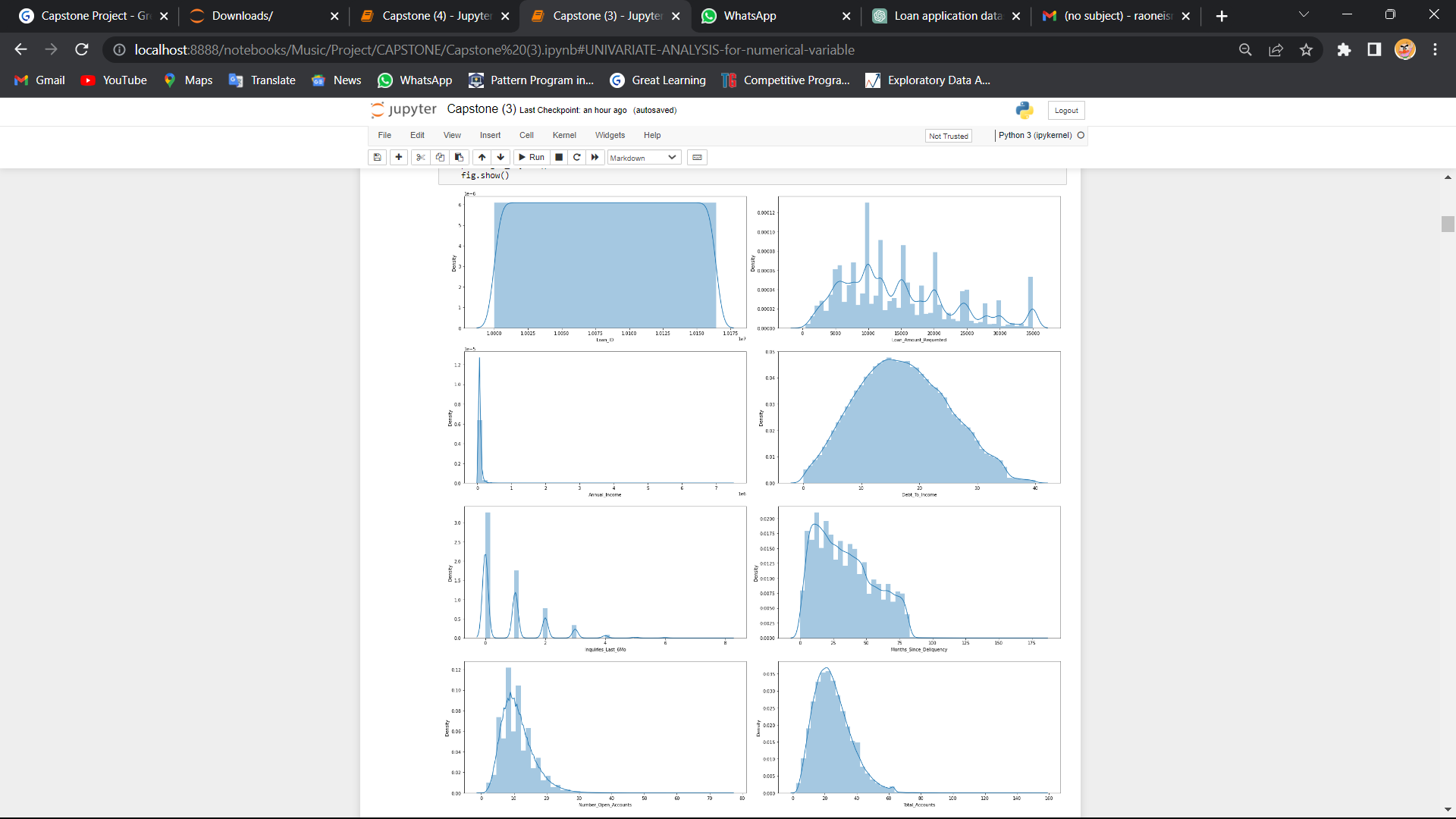
### inference -

'Annual\_Income' has a highly asymmetric distribution that is left-skewed. This means that the distribution has a longer tail towards the lower values and a majority of the observations are clustered towards the higher values.

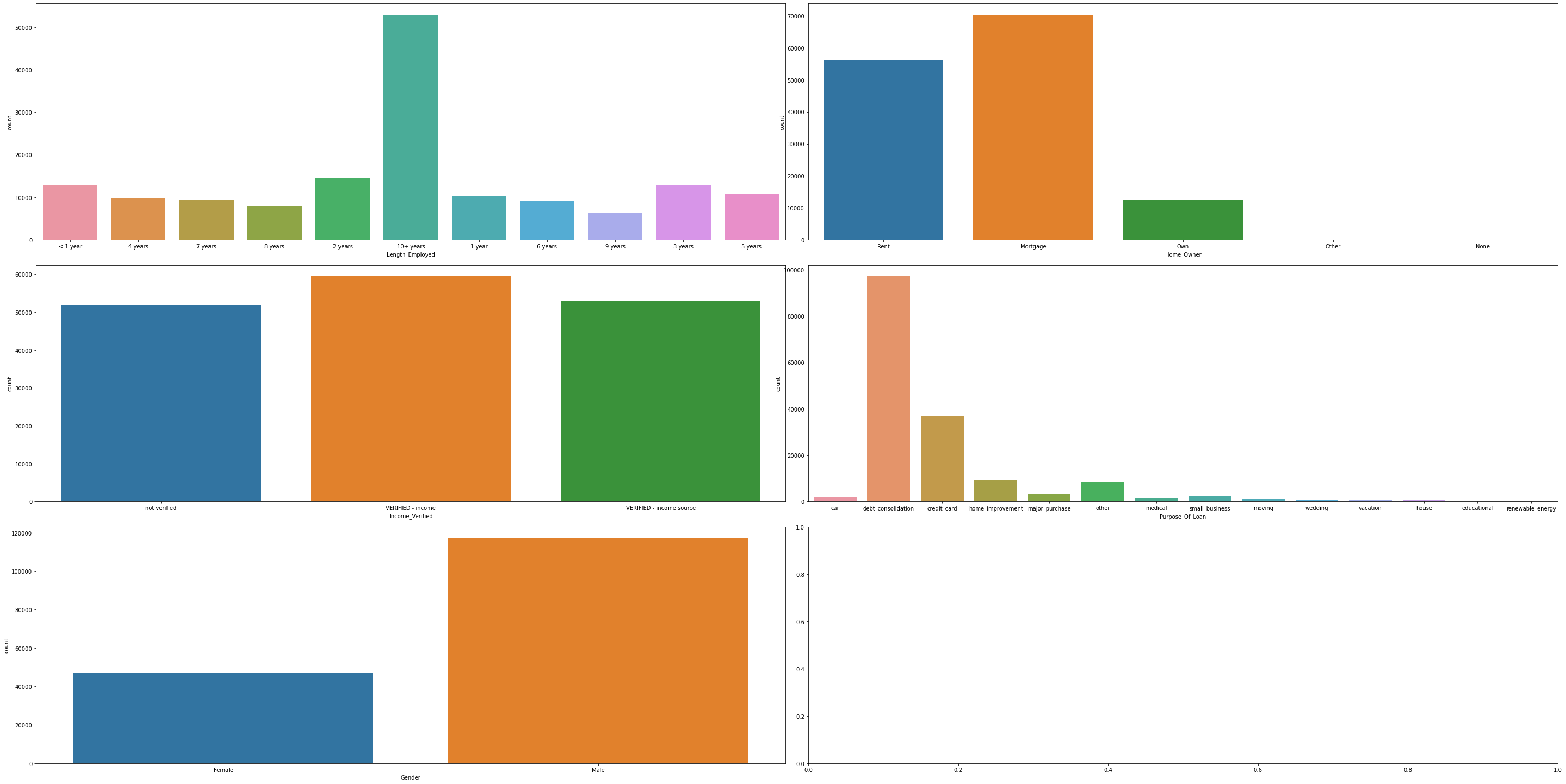
'Inquiries\_Last\_6Mo' and 'Number\_Open\_Accounts' are also left-skewed. This implies that the distribution has a longer tail towards the lower values and a majority of the observations are clustered towards the higher values.

'Interest\_Rate' and 'Loan\_Amount\_Requested' have a right-skewed distribution. This indicates that the distribution has a longer tail towards the higher values and a majority of the observations are clustered towards the lower values.

**Distriubution of Loan requested:-**

****

# Inference-



#Based on the information provided, we can infer the following:

#Among the different categories in 'Length\_Employed', '10+ years' has the highest number of entries.

#Majority of the entries in 'Home\_Owner' belong to the 'Mortgage' category.

#'VERIFIED - income' is the most common category in 'Income\_Verified'.

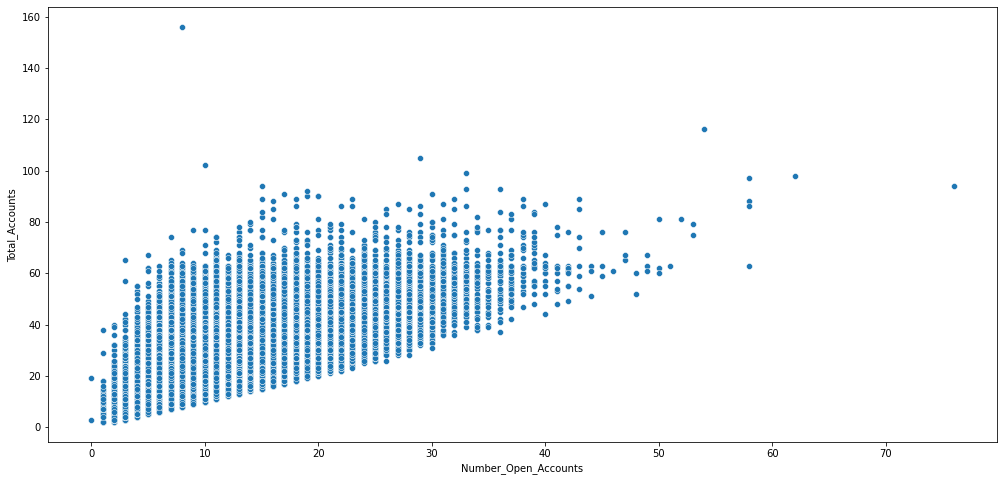
#'Debt consolidation' is the most common category in 'Purpose\_Of\_Loan', followed by 'credit\_card'.

#'Male' is the dominant gender class among the entries in the dataset.

#The dataset is divided into two parts, 'train' and 'test'. The 'train' part contains the data that will be used for training machine learning models, while the 'test' part contains the data that will be used for evaluating the performance of the models.

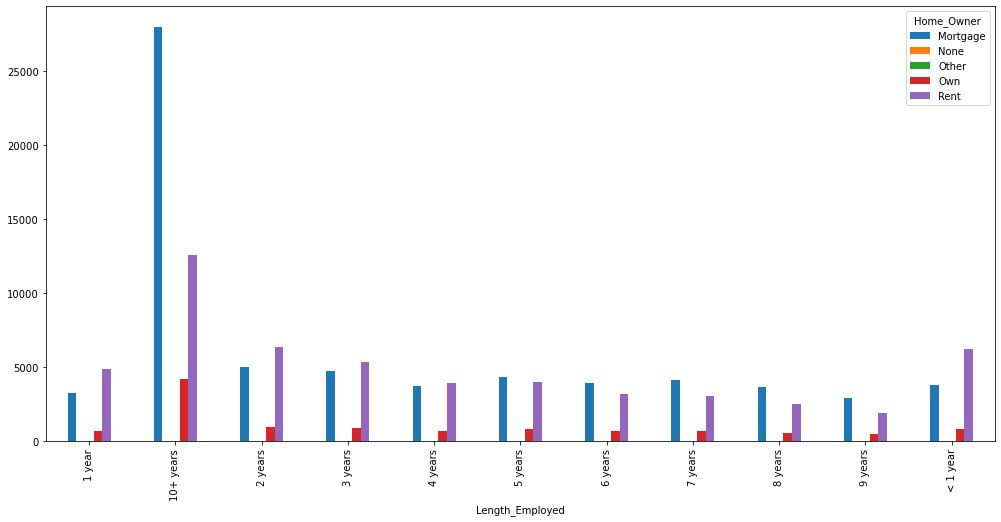
#These insights can be useful for feature engineering, exploratory data analysis, and modeling in the context of a loan prediction or credit scoring problem. For instance, knowing that debt consolidation and credit card are the most common loan purposes can help identify relevant features for prediction models. Similarly, knowing that males are the dominant gender class can help in identifying potential biases in the data and model evaluation.

# Bivariate analysis



### inference -

a positive correlation between 'Number\_Open\_Accounts' and 'Total\_Accounts' indicates that as the number of open accounts increases, the total number of accounts also tends to increase. This correlation makes intuitive sense because the total number of accounts is simply the sum of all open and closed accounts. Therefore, if the number of open accounts increases, the total number of accounts is likely to increase as well.

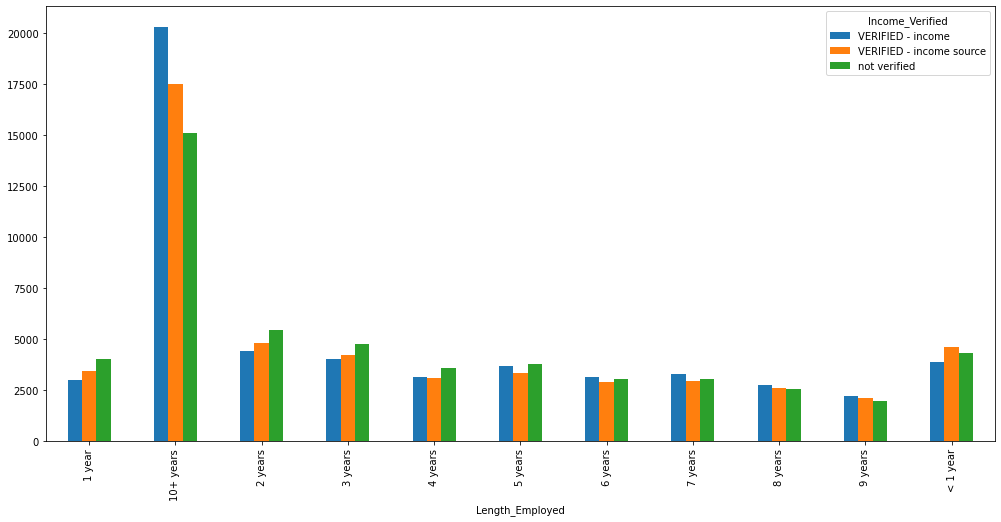
****

### inference -

The category '10+ years' is the most common value in the 'Length\_Employed' feature, indicating that a significant number of individuals in the dataset have been employed for more than 10 years.

Among the different categories in 'Home\_Owner', 'Mortgage' has the highest share, indicating that a majority of individuals in the dataset own a mortgaged property.

The category '9 years' has the lowest value in the 'Length\_Employed' feature, indicating that there are fewer individuals who have been employed for 9 years compared to the other categories.

****

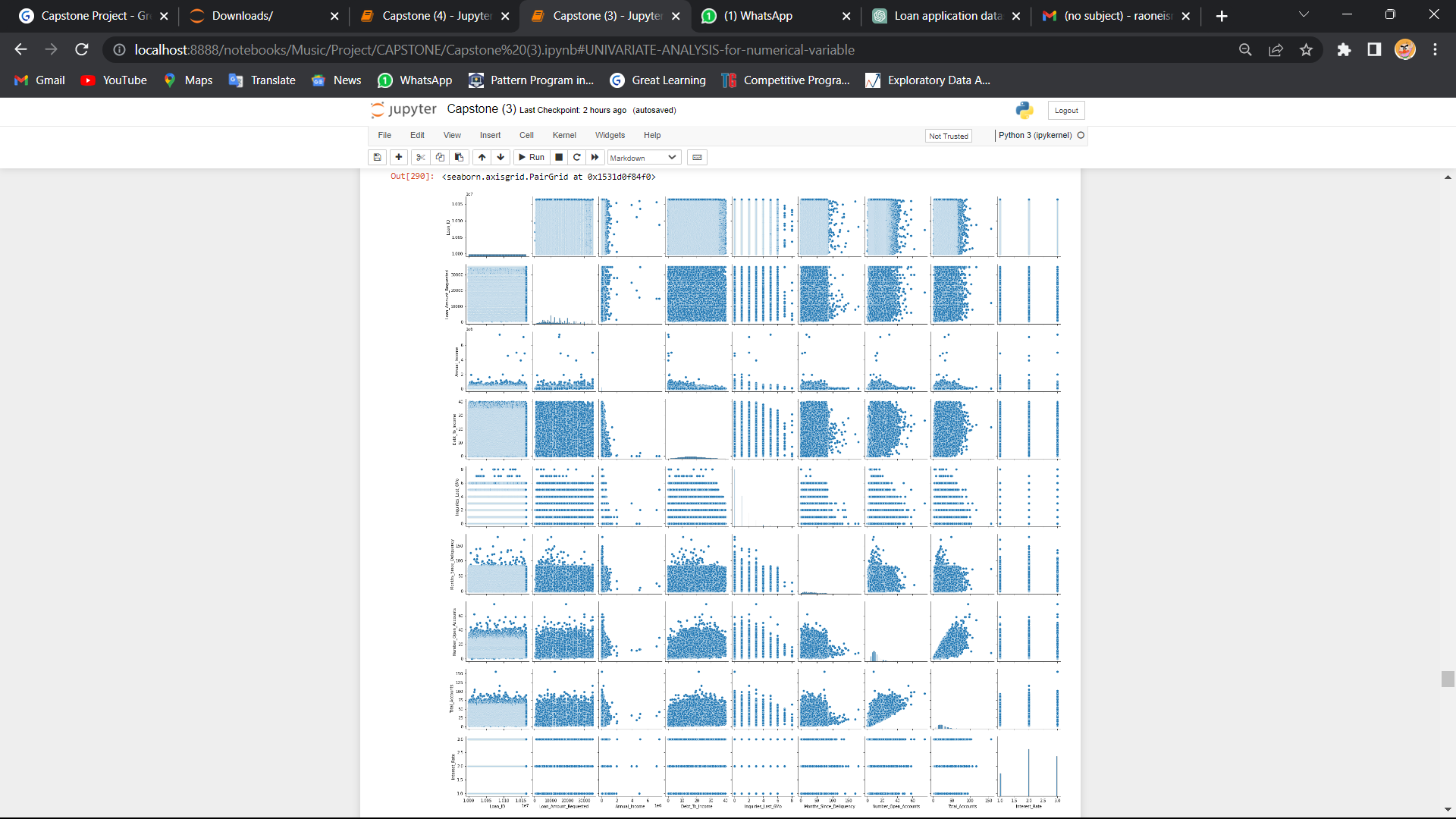
### inference -

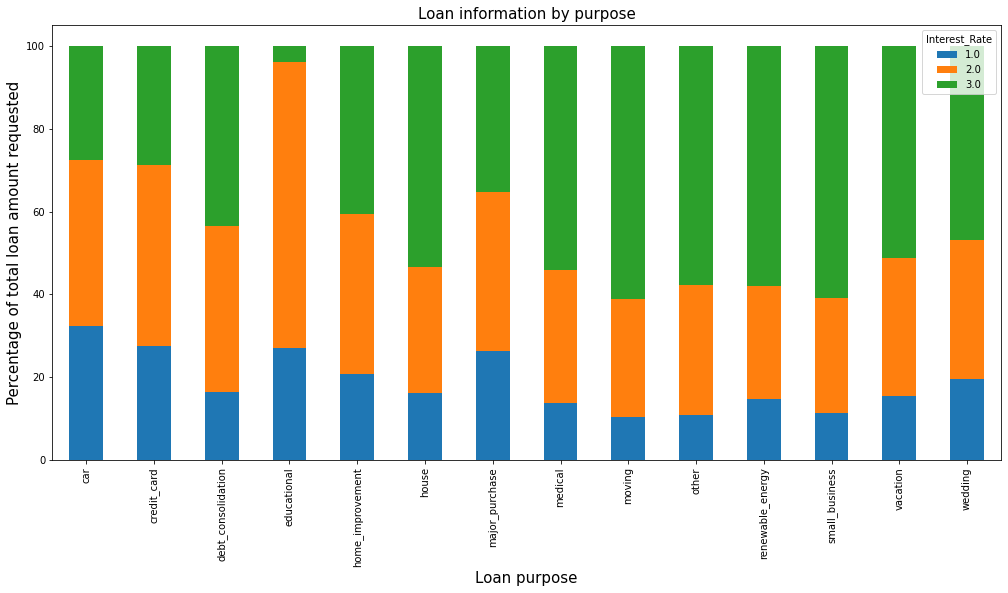
The category '10+ years' is the most common value in the 'Length\_Employed' feature, indicating that a significant number of individuals in the dataset have been employed for more than 10 years.

Among the different categories in 'Income\_Verified', 'VERIFIED - income' has the highest share, indicating that a majority of individuals in the dataset have their income verified.

The category '9 years' has the lowest value in the 'Length\_Employed' feature, indicating that there are fewer individuals who have been employed for 9 years compared to the other categories.

# Multivariate analysis

****

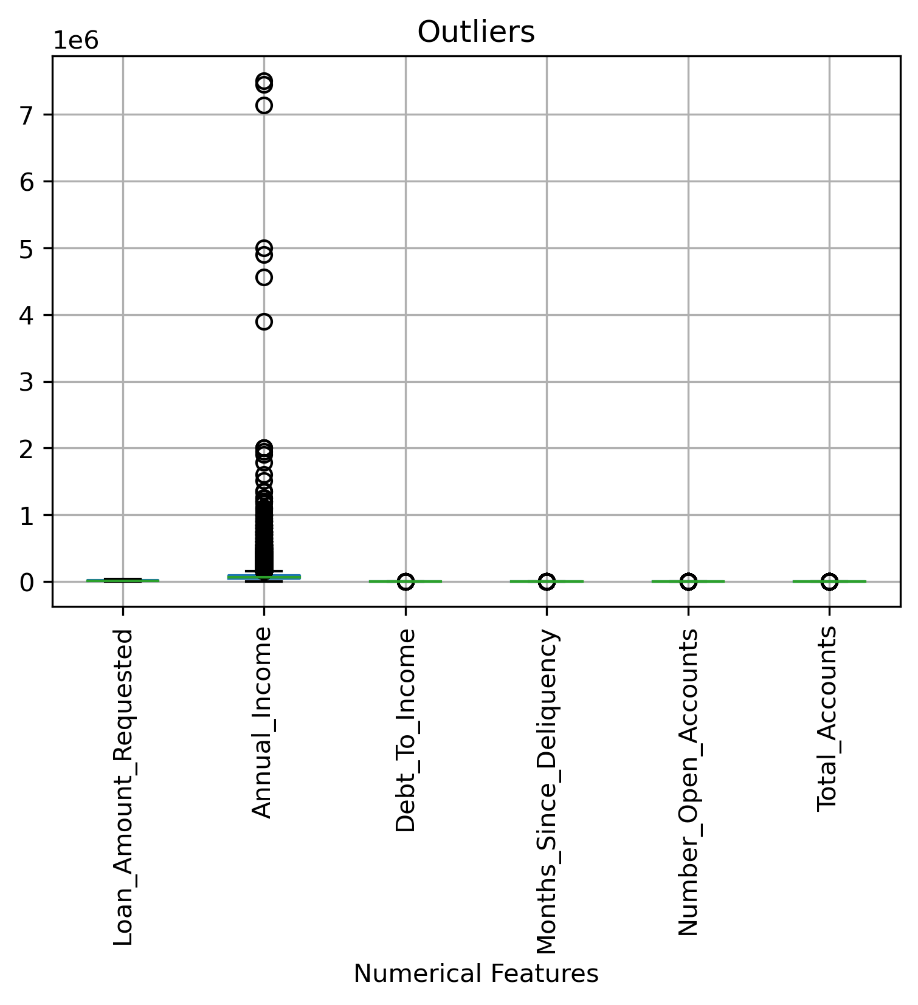
****

### Inference -

The stacked bar plot gives a clear indication of the distribution of interest rates across different loan purposes. It shows that certain loan purposes tend to have higher interest rates than others. This information can be useful for borrowers who want to make informed decisions about which loan to choose based on their financial goals and the purpose of the loan. The plot also highlights that educational and major\_purchase have lower interest rates, indicating that these loan purposes are less risky for lenders.

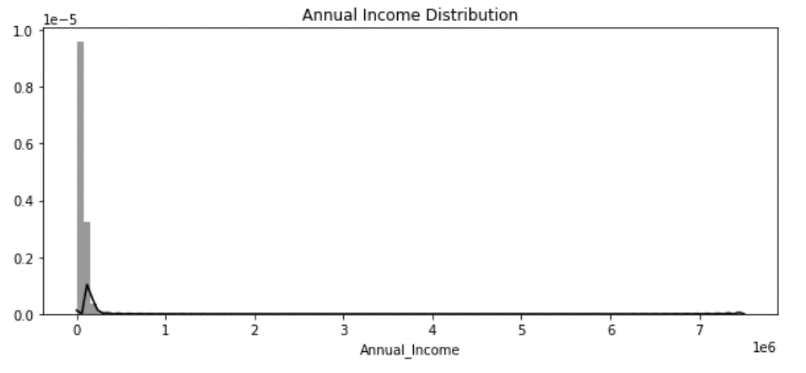
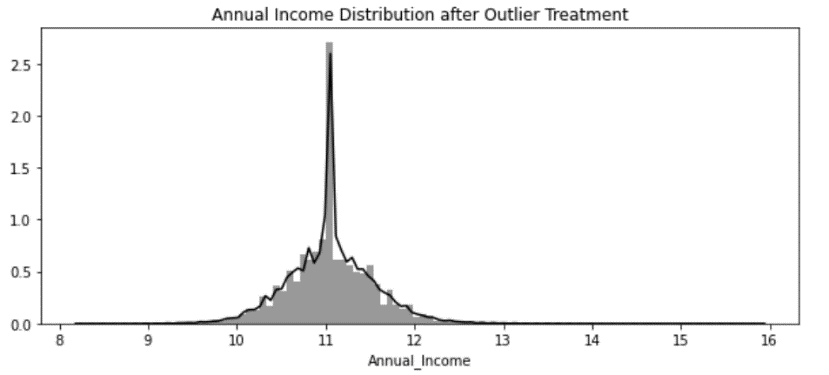
**OUTLIERS AND TREATMENT**

We use boxplots to visualise outliers present in the data.



Apparently the ‘Annual Income’ feature is having many outliers.

### Since the outliers are increasing the range of the data and the data is skewed, we perform Yeo Johnson Transformation to treat outliers.

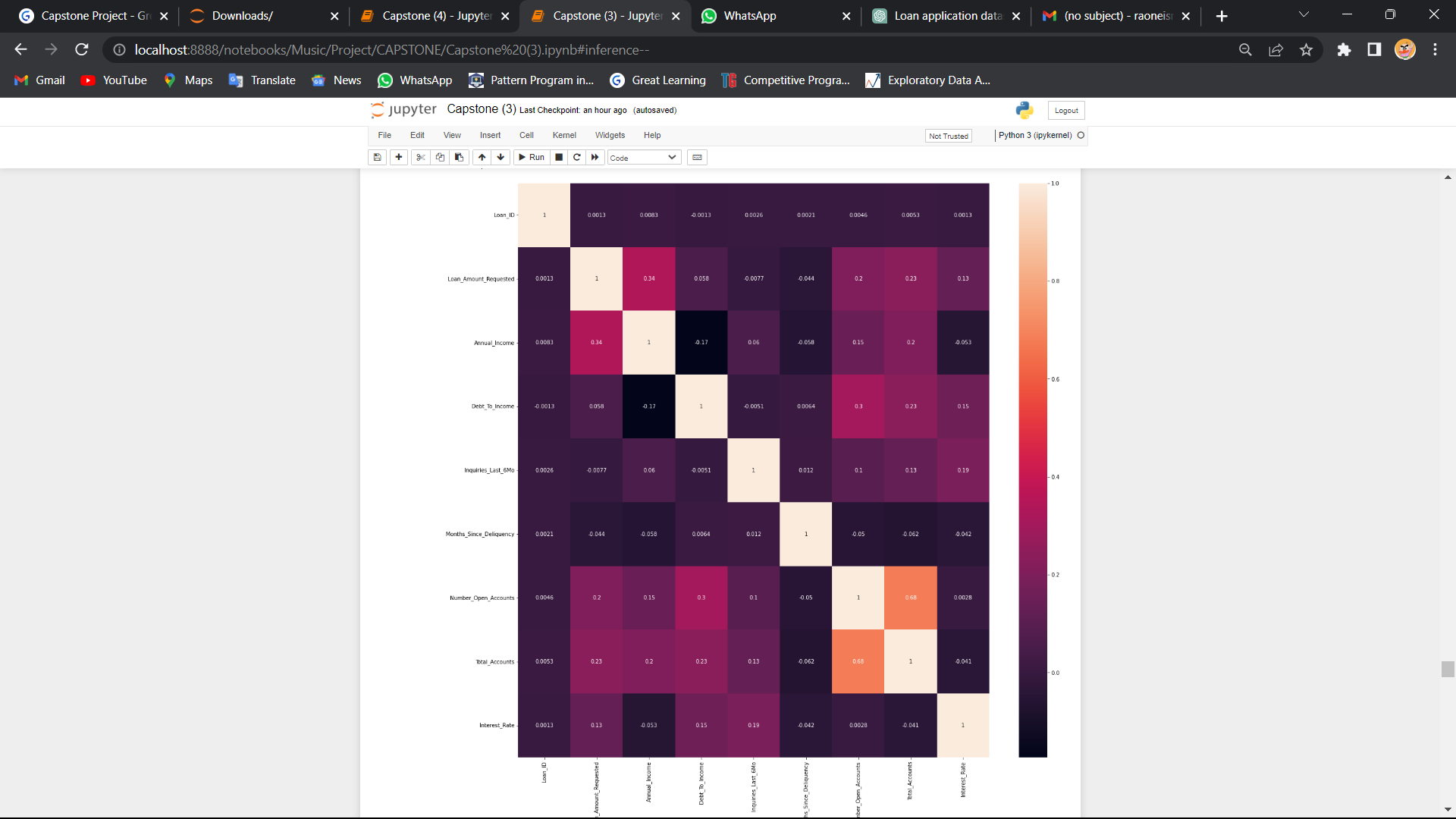
 

Before After

The histograms clearly show that now the data has less skewness and outliers as compared to earlier.

**FEATURE CORRELATION**

We use Pearson correlation to find correlation among features and plot them on a heatmap in Seaborn.

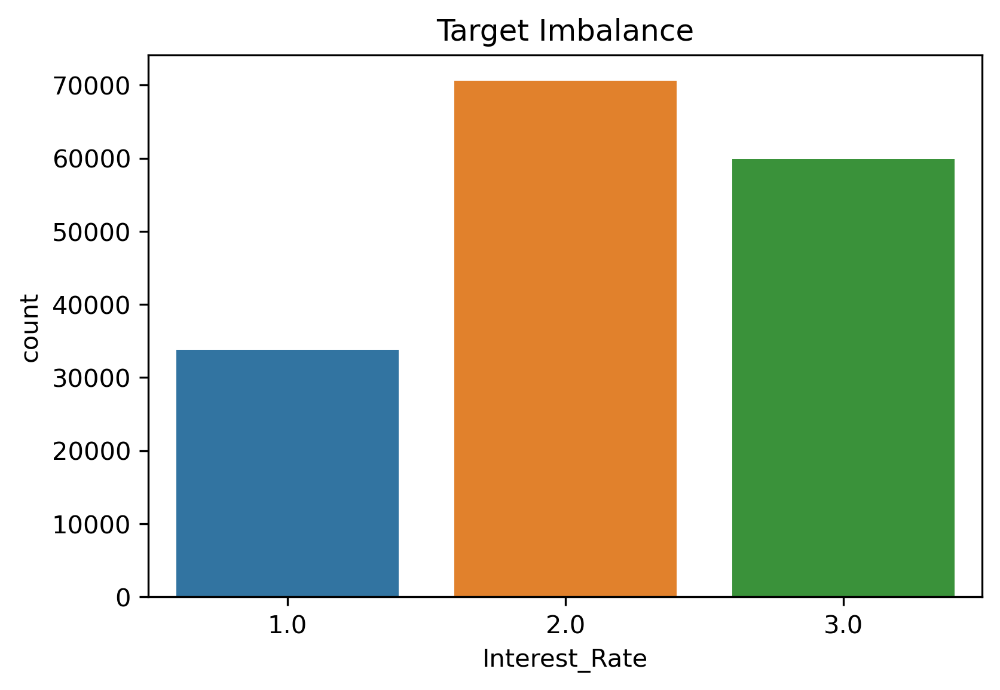


### inference -

there is a strong positive correlation between the loan amount requested and annual income, which suggests that people with higher annual income are more likely to apply for a higher loan amount. On the other hand, there is a weak correlation between the loan amount requested and inquiries last six months, which means that the number of inquiries made in the last six months does not have a significant impact on the loan amount requested. However, we cannot make any assumptions about causality or predictability based solely on the correlation values.

**CHECKING BALANCE OF DATA**

During our initial EDA it was evident that our Target is balanced.



|  |  |  |
| --- | --- | --- |
| Interest\_Rate | Count | Percentage |
| 1.0 | **33806** | **20.57%** |
| 2.0 | **70580** | **42.95%** |
| 3.0 | **59923** | **36.46%** |

### **Inference -**

Data is not imbalance so, there is no need to balance the data.

**BASE MODEL**

Since ours is a classification problem at first we will fit a base model to getter a rough idea of predictions.

Here we will use Logistic Regression algorithm with ‘multinomial’ argument under the multiclass parameter as we have more than two classes in the target.

The report is as follows:

**Classification Report**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target Class | **precision** | **recall** | **f1-score** | **support** |
|  |  |  |  |  |
| 1 | 0.45 | 0.57 | 0.56 | 21102 |
| 2 | 0.43 | 0.42 | 0.42 | 21308 |
| 3 | 0.48 | 0.48 | 0.48 | 21112 |
|  |  |  |  |  |
| **accuracy** |  |  | 0.45 | 63522 |
| **macro avg** | 0.49 | 0.49 | 0.49 | 63522 |
| **weighted avg** | 0.49 | 0.49 | 0.49 | 63522 |

Our model gave an overall accuracy of 45%.

Since our model accuracy is not up to our expectations we will go ahead

Algorithms considered:

1. Random Forest Classifier
2. Decision Tree
3. Naïve Bayes Classifier
4. KNN classifier
5. AdaBoost Classifier
6. Neural Network
7. SVM
8. Extra Trees
9. Light Gradient Boost
10. Ridge
11. Quadratic Discriminant
12. Linear Discriminant
13. Cat Boost

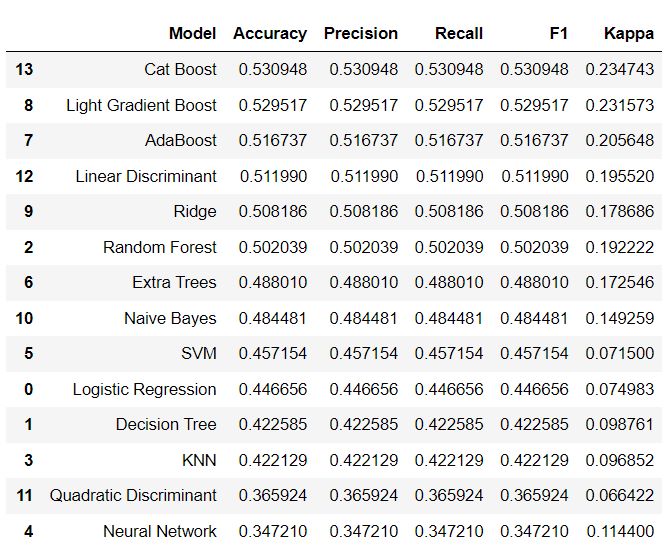
In order to determine the most effective algorithm for our purposes, we considered a range of options including Random Forest Classifier, Decision Tree, Naïve Bayes Classifier, KNN classifier, AdaBoost Classifier, Neural Network, SVM, Extra Trees, Light Gradient Boost, Ridge, Quadratic Discriminant, Linear Discriminant, and Cat Boost.

Rather than fitting each of these models individually, we chose to work with a Data Frame. We wrote a code that applies each of the models to the Data Frame and saves the results in a new Data Frame. To achieve this, we used a for loop to run each of the models on the Data Frame, allowing us to quickly compare the performance of each algorithm and identify the most effective approach. This approach allowed us to efficiently test a variety of models and determine the optimal solution for our needs.

**DataFrame**

**In the DataFrame we have Accuracy, Recall, Prec., F1\_score, Kappa And sorted by Accuracy**

**Model Comparison Table**

****

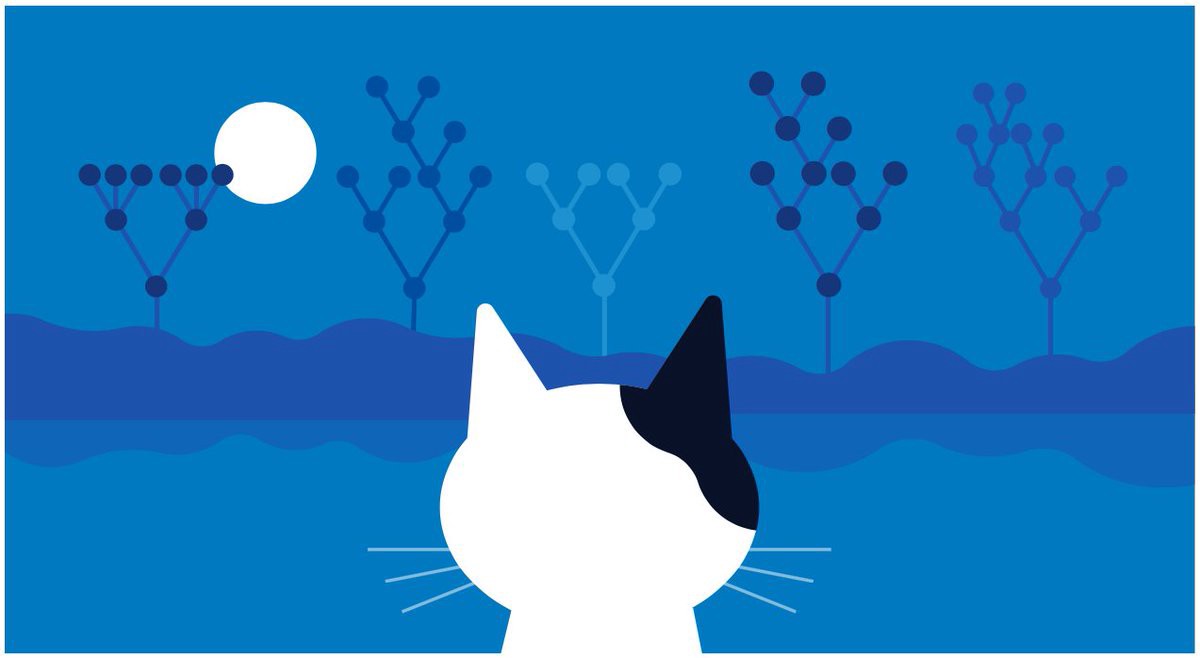
We found that the top three models are CatBoost Classifier, Gradient Boost Classifier and XGB Classifiers were giving the best accuracies.

So we decide to train these 4 models on our data, and the accuracies are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | CatBoost | Light  Gradient Boosting | Ada Boost | Linear Discriminant |
| Accuracy | 0.53 | 0.52 | 0.51 | 0.51 |
| Precision | 0.53 | 0.52 | 0.51 | 0.51 |

Hence it is clear that our final model will be CatBoost Classifier. Which gives accuracy score of 53% and f1 score of 51%.

**CatBoost**



CatBoost is based on gradient boosting. A new machine learning technique developed by Yandex that outperforms many existing boosting algorithms like XGBoost, Light GBM. The main difference being that CatBoost implements symmetric trees which helps in decreasing prediction time, which is extremely important for low latency environments.

CatBoost Process:

**Step 1:** Calculate residuals for each data point using a model that has been trained on all the other data points at that time. Hence we train different models to calculate residuals for different data points. In the end, we are calculating residuals for each data point that the corresponding model has never seen before.

**Step 2:** train the model by using the residuals of each data point as class labels.

**Step 3:** Repeat Step 1 & Step 2 (for n iterations).

So for big data sets its very complicated to do it for every data point Hence by default, instead of training different models for each data point, it trains only log(num\_of\_datapoints) models. Now if a model has been trained on n data points then that model is used to calculate residuals for the next n data points.

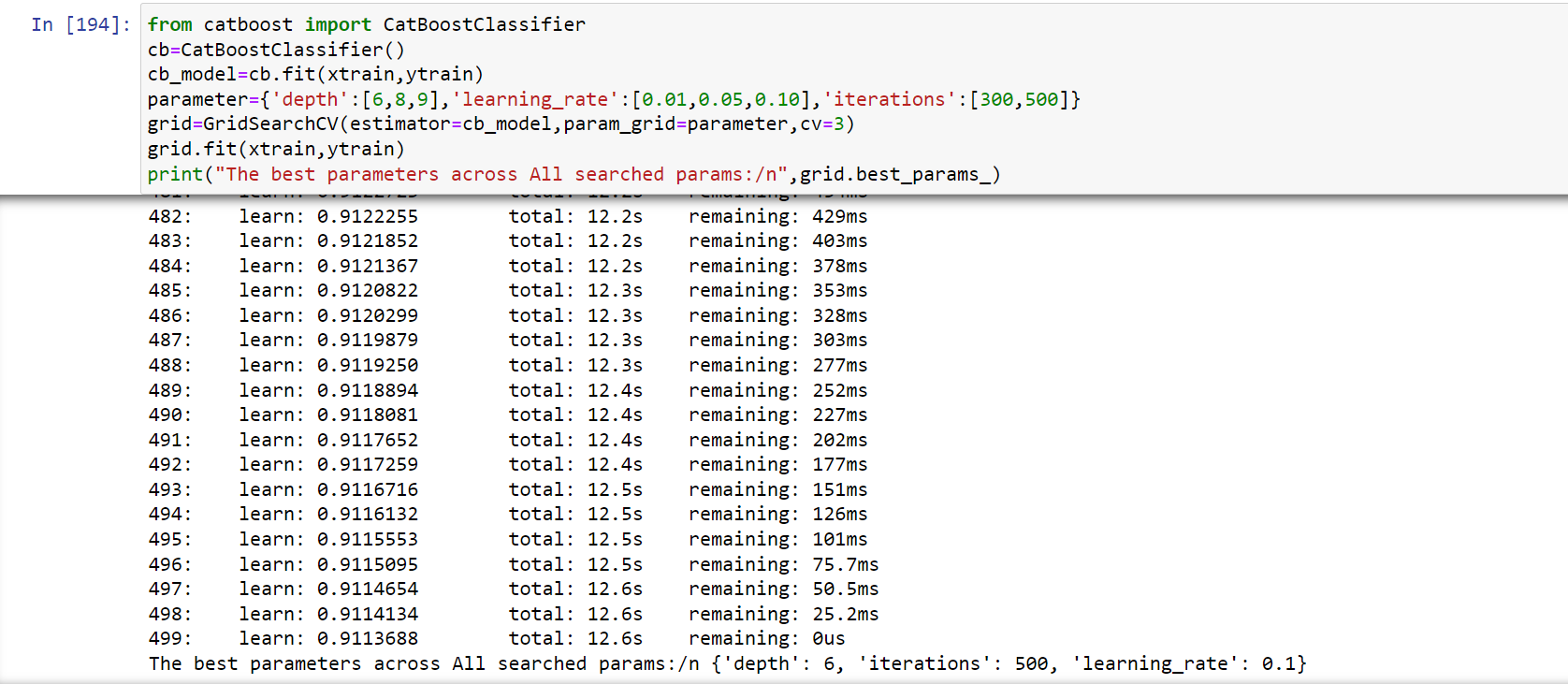
**Limitations**

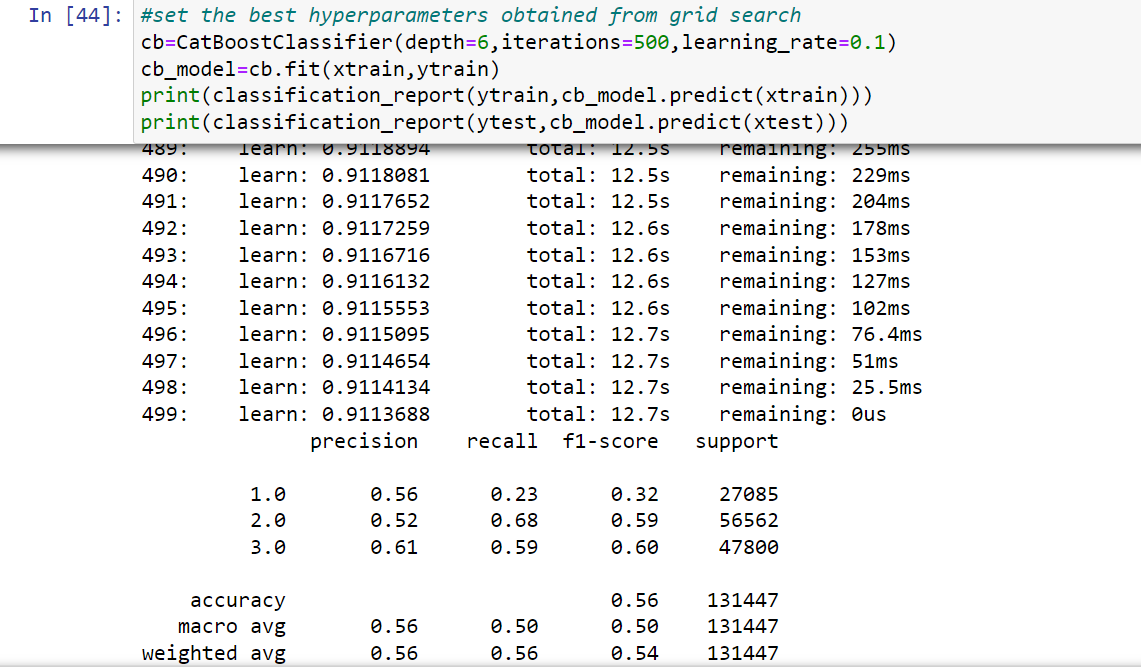
• CatBoost does not support sparse matrices.

• When the dataset has many numerical features, CatBoost takes more time to train than Light GBM.

**HYPERPARAMETER TUNING**

After using Randomised SearchCV we found that it was actually giving parameters which were not having any increase in the model performance, we use GridSearchCV which increased the accuracy and the Precision score to 0.5309, 0.5309 respectively.





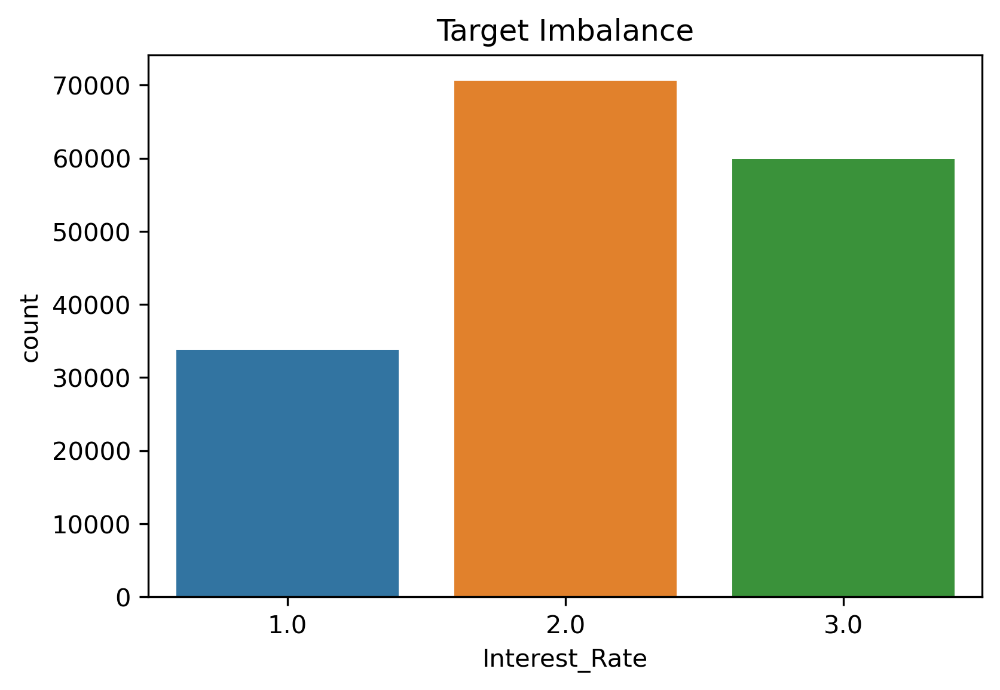
Our Final Model performance

Accuracy : 0.56

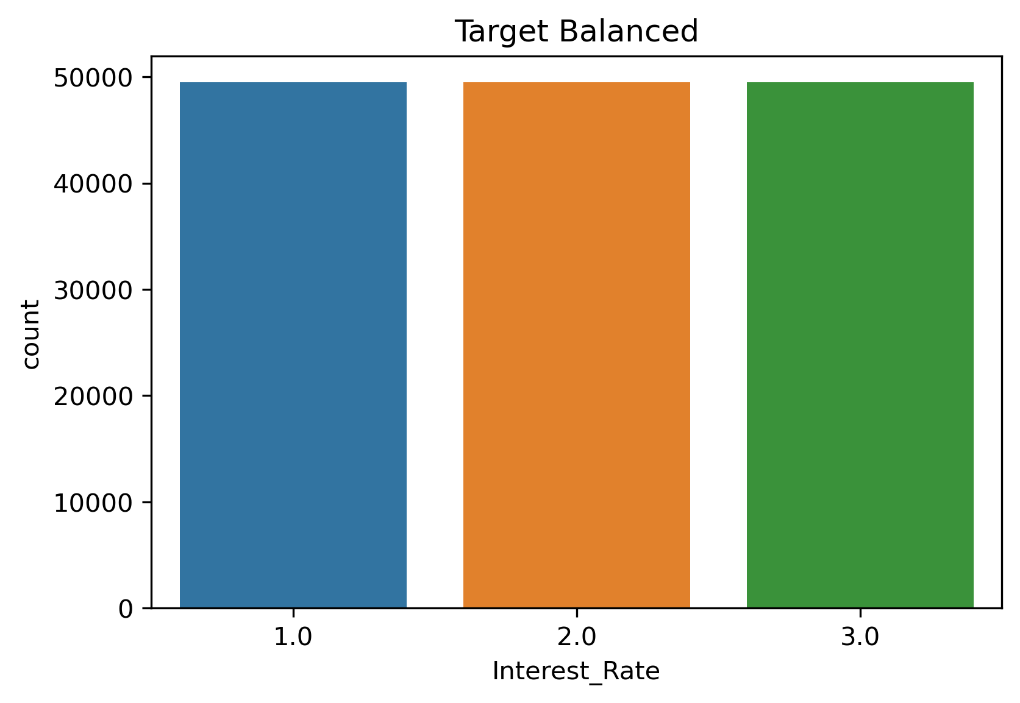
Precision:0.56 & Recall :0.23

**BALANCING OF DATA**

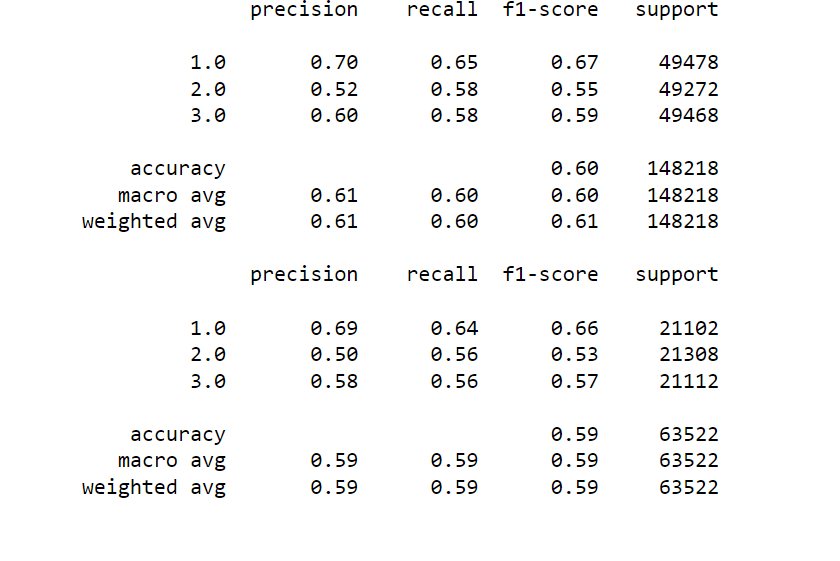
Just checking the final model and with smote model



Because of the cons related to simple undersampling and oversampling techniques we will use SMOTE technique to balance the target. Smote creates artificial data points near to minorities to balance the data.



Now it is evident that that our target is balanced and ready for training on models.



### Inference -

So include the result of smote at the end even though dataset has sufficient number of records for each class, smote has shown an improved performance.. Hence we require additional features to extract out the variability of the dependent variable

However, "we require additional features to extract out the variability of the dependent variable" implies that the existing features may not be capturing all the relevant information for predicting the dependent variable. This could be due to several reasons such as missing information, irrelevant or redundant features, or non-linear relationships between the features and the dependent variable. In such cases, it may be useful to explore alternative feature engineering techniques or consider collecting additional data to improve the predictive power of the models.

Overall, the improved performance of the models after applying SMOTE suggests that the dataset may have class imbalance, and addressing class imbalance can be an important step in improving the performance of the models. However, it is important to note that SMOTE alone may not be sufficient, and it is important to explore other techniques such as feature engineering, model selection, and hyperparameter tuning to further improve the performance of the models.

**BUSINESS APPLICATIONS**

Classifying the customer base on their past records we can provide them a rate of interest which they deserve and providing them benefits accordingly. This in turn can be used to improve the firm’s reputation and create a strong loyalty base.

In this project we can dividing our customers by assigning them different interest rates. This segmentation customers in 3 clusters will help in focusing on each cluster individually by means of creating policies, schemes and offers according to each specific cluster.

And in turn we can encourage or discourage shifting of customers from one segment to another.

Through the depiction of the customer segmentation through machine learning, Marketing team can apply various analysis over it and therefore they can plan there marketing scheme on that basis. For example:- If they observe there category 2 customers are more so then there marketing will be centric to them more compared to that of others.

Since the global financial crisis, risk management in banks has gained more prominence, and there has been a constant focus around how risks are being detected, measured, reported and managed. With the help of this banks can identify the customers who are good for their business. They can safeguard their interest by appropriate prediction of customers hence reducing the risk.

**CONCLUSION**

Although our predictive model did not cross our expectations in terms of performance, we were able to gain new insights about the factors which will help in better decision making and get an accuracy closer to the top performers. At the end we strongly feel that additions of features like credit score, lifestyle, savings etc could immensely help in better predictions, also lesser amount of missing data will further improve predictions.