PGP DSE-ONLINE BATCH DECEMBER 2020-21

Loan Status Prediction analysis

CAPSTONE PROJECT - FINAL REPORT

TEAM MENTOR- **VIDHYA KANNAIAH**

Group 7

Done by-

ASWIN.G

AAKASH GANDHI

PRASHANT RATHORE

ROHAN TAMBE

MOHANAKUMARAN.S

OVERVIEW: -

What is Customer validation for Loan Processing? –

The personal loan verification process includes validating all the details about an applicant including eligibility criteria, documents provided, repayment capacity, CIBIL score and more. A representative may visit your residence and your place of work to verify your details.

Business problem statement: -

1. Business Problem Understanding: - To Understand business/social opportunity If the bank is able to effectively predict the chance of loan default before the disbursement, then the future delinquency can be reduced significantly. This will help the bank to maintain good profitability and avoid any capital erosion.

2. Business Objective: - In order to make the process of providing loans effective, easy and risk free, customers who are meeting up certain criteria can take instant loans.

3. Approach: -

- Defining the problem statement and objectives –
- Cleaning and engineering of data
- Exploratory Data Analysis (EDA)-
- Initial modelling
- Iterations of steps 2/3/4 till satisfactory
- Results
- Analysis of findings and limitations

4. Conclusions: - Finally through a series of a analysis with the help of a number of algorithms we will be able to identify who are our target customers to whom loans can be sanctioned. This will help reduce the manual efforts for finding customers and the entire verification process.

TOPIC SURVEY IN BRIEF: -

1. Current solution to the problem: - Banks currently have a system of calling their customers directly or even running campaigns in various other campaigns in order to attract customers. This incurs a lot of financial loss for these banks in order to run these campaigns. Sometimes even some of these customers are unable to repay the loans in time which incurs further losses.

2. Proposed solution to the problem: - In our proposed solution we are trying to analyse that who are our potential customers who has the capacity to repay the loans based on various factors such as Loan amount, Annual Income, Purpose of loan, Debt to income etc. This will help us predict who are our target customers easily and hence saving a lot of money for advertising campaigns and other aspects.

CRITICAL ASSESSMENT OF TOPIC SURVEY: -

1. Find the key area, gaps identified in the topic survey where the project can add value to the customers and business: - Lending is generally considered risk-free as the bank might have done necessary due diligence including collateral requirement and credit score, re-payment. However, it has been recently witnessed that this segment has started to default, in turn impacting the revenue and profitability for the bank.

2. What key gaps are you trying to solve?

- Hence, there is a need to build a model to predict defaulters who are impacting the revenue and profitability of bank.
- Understanding business/social opportunity If the bank is able to effectively predict the chance of loan default before the disbursement, then the future delinquency can be reduced significantly. This will help the bank to maintain good profitability and avoid any capital erosion.
- Avoid certain customer.

Literature Survey-Past and undergoing research: -

- Try to predict loan defaulters using Logistic Regression models. Performance is measured using sensitivity and specificity. It tries to identify the right customers for granting loan.
- Default forecasting using data mining algorithms.

- Min-Max normalization with K-nearest neighbor classifier is used to create the credit score model for credit data of the customers.
- Clustering algorithms used to identify fraudulent activities. Quality of data is more crucial as it comes from varied sources. Principal feature analysis of financial data is implemented to extract relevant features, as the reduction of features will improve execution time without compromising the accuracy.
- Credit collectability prediction of debtor using dynamic k-Nearest neighbor using dynamic-k-nearest algorithm and distance and attribute weights.

Step-by-step walk through of the solution: -

a) Pre-processing

Initially the dataset is analysed with respect to the datatypes, null values and duplicates. Unwanted data are dropped. Loan amount requested is defined as object, and it is converted into float by removing the commas in the original value given in the dataset.

Null values are filled with appropriate values depending upon the nature of attributes. Count plot, hist plot is also used to identify the distribution of values for any particular attribute. Attribute Annual Income is filled by identifying a value 73000 after calculating the average annual income of individual which is grouped by Total number of experiences. On identifying the importance of values apt value is used for filling the missing data.

Relationship between related attributes is assessed using bar plot, scatterplot and pair plots. It is used to find the nature of correlation between different attributes.

Attributes with Categorical datatypes are label encoded to get numeric data. Outliers in the dataset is identified using box plots. Outliers in the dataset are removed using inter quartile range method.

b) Unsupervised Learning Models

Scaling of the data is done using Standard Scaler methods. Then model building starts with the base model K-means clustering technique. Within cluster sum of squares are realised using inertia function and it is used for finding the optimal number of clusters, WCSS is plotted with Scree plot and optimal number k=2 is identified. Silhouette Coefficient is also calculated and the cluster with the highest score which is nearest to 1 is finalised as the k value. Visualisation of bar graph indicates that k =2 is the best Silhouette score.

K means clustering with k =2 is implemented to group the given data in to two different clusters. Cluster size is calculated and it is shown using bar plot. Cluster 0 have 13120 observations and cluster 1 have 880 observations.

In order to improve the results PCA is used to for feature selection. For that Covariance matrix is calculated and that matrix is passed to get the Eigen values. PCA is applied on the all the attributes and pca explained variance is calculated and attributes with high variance is

assessed and it is used to finalise the number of significant features, in our project number of features selected is 5.

Now, k-means clustering is applied with the 5 principal components, and cluster number k =2. After clustering the count of observation in cluster 0 is 12535 and cluster 1 is 1465. The silhouette score for prediction is calculated it proves to be better now. Visualization of cluster data is also plotted using scatter plot by considering first 2 attributes.

We have implemented the pipeline concepts by creating three pipelines. One to perform pre-processing, second pipeline for performing clustering operation and the third one to execute both pre-processing and clustering operation. In pre-processing pipeline, we included scaling and pca operation. Pipelines are created and it is trained with train dataset and it is tested with test data. The prediction score is calculated with silhouette score. But, pipeline operation has not improved the prediction accuracy.

We tried to use DBscan clustering algorithm to group the observations in dataset, but the results showed the existence of more noise in the data and it is not able to perform grouping operation.

Then DBScan is implemented with principal components and it was able to create three cluster and noise observed when principal components is reduced to 8. The output was graphically represented. Prediction scores is assessed using silhouette, calinski harabasz and davies ouldin scores and all of them showed better results.

K-means algorithm is executed in mini batch mode. Hyper parameters used block size as 50, number of clusters = 4. Prediction scores is calculated after performing clustering operation.

We tried to use hierarchical clustering with Agglomerative Clustering technique with the pre-processed dataset, but since the size of dataset was relatively large, execution could not be completed. So, we tried to execute with the principal components calculated. ie. With 5 principal components.

The number of clusters here is decided based on the dendrogram, and plot reveals the optimal number of clusters as 2. Then we tried to implement Agglomerative Clustering on dataset with five principal attributes and cluster number as 2. It created 2 clusters with Cluster 0 having 11558 observations and Cluster 1 having 8442 observations. Visualisation and analysis of the cluster data reveals that cluster 1 can be labelled as loan can be approved and cluster 0 as loan cannot be approved.

4. Model evaluation

Since the dataset does not have labels, we used unsupervised learning algorithms. First Basic K-means algorithm is implemented with the pre-processed and scaled dataset.

K value is finalised with elbow curve using Within cluster sum of squares values plotted in Scree plot. Silhouette Coefficient is also used in order to confirm the optimal number of clusters to be used in K means algorithm. Then K means algorithm is implemented with the optimal hyper parameter's values.

The dataset used to 37 attributes and in order to find the best attributes PCA is implemented and optimal number of attributes is interpreted from Scree plot using explained_variance_ratio of all the attributes. The original set is reduced to 5 components. IT is again applied with k-means clustering algorithm which produced better results.

DBScan algorithm was implemented but it failed to produce better results and it created four clusters. Since, our problem requires to find whether loan can be approved or not, grouping the data into 4 clusters would not be beneficial.

K-means batch algorithm in implemented with batch size 50 and here also we used k as 4 So, this method did not produce good results.

Finally, hierarchical clustering is used with the scaled dataset. Here we implemented Agglomerative Clustering with five principal components. Linkage between datapoints is calculated. Optimal number cluster is decided with dendrogram which reveals k=2. It is most suitable for our problem. Then clustering algm created 2 clusters with improved cluster counts. On analysis of the cluster data cluster 1 represents individuals whose loan can be approved and cluster 0 represents individuals whose loan cannot be approved.

Among all the above clustering technique AgglomerativeClustering() could be considered as the best situable one for his project. as it gives more reasonable way of clustering

The attributes debit_to_income, Annual income, Length Employed > 1o years signals the strength to repay the loan and Months_Since_Deliquency which indicate the person who have failed to repay the loan in the past. These statistical data about these attributes in the ckuster are studied and analysed to make inferences.

In AgglomerativeClustering cluster 1 have high average values for debit_to_income, Annual income, Length Employed > 1o years and less value for Months_Since_Deliquency, So, the count of individuals in cluster 1 with 8442 observations can be approved and the count of individuals with 11558 observations in cluster 0 cannot be approved.

Describe the final model (or ensemble) in detail. What was the objective, what parameters were prominent, and how did you evaluate the success of your models(s)? A convincing explanation of the robustness of your solution will go a long way to supporting your solution.

.

5. Comparison to benchmark

How does your final solution compare to the benchmark you laid out at the outset? Did you improve on the benchmark? Why or why not?

6. Visualization(s)

In addition to quantifying your model and the solution, please include all relevant visualizations that support the ideas/insights that you gleaned from the data.

7. Implications

How does your solution affect the problem in the domain or business? What recommendations would you make, and with what level of confidence?

8. Limitations

What are the limitations of your solution? Where does your model fall short in the real world? What can you do to enhance the solution?

DATASET AND DOMAIN: -

Data Dictionary

1.LOAN_ID = LOAN ID OF PERSON WHO REQUESTED FOR A LOAN.

2.LOAN_AMOUNT_REQUESTED = AMOUNT OF LOAN IS REQUESTED FOR DIFFERENT PURPOSE.

3.LENGTH_EMPLOYED = FOR HOW MUCH PERIOD OF TIME THE PERSON IS WORKING.

4.HOME OWNER = PERSONS TAKING LOAN, WHETHER THEY HAVE THEIR OWN HOME, ON RENT, ON MORTGAGE.

5.ANNUAL INCOME = DESCRIBES THE ANNUAL INCOME OF PERSON.

6. INCOME VERIFIED = TO THE PERSONS WHO ARE TAKING LOANS WHETHER THEY HAVE VERIFIED INCOME OR NON-VERIFIED INCOME.

7.PURPOSE OF LOAN = PEOPLE DEMANDING LOAN FOR DIFFERENT PURPOSES SUCH AS DEBT CONSOLIDATION, CREDIT CARD OR FOR MEDICAL PURPOSE AND OTHER.

8. DEBT TO INCOME = IT MEASURES THE PERSON ABILITY TO MANAGE THE MONTHLY PAYMENTS TO REPAY THE MONEY HE/SHE PLANS TO BORROW AS A LOAN.

9. INQUIRIES LAST 6 MONTHS = NUMBER OF INQUIRIES FOR A PERSON IN LAST 6 MONTHS

10. MONTHS SINCE DELIQUENCY =

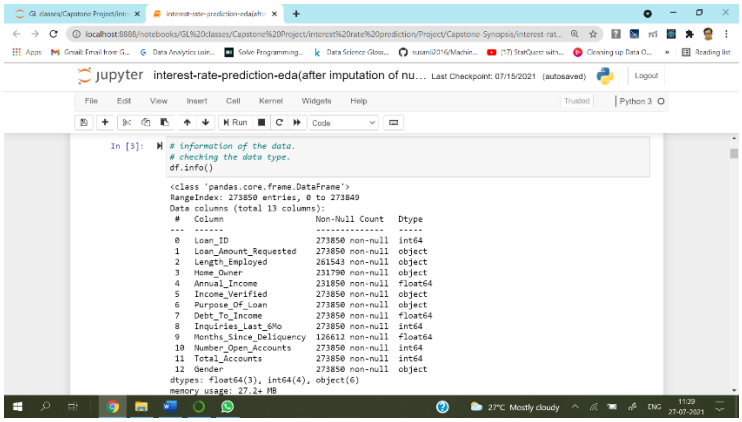
11.NUMBER OF OPEN ACCOUNTS = DESCRIBES THE NUMBER OF ACCOUNTS OPEN FOR A PERSON TAKING LOAN.

12. TOTAL ACCOUNTS = NUMBER OF TOTAL ACCOUNTS A PERSON HAVE WHO REQUESTED FOR A LOAN

13. GENDER = WHETHER THE PERSON IS MALE OR FEMALE.

14. INTEREST RATE = ON WHAT RATE THE LOAN IS GIVING TO PERSON.

1- Variable Categorization (count of numeric and categorical): - We have total 13 variables in dataset out of which 6 variables are categorical and remaining 7 are numerical.

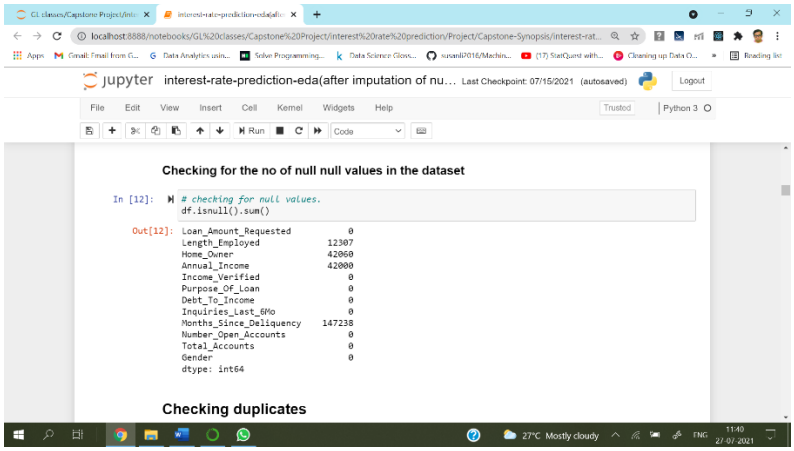


There are 273850 records in the dataset.

2- Pre-Processing Data Analysis (count of missing/null values, redundant column): -

We have null values in as follows:

LENGTH_EMPLOYED -	12307
HOME_OWNER -	42060
ANNUAL_INCOME -	42000
MONTHS_SINCE_DELIQUENCY -	147238



Project Justification: -loan prediction works correctly and fulfils all requirements of bankers to verify and approve or reject the applicant based on the documents provided and other factors. This system properly and accurately calculates the result. It predicts the loan is approved or rejected to the applicant very accurately. It also helps to detect the scams which are happening for e.g. (Fake documents) which are provided by applicant to get loan.

DATA EXPLORATION(EDA): -

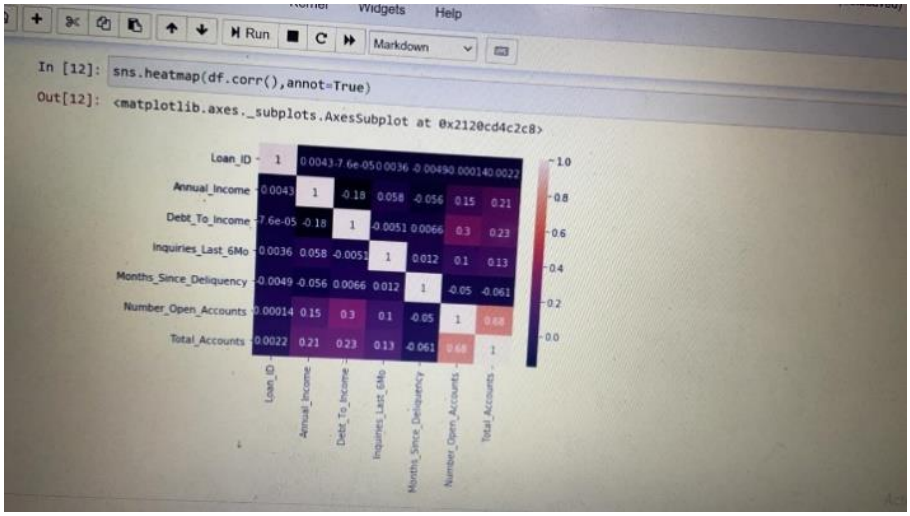
Between Variables: - By checking the correlation between the variables we get to know that

I. ANNUAL_INCOME and debt to Income are negatively correlated which means if ANNUAL_INCOME will increase DEBT_TO_INCOME will decrease.

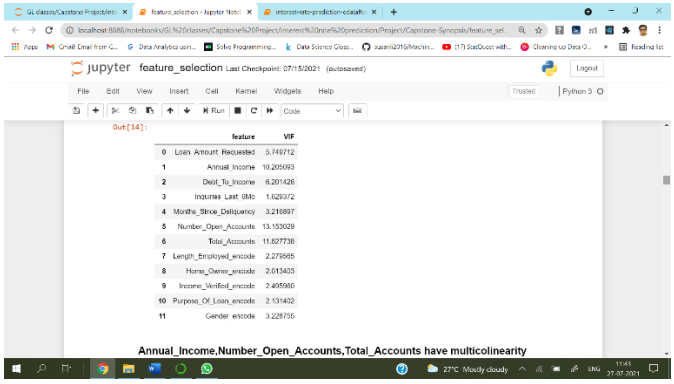
II. TOTAL_ACCOUNTS and NUMBER_OPEN_ACCOUNTS have highly correlated relationship between them.

III. INQUIRIES_LAST_6MO and NUMBER_OPEN_ACCOUNTS has very less correlation between them

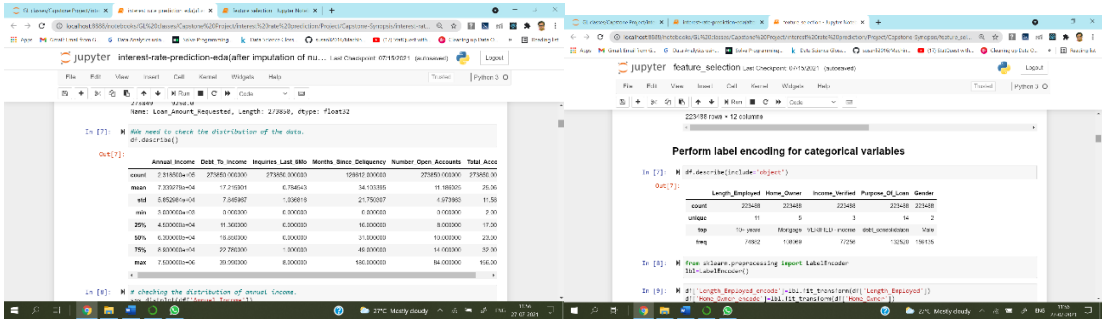
IV. NUMBER_OPEN_ACCOUNTS and TOTAL_ACCOUNTS are highly correlated.



1- MULTI-COLLINEARITY - Through VIF FACTOR we have checked the Multicollinearity and we get to know that ANNUAL_INCOME, NUMBER_OPEN_ACCOUNTS, TOTAL_ACCOUNTS have multicollinearity.



2- DISTRIBUTION OF VARIABLES-



The 7 numeric variables and 6 categorical variables are distributed in the as mentioned in the figure.

3- PRESENCE OF OUTLIER AND ITS TREATMENT- We have outliers in the following columns -

ANNUAL_INCOME

DEBT_TO_INCOME

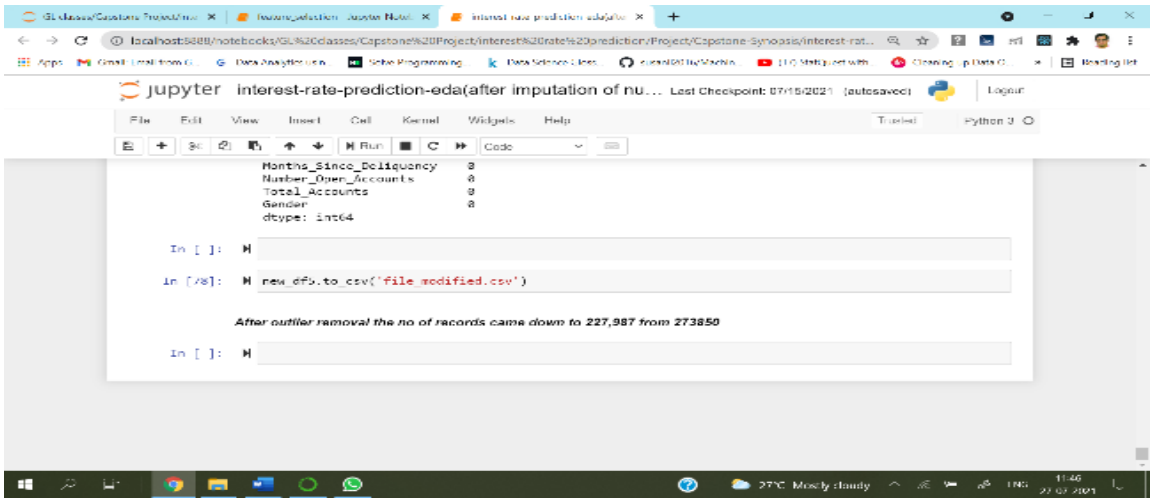
INQUIRIES_LAST_6MO

MONTHS_SINCE_DELIQUENCY

NUMBER_OPEN_ACCOUNTS

TOTAL_ACCOUNTS

We have treated the OUTLIERS by IQR METHOD. After the IQR treatment total records reduced from 273850 to 227987.



STATISTICAL SIGNIFICANCE OF VARIABLES AND CLASS IMBALANCE IS NOT REQUIRED BECAUSE WE ARE WORKING ON CLASSIFICATION PROBLEM.

FEATURE ENGINEERING: -

- Whether any transformations required:

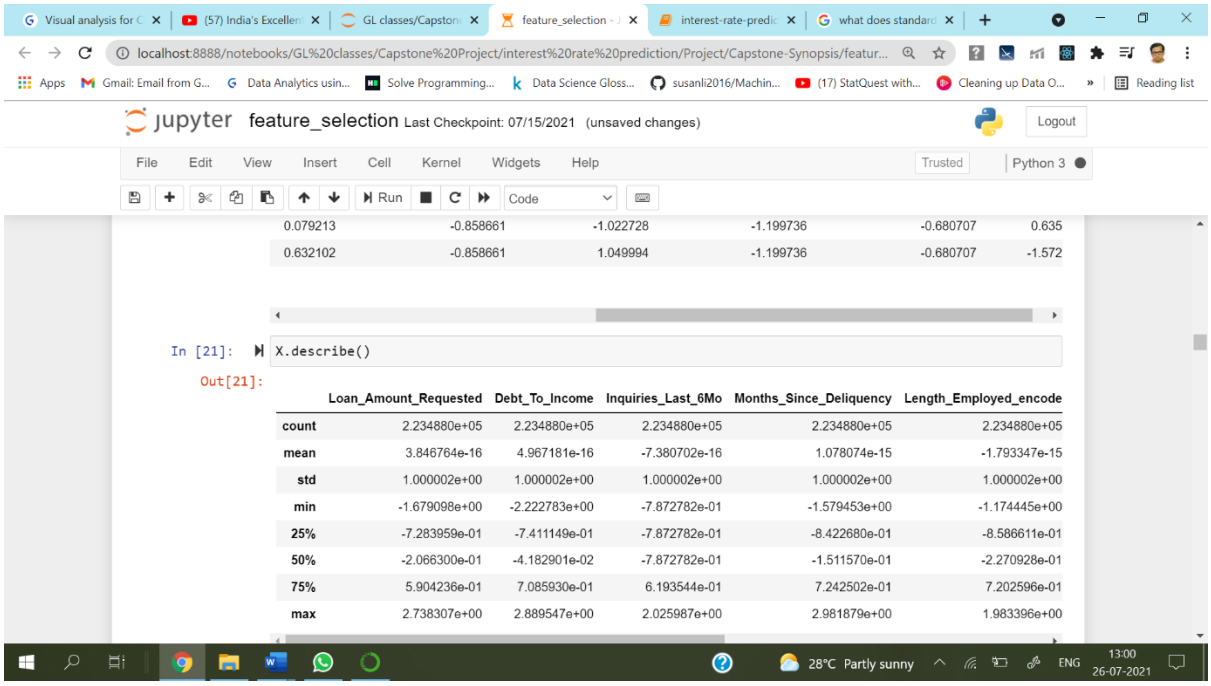
Since we need to perform unsupervised learning algorithm, we need to perform transformations. Annual income, debt to income, inquiries last 6month, number open accounts, total accounts are in a different range and there are various categorical variables in this dataset. Hence, we need to perform transformation to make all the variables continuous.

The transformations performed on this particular dataset is Label encoding and one hot encoding. Loan amount requested is in string format. This is converted into float format using pandas to numeric function. The following columns are encoded using label encoding. Length employed, home owner, income verified, Purpose of loan are label encoded. Gender is one hot encoded.

- Scaling the data:

Scaling the data is necessary because u need all the data to be in the same range. After transformation all data are transformed into numerical format. Now we need to ensure that all the features are in specific range so that our predictions don't go wrong. Hence, we applied standard scaler on the entire dataset to convert the whole data set into the same range.

Standard scaler will transform the data such we have data where the mean is 0 and standard deviation is 1. Hence, we obtain the required data in a specific range rather than scattered across various ranges.



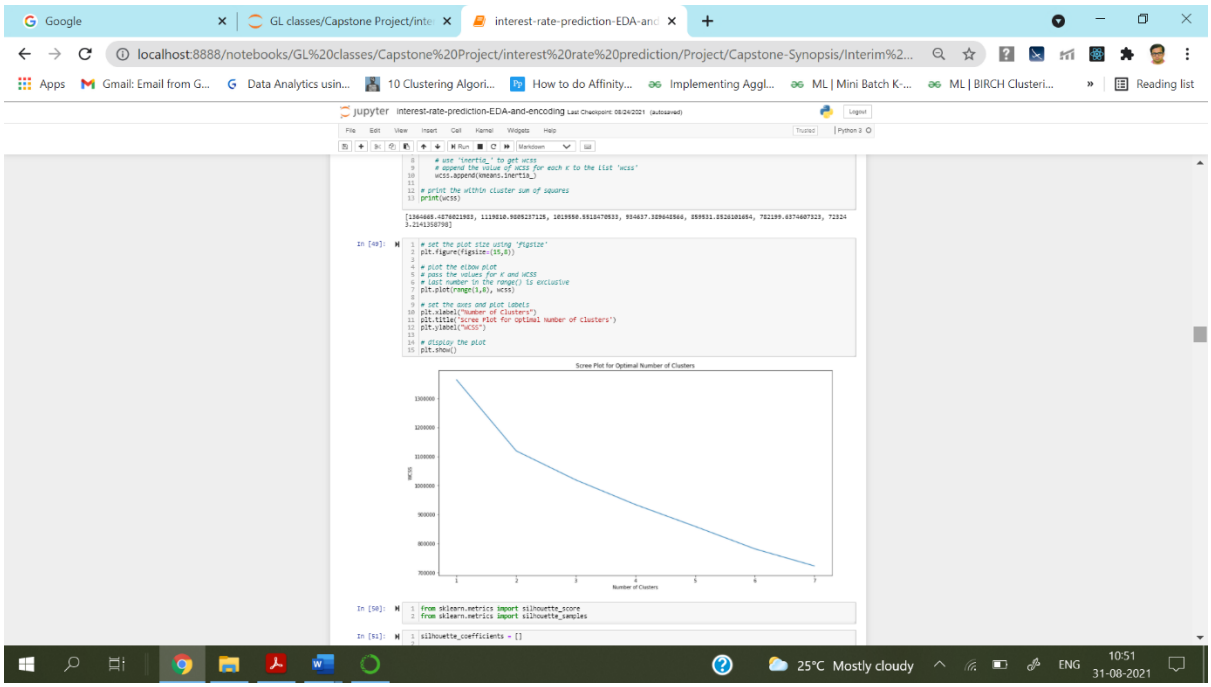
Here we can see that the entire dataset has mean 0 and standard deviation as 1. This ensures there is no complexity in data ranges.

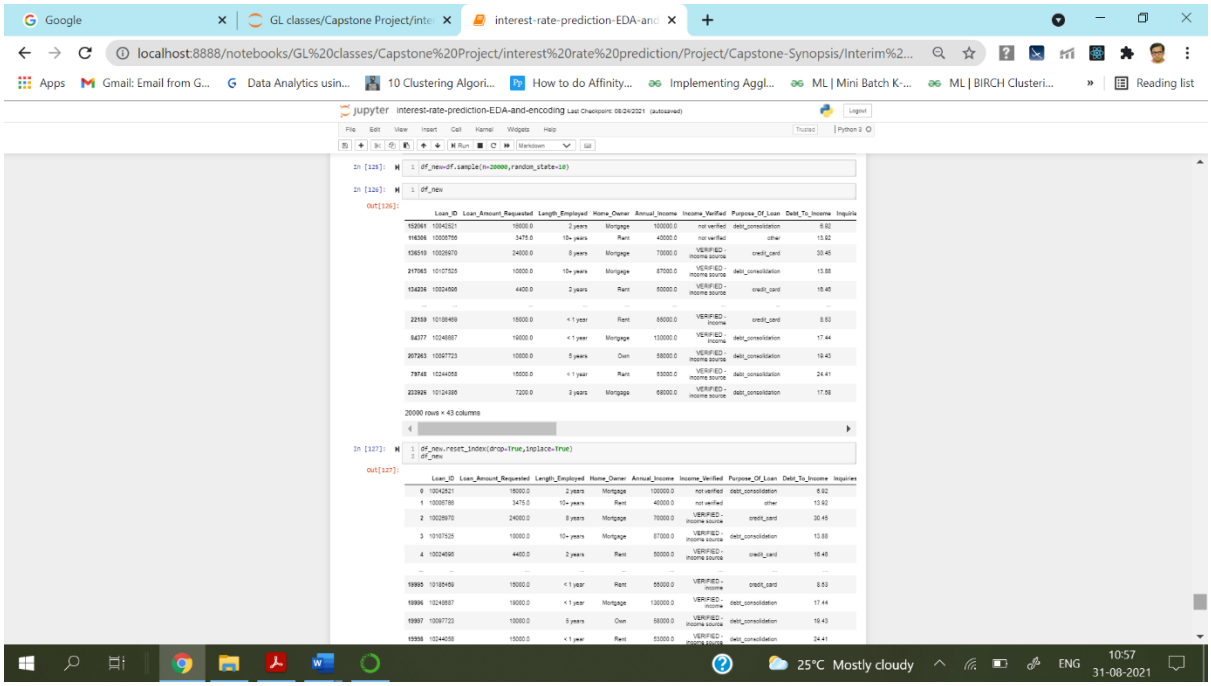
The first thing in training a machine learning model is to split the train and test sets. As researched, we found the dataset is balanced. A normal split is a good option here it serves the ratio between classes in both train and test split.

Base Model and Subsequent Models

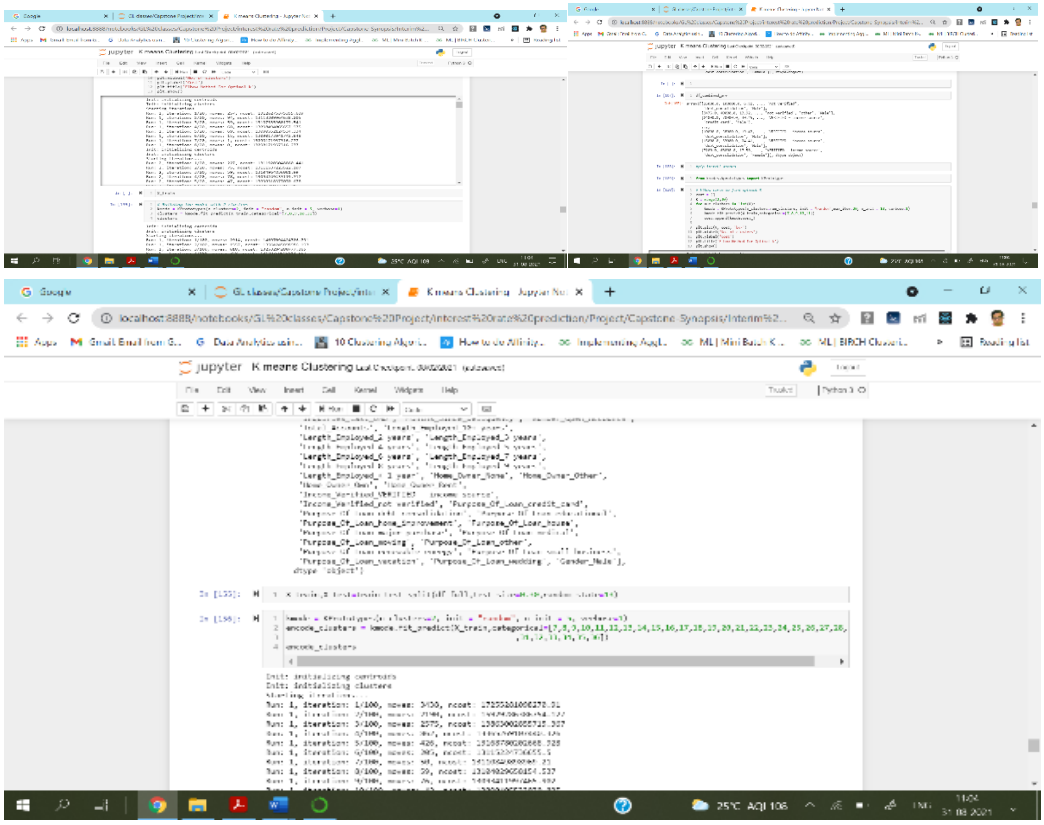
We have built a k means clustering model as our initial model with only numeric features. The model was formed before removal of outliers and it did not perform so well during testing phase and hence, we went for outlier removal.

After outlier removal the no of records came down to 227,987 from 273850.Since we cannot analyze 2 lakhs of records on our local machines with limited memory, we reduced the no of records further to 20000 which were selected randomly from the 2 lakh records.





Further we have built models with numerical and encoded categorical features and further we have also built a model with K prototypes algorithm of k-modes package without encoding categorical features.



Numerical and Categorical Feature selection

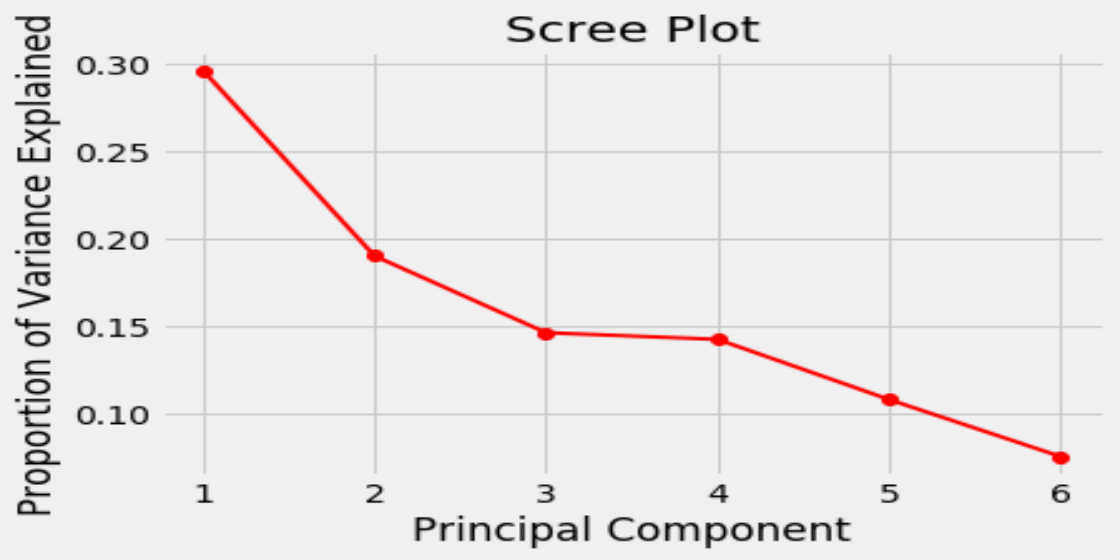
For numeric feature selection we have used Principal Component Analysis (PCA).

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

PCA converts features into eigen vectors and values where each eigen values explain the captured variance by those particular features. Hence in this way we capture about 95% of variance from the numeric data.

After our analysis we got around 6 features as our corresponding PC's which corresponds to 95% variance captured.

- Perform PCA for feature selection

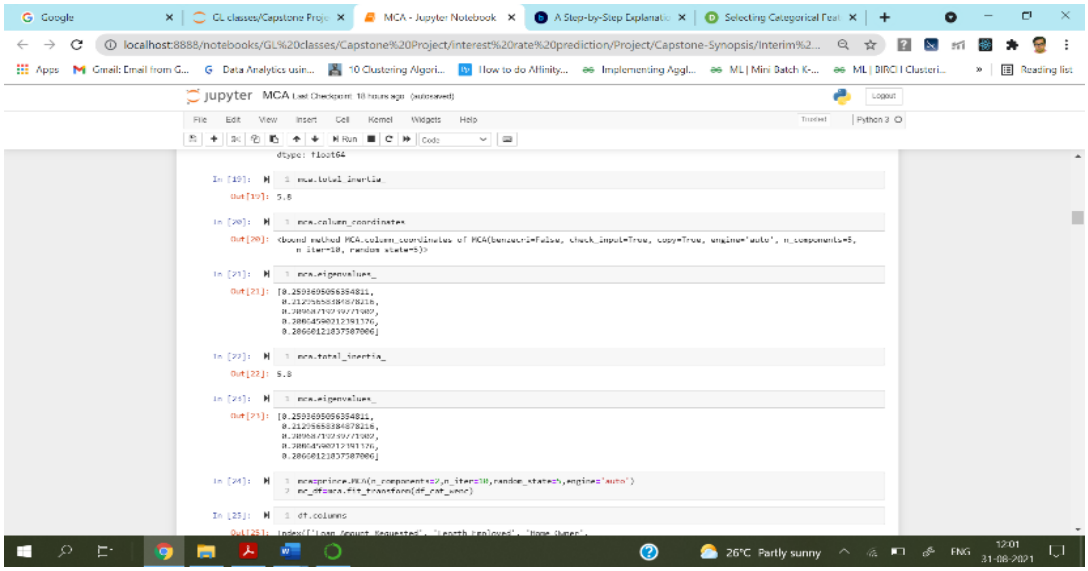


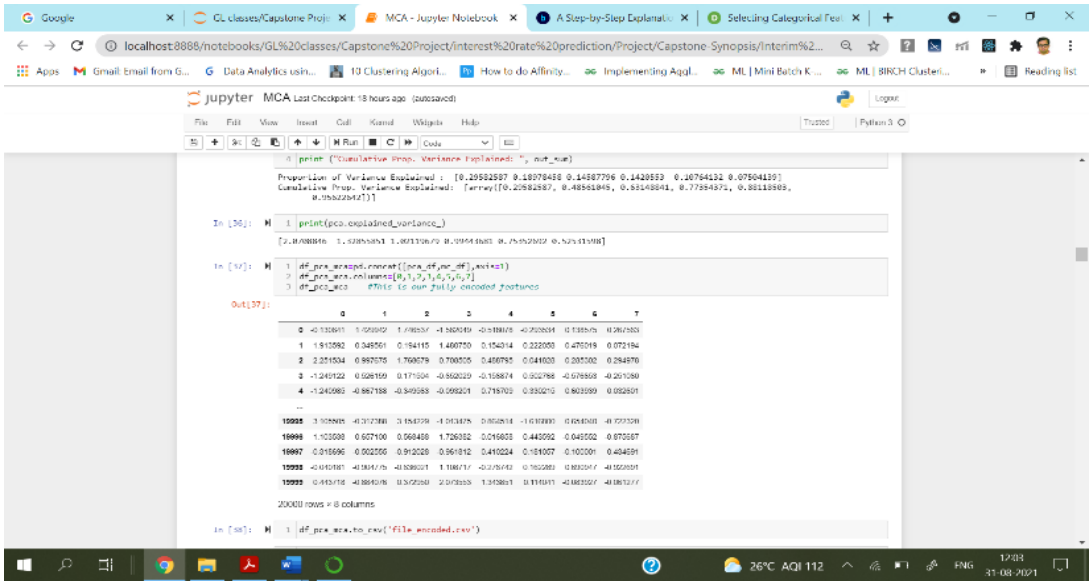
MCA or Multiple correspondence analysis is used for selection of important categorical variables from available variables.

Multiple correspondence analysis is a multivariate data analysis and data mining tool concerned with interrelationships amongst categorical features. For categorical feature selection, the scikit-learn library offers a select K Best class to select the best k-number of features using chi-squared stats (chi2).

Such data analytics approaches may lead to simpler predictive models that can generalize customer behavior better and help identify at-risk customer segments. Such prescriptive analytics efforts may also help identify customer segments.

From our analysis we have got 2 MCA components that gives a variance of 5.8.





As we combine the 6 components obtained from numeric features and 2 components obtained from categorical features we obtain a new dataset that is called features encoded. And this new features can be used for analysis of further models with different algorithms .

Further we have built Agglomerative clustering models, Birch clustering models, DB Scan models, Affinity Propagation models as well as Gaussian mixture models with predominantly these 3 datasets only:

a) Numeric features

b) Numeric features and categorical features encoded

c) PCA and MCA features together

After all these models we have evaluated them and got the following results.

Evaluation and Metrics

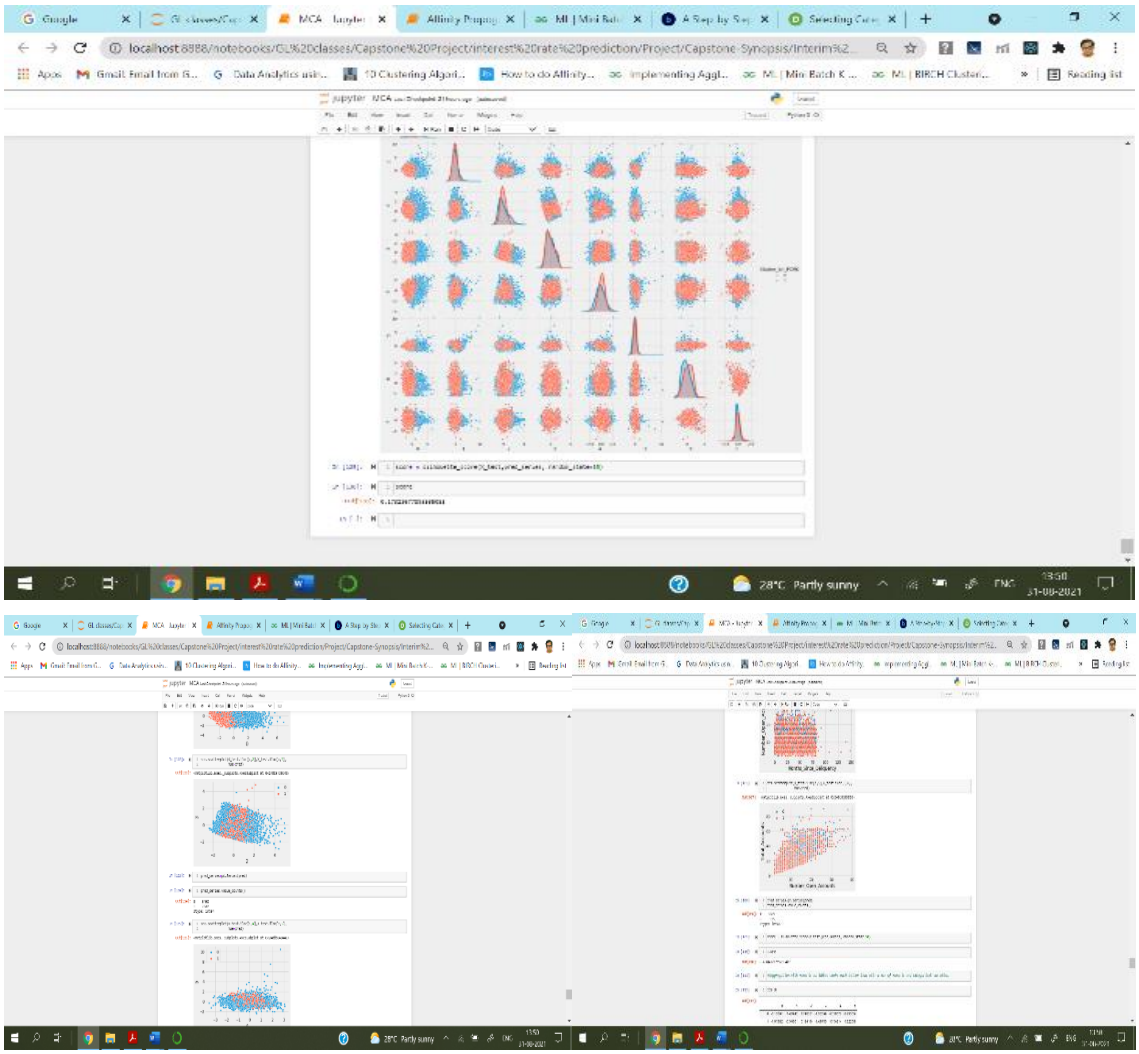
We have used Silhouette score to evaluate all these models. The results are shown below.

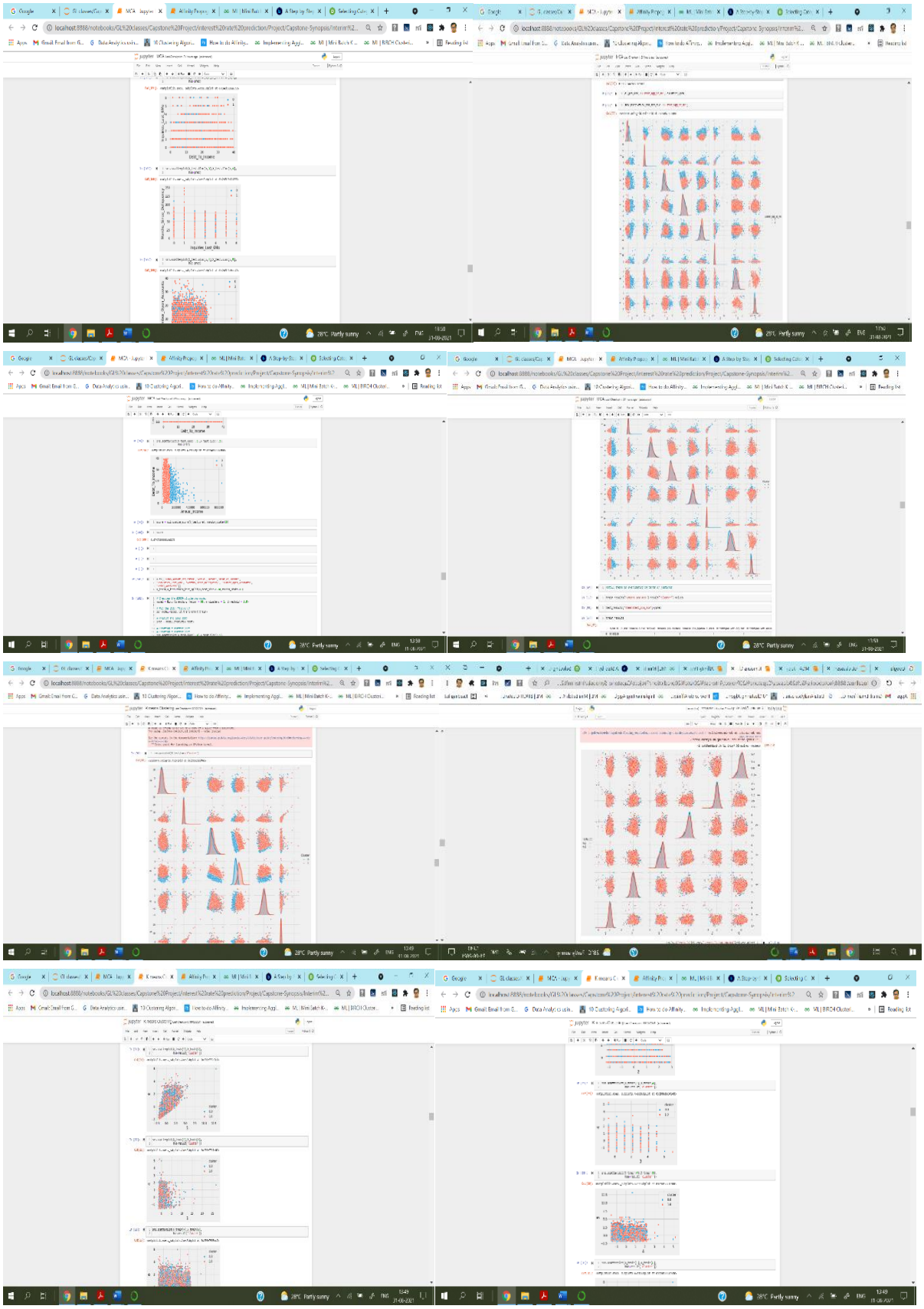
Model	Silhouette Score
K-means model with numeric features	0.196
K-means model after outlier removal	0.2305
K-means with PCA	0.205
K-means with Pipeline	0.1997
K Prototype with encoding categorical	0.6903
K-means with PCA and MCA	0.1906
Agglomerative with encoded categorical and numeric features.	0.679
Agglomerative with PCA and MCA	0.1502
DB Scan with categorical encoded and numeric features	-0.211
Birch clustering with categorical encoded and numeric features	0.674
Birch with numerical only	0.6304
Birch with PCA and MCA	0.173

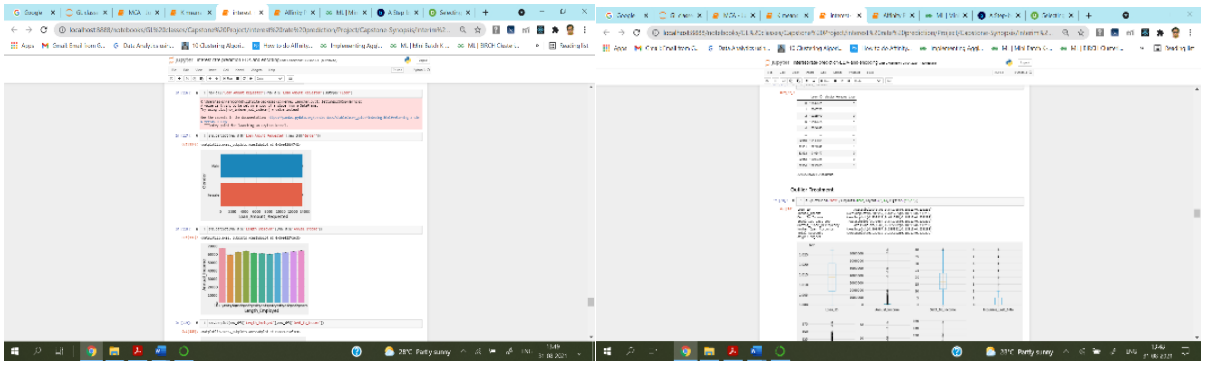
Gaussian Mixture model with numeric features	0.235
Gaussian Mixture model with PCA and MCA features	0.234
Gaussian Mixture model with numerical and encoded categorical features	0.021

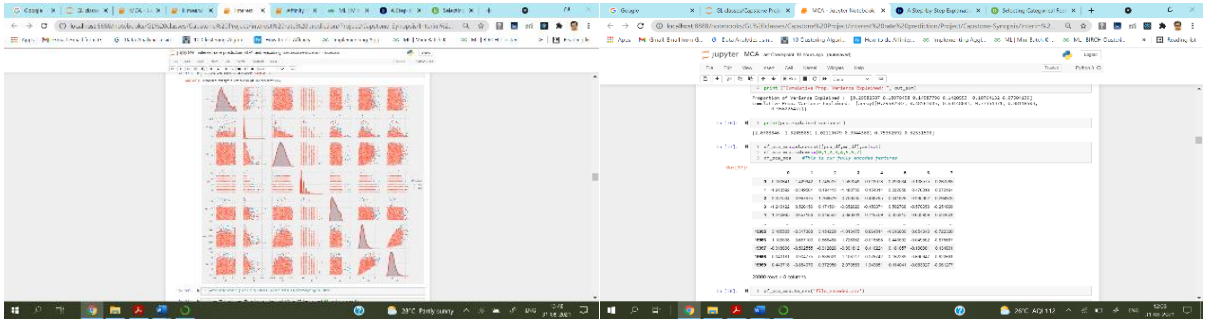
Apart from these models we have constructed K prototypes model with numerical features and categorical original features without encoding or scaling which gave us 2 clusters.

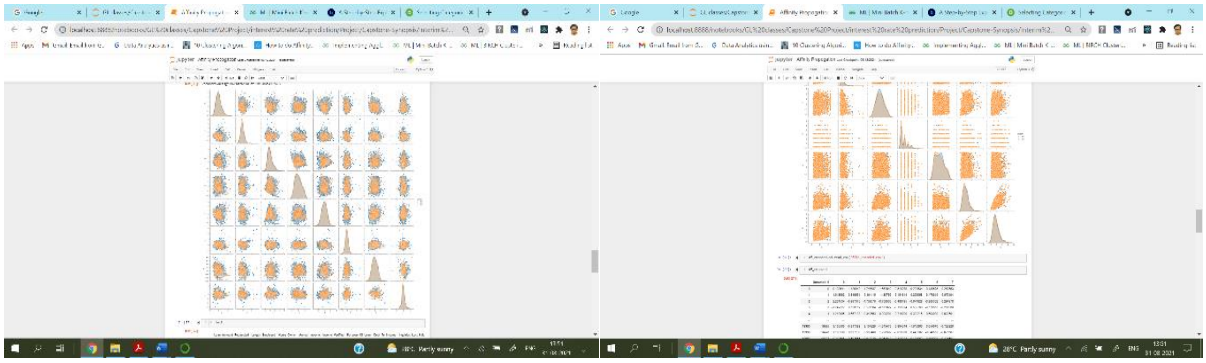
Visualizations using pair plots:

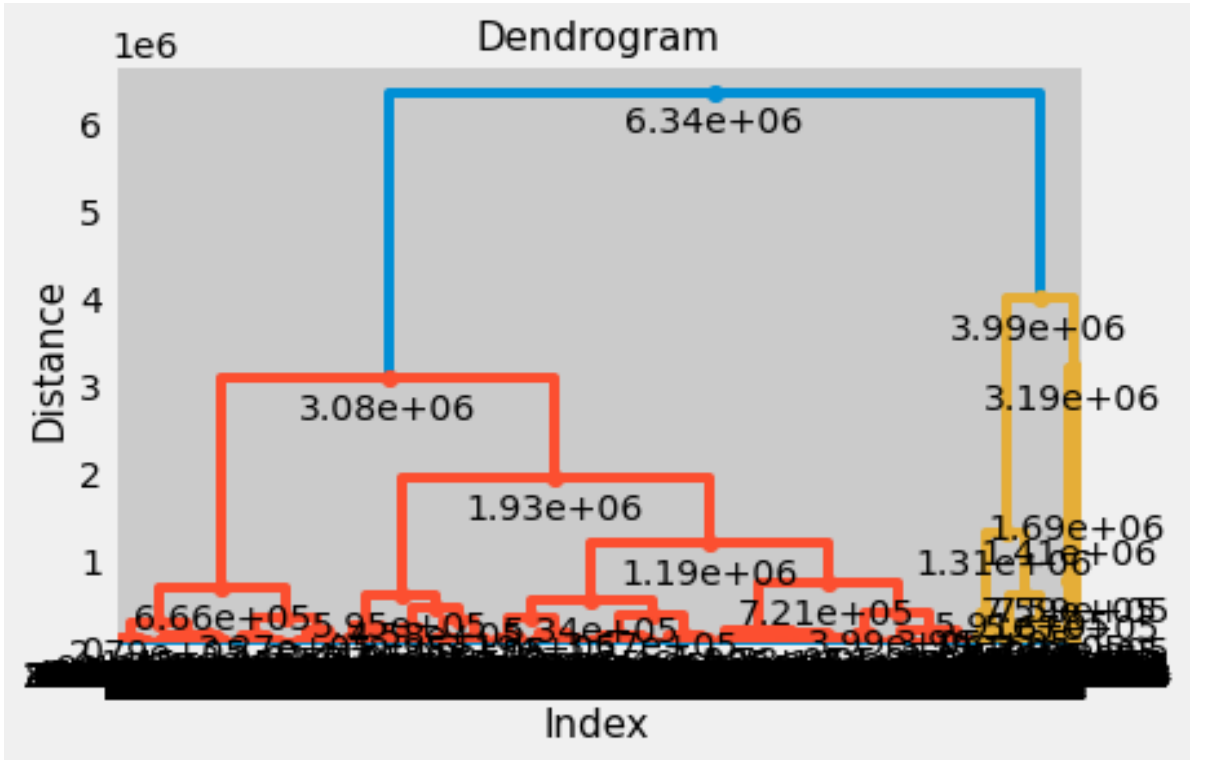










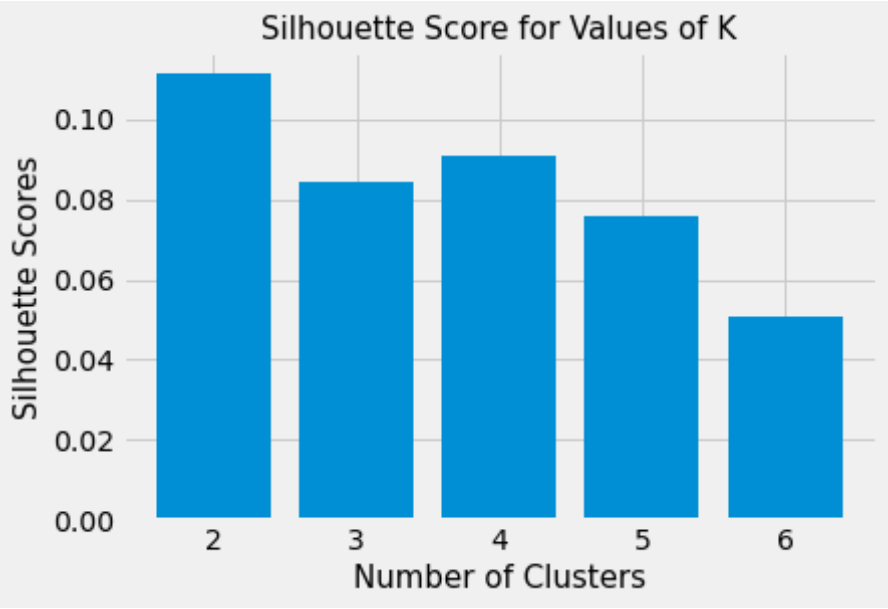

Dendrogram
1e6
Distance
Index
6.34e+06
3.99e+06
3.19e+06
3.08e+06
1.93e+06
1.69e+06
1.41e+06
1.31e+06
1.19e+06
7.21e+06
6.66e+05

- 1) x-axis contains the observations and the y-axis contains the distances computed using the 'ward' method.

- 2) Horizontal lines show the merging of the clusters.

- 3) The topmost line in the dendrogram refers to a single cluster of all the data points.

Silhouette score –



As we can see on Silhouette score number of clusters can be fixed as 2, as it have the highest score. And after retrieve the cluster we use agglomerative Clustering to form two clusters. And observation we got segregation we found is good.

On analyzing cluster 0 and 1, it is clear that cluster 1 can be given the label As 'Loan Can be Approved' and cluster 0 can be given the label 'Loan cannot be approved'. Decision is taken based on the average values of debit_to_income, Annual income, Length Employed > 1o years. Cluster 1 have the higher average values for all the above fields and Cluster 0 have an lower average values and for certain attributes values are less than 0. with respect to the field Months_Since_Deliquency - which refers to defaulters - cluster 0 have a higher average value and cluster 1 have a average value less than 0.

K Prototypes model with encoded categorical features gives the best results and hence considered the best model among all.

WHAT IS THE LIMITATION OF YOUR SOLUTION? WHERE DOES YOUR MODEL FALL SHORT IN THE REAL WORLD? WHAT CAN YOU DO TO ENHANCE THE SOLUTION?

- So as the future aspect we tried many clustering model techniques, the dataset we prepared the Agglomerative Clustering the only gave a good result but, if we work with so many missing data and mixed data types the model doesn't work very much efficiently.

- The dataset we get contains so much irregularity. We can explore the dataset more and extract more relevant and best features like decision we took on the basis of columns like debt to income, annual income, Months Since Delinquency etc.

Conclusion

From the above analysis of the loan prediction dataset, we were able to get some ideas regarding the most significant variables which are effective in predicting whether the loan needs to be approved or not. We know that there are many kinds of financial frauds that are persistent in our society. Hence a loan prediction system will eventually help in reducing such crimes. It is predicted based on the features like annual income, amount requested, months since delinquency, purpose of loan, total account etc.

Hence, we will be able to design a robust loan disbursal system in the future with the help of this project.

Datasets used and Solution Files

Datasets

1) file.csv – contains the original dataset
2) file_modified.csv – contains a mix of original dataset and encoded categorical features after removing outliers and select random 20000 records.
3) file_encoded.csv- contains the datasets selected after PCA and MCA together.

Solution Files

1) train_base_model.csv-Base model train result.
2) test_base_model.csv-Base model test results.
3) train_results.csv-train clusters formed.
4) test_results.csv-test clusters formed.
5) Agg_db_results.csv-clusters of agglomerative and dB scan clustering.

References

Reference for Dataset

https://www.kaggle.com/mayankgupta96/interest-rate-prediction

References for coding.

https://datascienceplus.com/selecting-categorical-features-in-customer-attrition-prediction-using-python/

https://www.geeksforgeeks.org/ml-mini-batch-k-means-clustering-algorithm/

https://machinelearningmastery.com/clustering-algorithms-with-python/

https://www.dezyre.com/recipes/do-affinity-based-clustering-in-python