Project Summary

Batch details	PGPDSE-FT Gurgaon april,2023
Team members	Balvinder Singh, Riya Soni, Ankit Kumar, Aasim Mirza, Aryan Vats
Domain of Project	Financial Service and Credit Profile Assessment
Proposed project title	Credit-Profile Analysis (for Two-Wheeler Loans)
Group Number	Group-5
Team Leader	Balvinder Singh
Mentor Name	Animesh Tiwari

Date: 24-December-2023

Animesh Tiwari

Signature of the Mentor

Balvinder Singh

Signature of Team Leader

Table of Contents

PROJECT DETAILS

OVERVIEW

In the bustling streets of India, two-wheeler vehicles have emerged as a symbol of both convenience and economic empowerment. For countless individuals, the prospect of owning a two-wheeler represents not just a mode of transportation but an opportunity to access better economic prospects. In pursuit of this aspiration, many turn to financial institutions for two-wheeler loans, thereby igniting the engine of a growing market. However, as the demand for such loans escalates, lending institutions are faced with a critical challenge how to effectively evaluate the creditworthiness of loan applicants. The decision to approve or deny a loan carries significant consequences, not only for the institution but, more crucially, for the livelihoods and dreams of countless individuals. Striking a balance between facilitating financial inclusion and mitigating risk is paramount. This project delves into the heart of this challenge, using machine learning techniques to assess the credit profiles of potential two-wheeler loan applicants in India. By using a dataset, tailored to the Indian demographic, we aim to develop a predictive model that empowers lending institutions to make well informed lending-decision.

BUSINESS PROBLEM STATEMENT:

In India, many people want two-wheeler loans to buy scooters or motorcycles. We have a way to help banks figure out who can get these loans and who can't. The problem is that banks need to make sure they're giving loans to people who can pay them back. They need a way to figure out if someone is a good fit for a loan or not.

1. What would you achieve by this project?

The goal of this project is to utilize a dataset containing detailed information about potential loan applicants in India. This dataset includes a variety of features such as demographic details and financial information. The aim is to leverage this data to assess the creditworthiness of individuals applying for loans, helping lenders make informed decisions on whether to approve or deny loan applications.

2. How would this help Business and Clients?

For Business:

Risk Mitigation: Businesses can better assess the credit risk associated with loan applicants by analyzing the comprehensive dataset. This helps in reducing the likelihood of default and potential financial losses.
Informed Decision-Making: Access to a diverse set of applicant details allows businesses to make more informed decisions on approving or denying loan applications, contributing to a more effective and efficient lending process.
Tailored Products: The insights gained from the dataset enable businesses to design and offer more tailored financial products that meet the specific needs and capabilities of their target demographic.

For Clients:

Increased Approval Chances: Clients benefit from a more nuanced evaluation of their creditworthiness, potentially increasing their chances of loan approval as the decision is based on a comprehensive understanding of their financial profile.
Fair and Transparent Evaluation: The use of detailed demographic and financial information ensures a fair and transparent assessment process, providing clients with clarity on why their loan application was accepted or rejected.
Customized Offerings: As businesses tailor their financial products based on the dataset, clients may have access to more personalized loan options that align with their unique financial situations and goals.

3. What is the further scope of this project?

The project can evolve to meet the dynamic needs of the market, contribute to financial inclusion, and provide valuable insights for both lending institutions and loan applicants in the context of two-wheeler financing in India
Dynamic Risk Assessment: Implement a system for dynamic risk assessment that adapts to changing economic conditions, market trends, and borrower behavior over time. This could involve incorporating real-time data feeds or regularly updating the model with the latest information.
Behavioral Analytics: Explore the integration of behavioral analytics to gain insights into the financial habits and patterns of loan applicants. This can provide a more holistic view of an individual's creditworthiness beyond traditional financial metrics.
Global Expansion: Consider adapting the model for use in other global markets, tailoring features and criteria to suit diverse demographics and regulatory environments.

4. Limitation of the project?

Data Accuracy: Reliability of decisions depends on the accuracy of the data, and inaccuracies may lead to flawed assessments.
Privacy Concerns: Collection of detailed personal and financial information raises privacy issues, necessitating robust security measures.

Dynamic Economic Factors: Economic conditions can rapidly change, impacting the relevance of historical financial data for predicting creditworthiness.

Cultural and Social Factors: The dataset may not fully capture cultural and social nuances affecting individuals' financial behaviors.

Algorithmic Bias: The model's predictions may reflect biases present in historical data, potentially leading to unfair outcomes for certain demographic groups.

TOPIC SURVEY IN BRIEF

1. Problem understanding

The problem at hand revolves around the efficient and responsible assessment of the creditworthiness of individuals applying for two-wheeler loans in India. Can we develop a robust machine learning model that accurately evaluates the credit profiles of two-wheeler loan applicants, enabling lending institutions to make informed and ethical lending decisions, while ensuring fairness and adherence to regulatory standards?

2. Proposed solution to the problem?

To make the project better, we should do a few things. First, we need to check and make sure the information in the dataset is accurate. This means we want to be sure that the details about people are correct. Second, we have to be very careful with people's private information. We need strong security measures to keep this data safe. Third, because the economy can change, we should regularly update our information to keep it useful. Fourth, we must be fair to everyone. This means we need to check if our system is treating people differently because of their background. Finally, we should keep learning about different cultures and lifestyles to understand everyone better. By doing all these things, we can make our project more accurate, safe, and fair for everyone.

3. Reference to the problem

Cite blogs, dataset, articles of this domain.

CRITICAL ASSESSMENT OF TOPIC SURVEY

Q. Find the key area, gaps identified in the topic survey where the project can add value to the customers and business, what key gaps are you trying to solve?

Predicting two-wheeler loan profile scores involves a critical assessment of various factors to determine the creditworthiness of an individual applying for the loan, here are some key aspects to consider in the critical assessment:

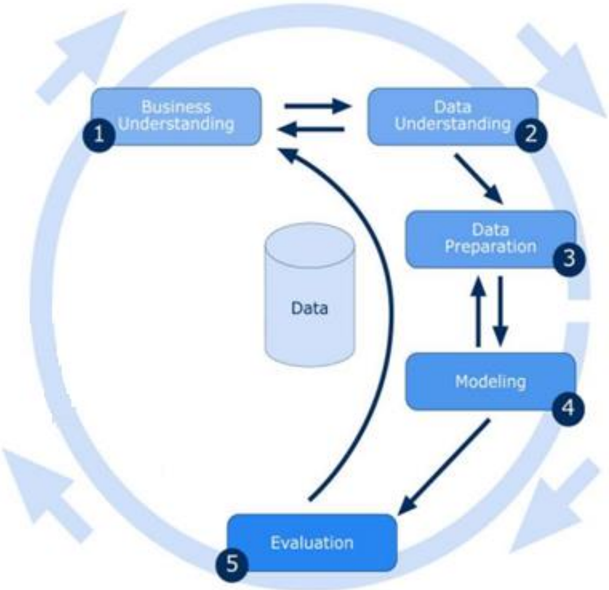
1. Credit History:
   - Review the applicant's credit history to assess their track record in repaying previous loans and managing credit.
   - Check for any defaults, late payments, or outstanding debts that might indicate a higher risk.
2. Income and Employment Stability:
   - Evaluate the applicant's income level and stability of employment.
   - A stable and sufficient income is crucial for timely loan repayments.
3. Debt-to-Income Ratio:
   - Analyze the ratio of the applicant's debt to their income. A high ratio may suggest that the individual is overleveraged.
4. Loan Amount and Loan-to-Value Ratio:
   - Assess the loan amount requested in related to the value of the two-wheelers being financed.
   - Evaluate the LTV ratio to understand the level of risk associated with the loan.
5. Credit Score:
   - Consider the applicant's credit score as it provides a numerical representation of their creditworthiness.
   - A higher credit score generally indicates a lower credit risk.
6. Employment Stability:
   - Evaluate the stability of the applicant's job or source of income.
   - A consistent employment history may indicate a lower risk of default.
7. Geographical Location:
   - Consider the applicant's geographical location, as economic conditions and regional factors can impact the ability to repay loans.

8. Age and Demographics:
    - Analyze the applicant’s age and demographics, as these factors can provide insight into their financial stability and responsibility.
9. Payment History:
    - Examine the applicant’s payment behavior on other bills and financial obligations.
    - Consistent and timely payments demonstrate responsible financial management.
10. Purpose of the Loan:
    - Understand the purpose for which the two-wheeler loan is being taken.
    - Some purpose may be considered more financially responsible than others.
11. Fraud Prevention:
    - Implement measures to detect and prevent fraud, such as verifying the authenticity of documents provided by the applicant.
12. Data Accuracy:
    - Ensure that the data used for assessment is accurate and up-to-date to make informed decisions.

METHODOLOGY TO BE FOLLOWED

HIERARCHICAL- APPROACH – Evaluation Learning

- Business Understanding
- Data Understanding
- Data Preparation
- Code For Similarity
- Evaluation



Business Understanding

Operating a successful two-wheeler loan business in India requires a comprehensive understanding of the market, customer dynamics, and regulatory landscape. This involves analyzing demographics, market size, and trends, while staying compliant with regulatory requirements set by entities like the Reserve Bank of India. To manage credit risk, businesses must develop effective credit scoring models and implement risk mitigation strategies. Additionally, offering competitive interest rates, flexible loan terms, and leveraging technology for efficient operations are crucial for staying ahead in the market. Establishing strong partnerships with two-wheeler dealerships and diversifying distribution channels, including online platforms, enhances market reach. A focus on positive customer experiences, effective marketing, and brand building contributes to building trust. Continuous monitoring of the loan portfolio, adapting to economic factors, and implementing proactive collection strategies are essential for long-term success. Embracing sustainability initiatives and considering the social impact of the business further contribute to responsible and competitive operations.

PROBLEM STATEMENT:

In India, there is a growing demand for two-wheeler loans, reflecting people's aspirations to own scooters or motorcycles. However, banks face the challenge of ensuring responsible lending by identifying individuals with the capacity to repay these loans. The issue at hand is developing an effective system that enables banks to make informed decisions about loan approvals, distinguishing between applicants who are suitable candidates for loans and those who may pose a higher risk of non-repayment. The goal is to establish a reliable method for assessing the creditworthiness of individuals seeking two-wheeler loans, thus facilitating responsible and inclusive financial practices in the banking sector.

Data Understanding

Credit Profile Dataset

Number of records: 100000

Number of variables: 15

This dataset provides a comprehensive overview of potential loan applicants' profiles, specifically tailored for the Indian demographic. It encapsulates a range of features, from basic demographics to financial details, that can be instrumental in assessing the creditworthiness of an individual.

Feature Name	**Type**	**Description**	**Range**
Age	Integer	Represents the age of the applicant. Indicate the applicant maturity level.	18 to 70
Gender	Categorical	Gender of the applicant.	Male, Female and Other
Income	Integer	The applicant's income, which is critical in assessing their ability to repay the loan.	Multiples of 1000's
Credit Score	Integer	A score quantifying the applicant's creditworthiness based on their credit history	300 to 850
Credit History Length	Integer	Represents the number of months since the applicant's first credit line.	Months
Number of Existing loan	Integer	The number of loans the applicant currently active	0 to 10
Loan Amount	Integer	The amount of money the applicant is requesting	0 to 1,50,000
Tenure	Integer	The number of months the applicant wants to repay the loan over.	Months

Existing customer	Categorical	Whether the applicant is an existing customer or not	Yes, No
State	Categorical	The state in India where the applicant resides	Maharashtra, Delhi, Karnataka, Tamil Nadu, West Bengal, Uttar Pradesh, Gujarat, Rajasthan, Kerala, Telangana, etc.
City	Categorical	The city or village in India where the applicant resides.	Mumbai, Pune, New Delhi, Bengaluru, Chennai, Kolkata, Ahmedabad, Jaipur, Kochi, Hyderabad, and various villages.
LTV Ratio	Float	The Loan-to-value ratio, represents the ratio of the loan amount to the appraised value of the asset. Higher LTVs can indicate higher risk.	40% to 95%
Employment Profile	Categorical	General employment category of the applicant	Salaried, Self-employed, freelancer, Unemployed, Student.

Occupation	Categorical	Specific occupation or job title of the applicant	Software Engineer, Doctor, Teacher, Business Owner, Writer, etc.
Profile Score	Integer	A score represents the overall profile of the applicant based on the actual loan repayment data. Higher values indicated better profiles.	0 to 100

Target Variable

- Profile score is our target variable
- Distribution – Left skewed

Derived column

- Zone
- EMI_per_month

Existing column

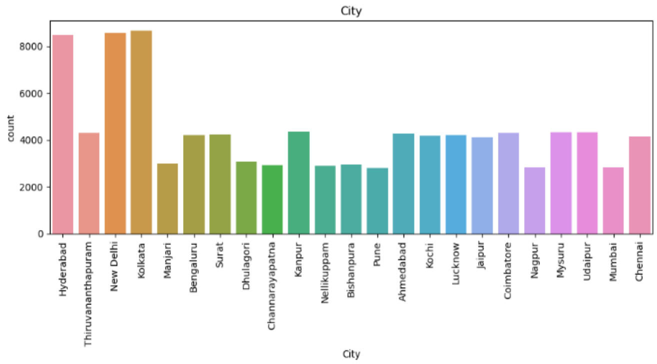
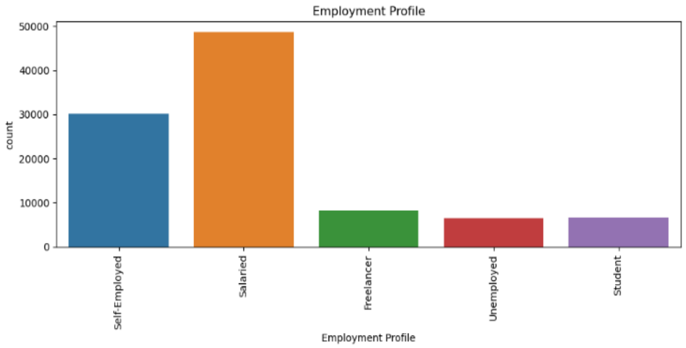
- Age
- Gender
- Income
- Credit Score
- Credit History length
- Number of existing loans
- Loan amount
- Loan tenure
- Existing customer
- State
- City
- LTV Ratio
- Employment profile
- Profile score
- Occupation

Number column- Age, Income, Credit Score, Credit History Length, Loan amount, Loan Tenure, LTV Ratio, Profile Score

Category column- City, Employment profile, Existing Customer, Gender, Number of existing loan, Occupation, State

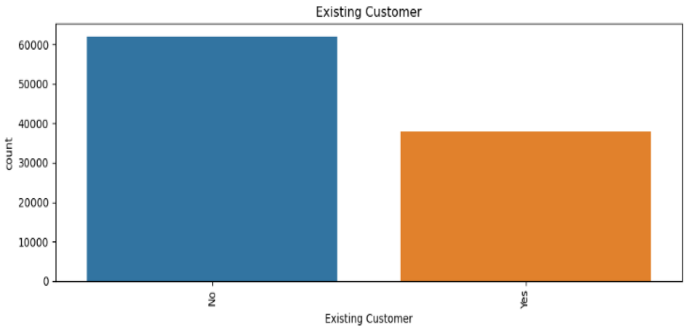
Univariate Analysis

Category

City Hyderabad, New Delhi and Kolkata has highest count	
Employment Profile Salaried>Self-employed> Freelancer> unemployed and Student	

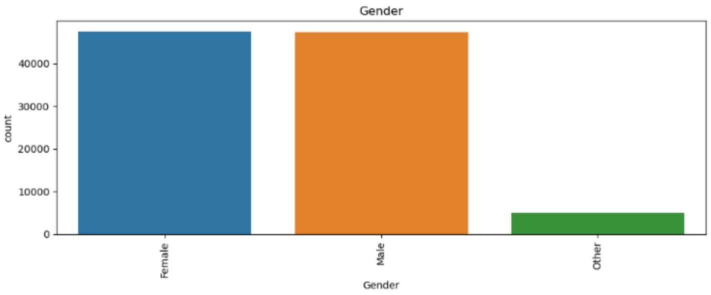
Existing Customer

No has the highest count



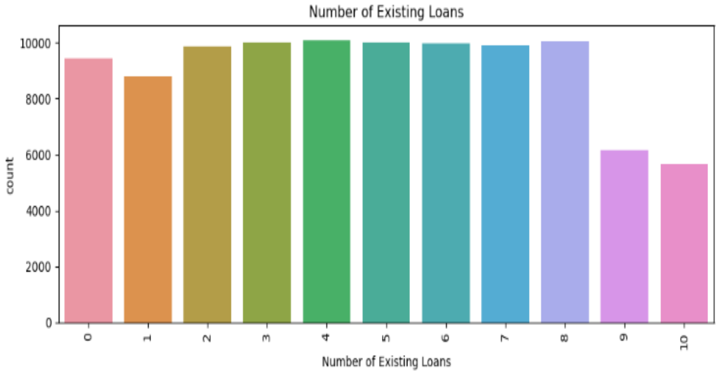
Gender

Male and Female has almost similar count and more than others



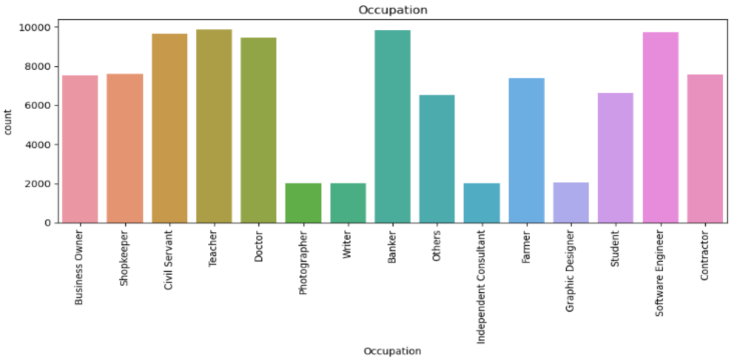
Number of Existing Loans

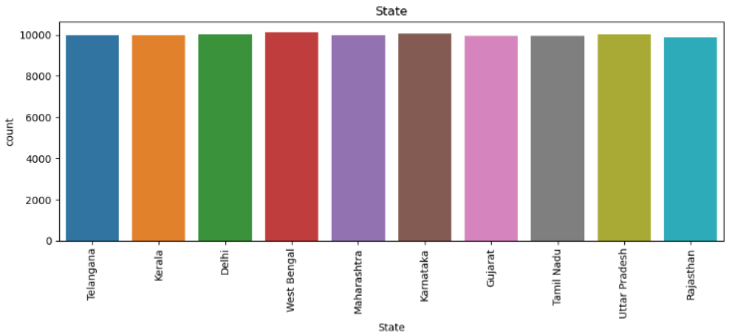
9 & 10 has lowest count



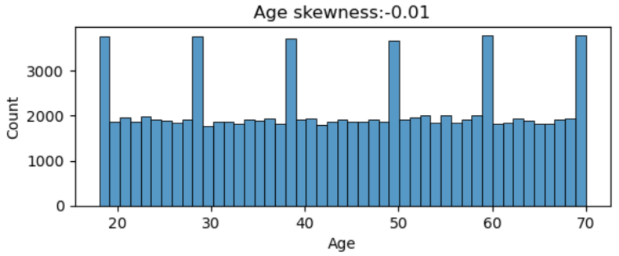
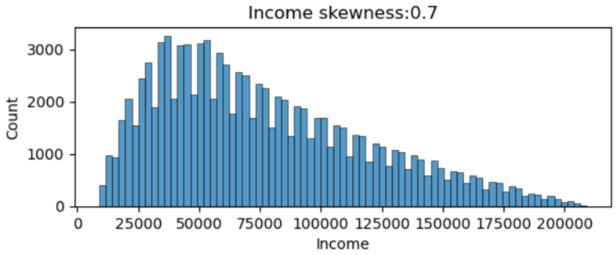
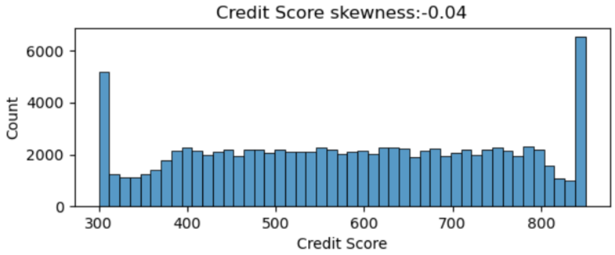
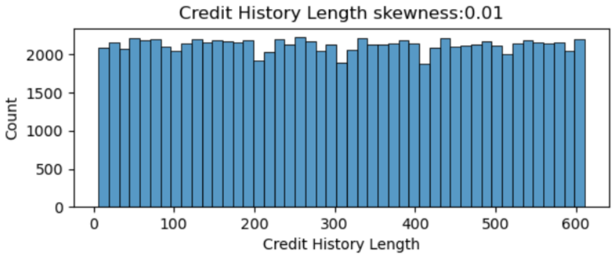
Occupation

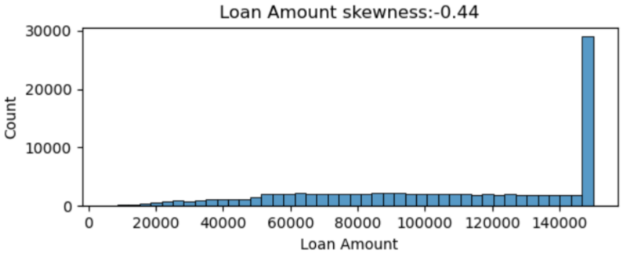
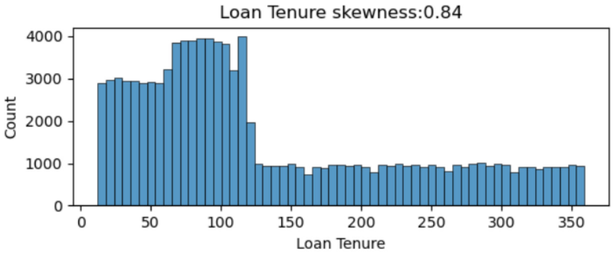
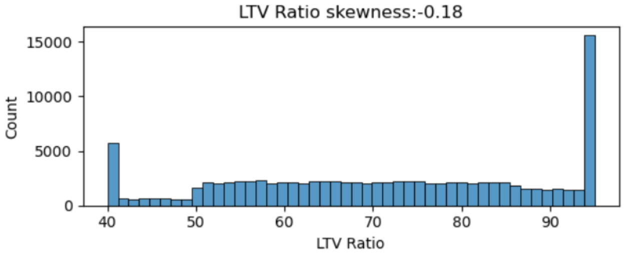
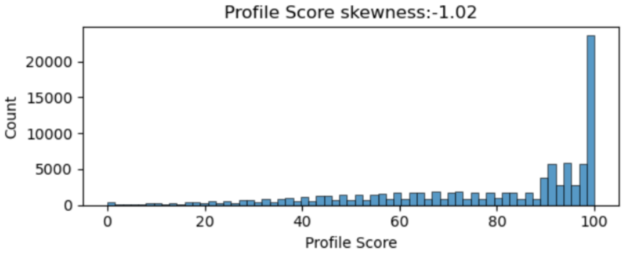
Civil service, Teacher, Doctor, banker and software engineer has highest count



State All states have similar counts	

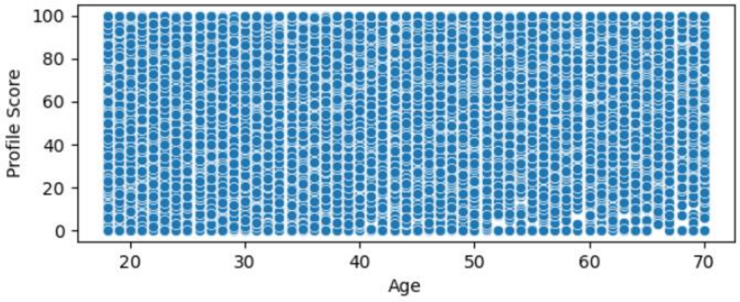
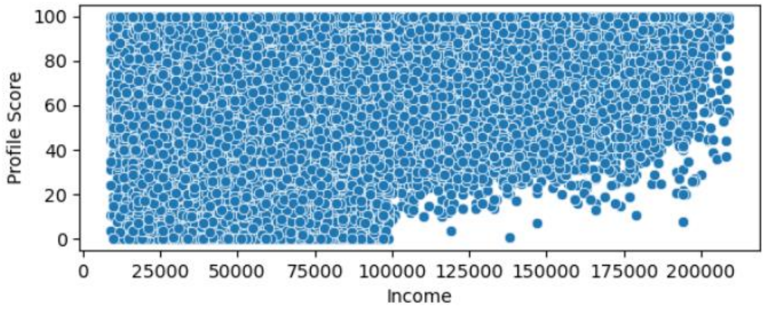
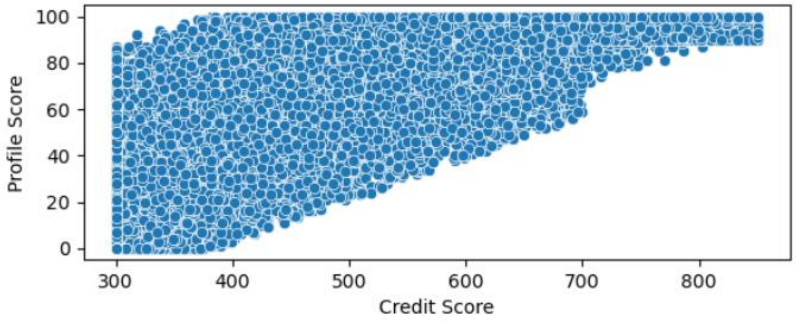
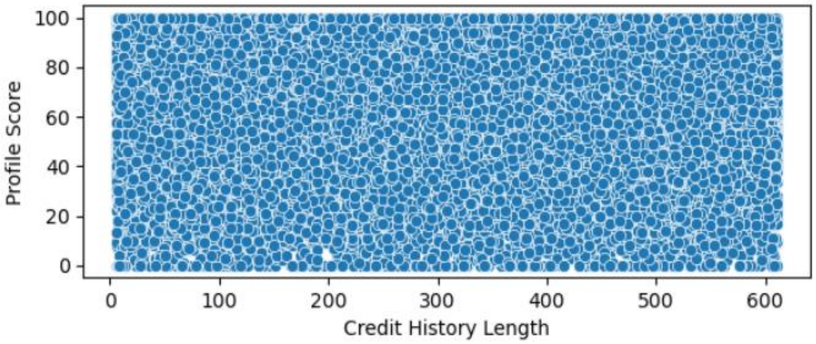
Number

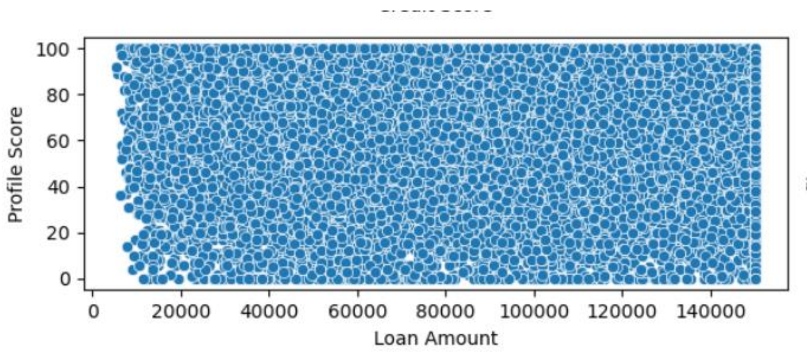
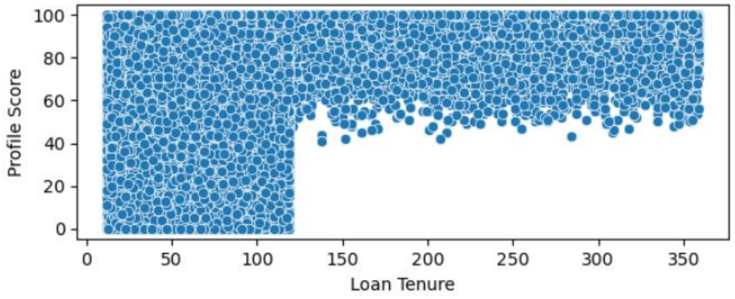
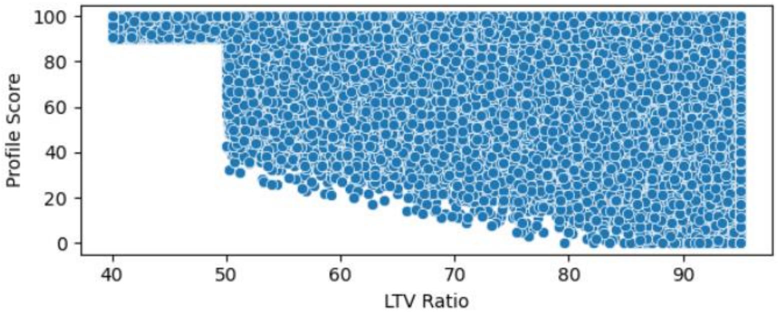
Age Skewness = -0.01	
Income Skewness = 0.7	
Credit Score Skewness = -0.04	
Credit History Length Skewness = 0.01	

Loan Amount Skewness = -0.44	
Loan Tenure Skewness = 0.84	
LTV Ratio Skewness = -0.18	
Profile Score Skewness = -1.02	

Bivariate Analysis

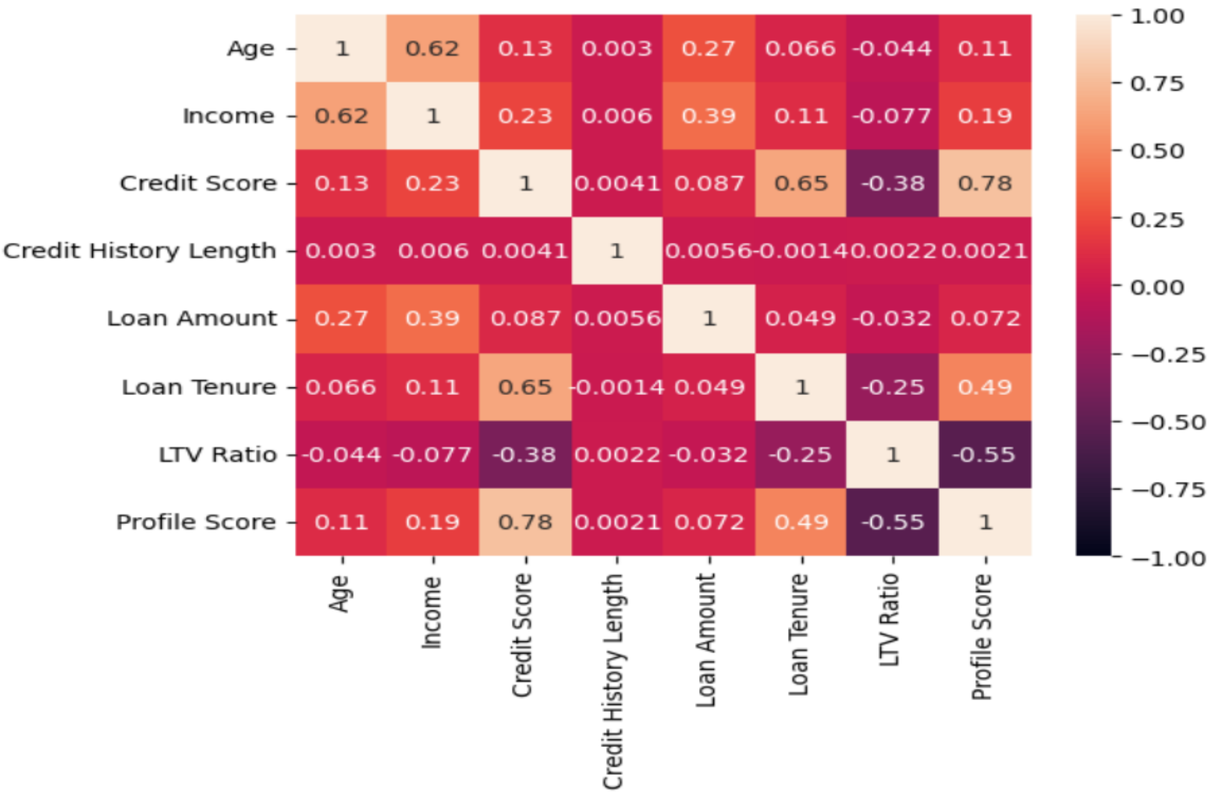
NUMBER vs TARGET

Profile Score vs Age	
Profile score vs Income	
Profile Score vs Credit Score	
Profile Score vs Credit History Length	

Profile Score vs Loan Amount	
Profile Score vs Loan tenure	
Profile Score vs LTV Ratio	

Inference: As all the data points are uniformly distributed. Therefore, there is no such significant relation between Number and target column.

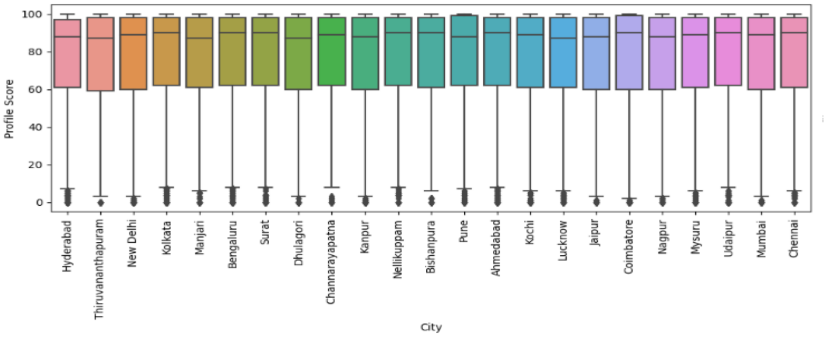
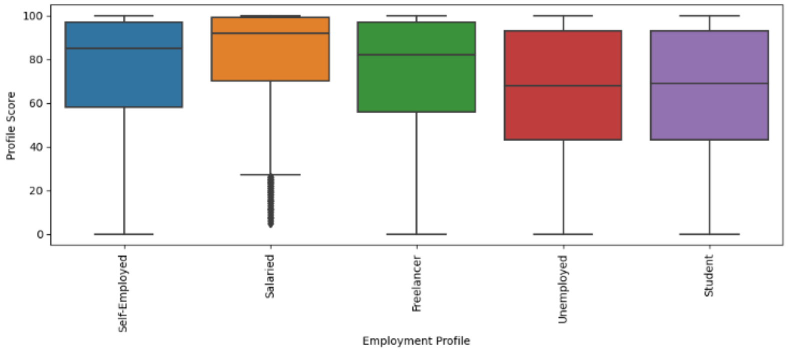
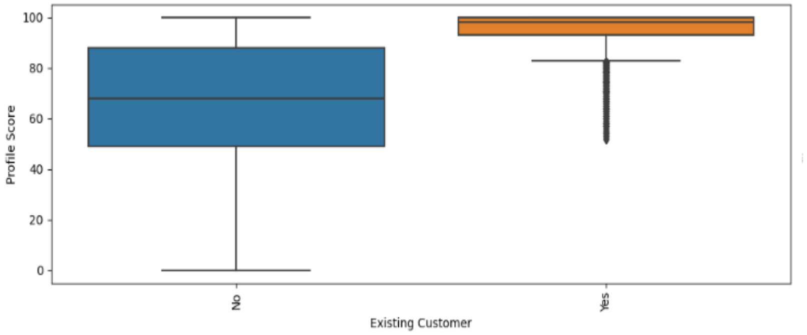
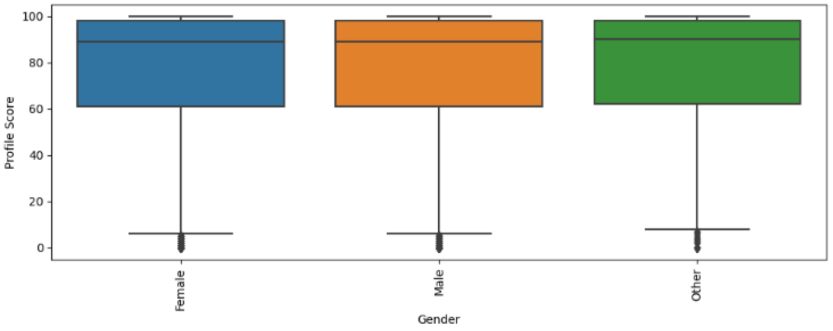
HEAT MAP

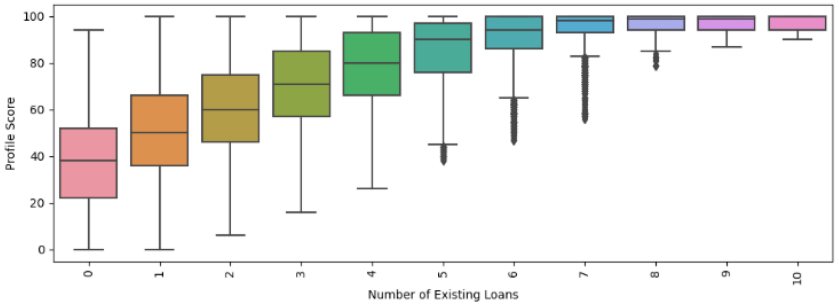
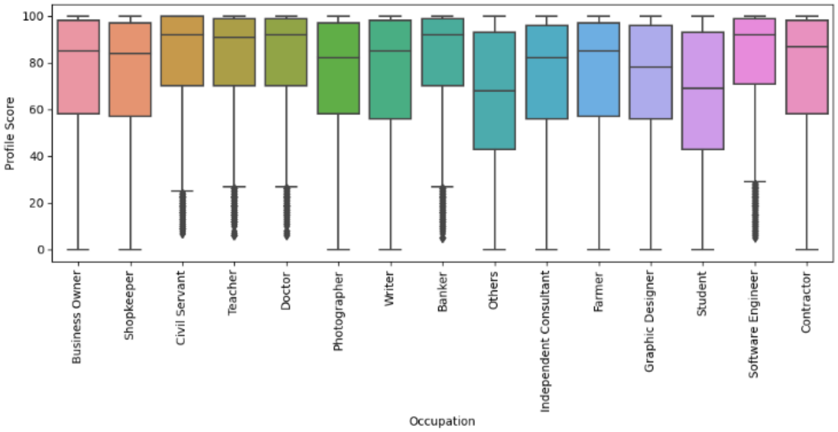
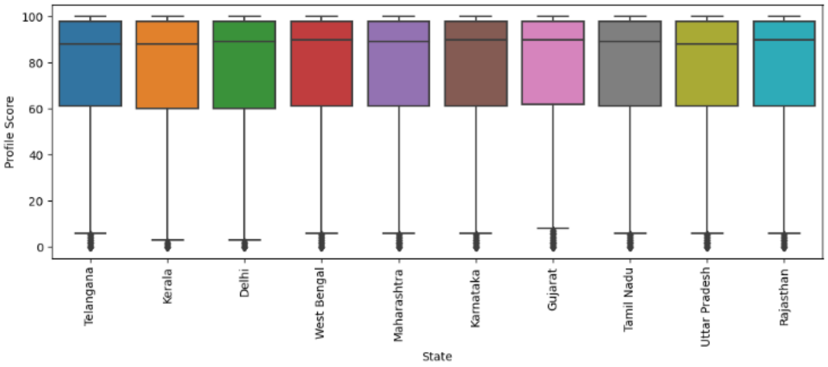


INFERENCE

- Income and age are positive correlated (62%)
- Loan Tenure and Credit score are positive correlated (65%)
- Profile score and Credit score are positive correlated (78%)
- Profile Score and LTV ratio are negative correlated (55%)

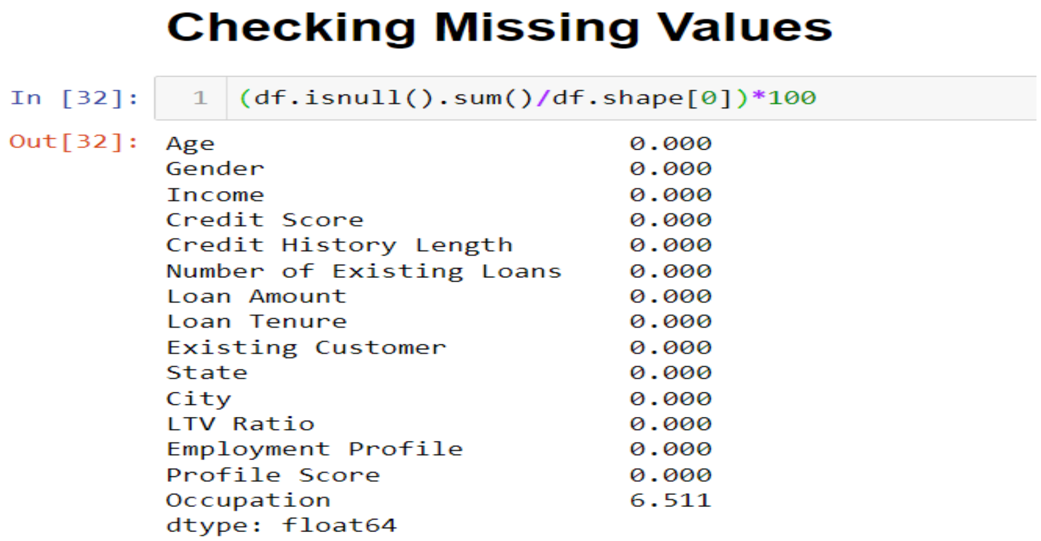
CATEGORY vs TARGET

Profile score vs City	
Profile score vs Employment profile	
Profile Score vs Existing Customer	
Profile Score vs Gender	

Profile Score vs number of Existing loan	
Profile Score vs Occupation	
Profile Score vs State	

Data Preparation

Missing Values-



Checking Missing Values

```
In [32]: (df.isnull().sum()/df.shape[0])*100
```

```
Out[32]: Age                            0.000
         Gender                         0.000
         Income                         0.000
         Credit Score                   0.000
         Credit History Length          0.000
         Number of Existing Loans       0.000
         Loan Amount                    0.000
         Loan Tenure                    0.000
         Existing Customer              0.000
         State                          0.000
         City                           0.000
         LTV Ratio                      0.000
         Employment Profile             0.000
         Profile Score                  0.000
         Occupation                     6.511
         dtype: float64
```

Occupation- 6% of the total data is missing in occupation column.

Approach-

Tried to find the pattern on the basis of which we can impute the missing values in the occupation column. There seems to be a relationship between occupation and employment profile and checked where occupation is null what is the status of employment profile there checked the value counts of occupation where employment profile is unemployed so that we can find the pattern, but there's no such pattern found as where the employment profile is unemployed all the rows in the occupation column is null.

No such pattern found on the basis of which we can impute the occupation null values, therefore, imputing the null values with OTHERS.

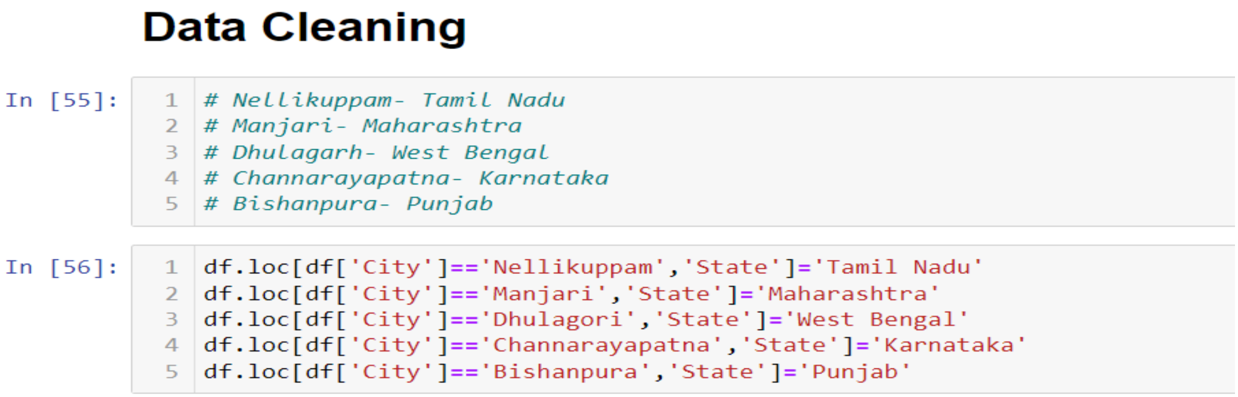
Outliers-

There are no such potential outliers found in the numerical columns that could be treated

Data Cleaning-

City- under City column, we have observed some cities appear in multiple states in the dataset. Ensure proper handling to maintain data consistency.

Cleaning process - Addressing cities listed in multiple states to ensure accurate analysis. Decisions on handling duplicates (e.g.- Bishanpura, Nellikuppam city is listed in multiple state) were made during this step.



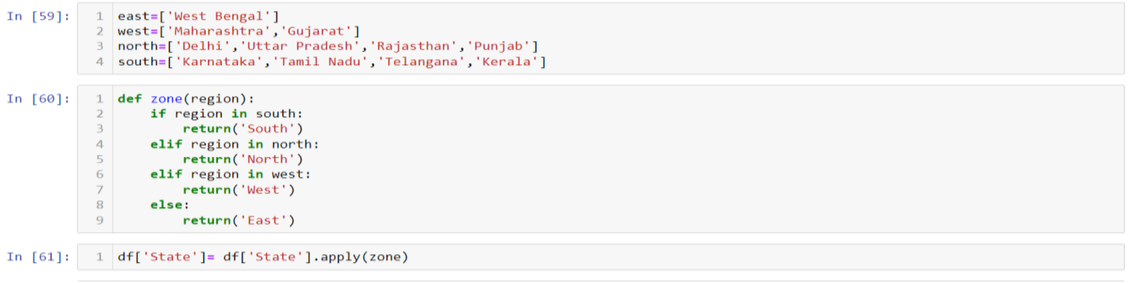
Data Cleaning

```
In [55]:
# Nellikuppam- Tamil Nadu
# Manjari- Maharashtra
# Dhulagarh- West Bengal
# Channarayapatna- Karnataka
# Bishanpura- Punjab
```

```
In [56]:
df.loc[df['City']=='Nellikuppam','State']='Tamil Nadu'
df.loc[df['City']=='Manjari','State']='Maharashtra'
df.loc[df['City']=='Dhulagori','State']='West Bengal'
df.loc[df['City']=='Channarayapatna','State']='Karnataka'
df.loc[df['City']=='Bishanpura','State']='Punjab'
```

Feature Engineering-

Zone- States were converted into zones as north, south, east and west

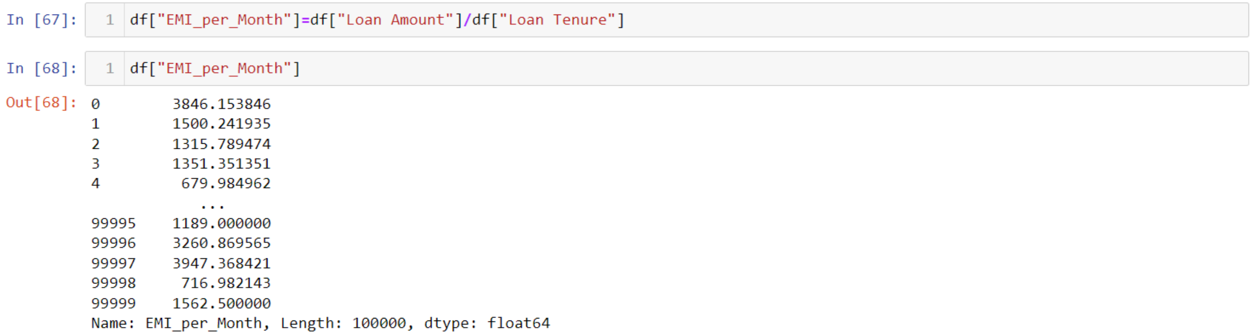


```
In [59]:
east=['West Bengal']
west=['Maharashtra','Gujarat']
north=['Delhi','Uttar Pradesh','Rajasthan','Punjab']
south=['Karnataka','Tamil Nadu','Telangana','Kerala']
```

```
In [60]:
def zone(region):
    if region in south:
        return('South')
    elif region in north:
        return('North')
    elif region in west:
        return('West')
    else:
        return('East')
```

```
In [61]:
df['State']= df['State'].apply(zone)
```

EMI_per_Month- EMI per month was calculated with the help of the columns Loan Amount and Loan Tenure (Loan Amount/Loan Tenure)



```
In [67]:  df["EMI_per_Month"]=df["Loan Amount"]/df["Loan Tenure"]

In [68]:  df["EMI_per_Month"]

Out[68]: 0        3846.153846
         1        1500.241935
         2        1315.789474
         3        1351.351351
         4         679.984962
                     ...
         99995    1189.000000
         99996    3260.869565
         99997    3947.368421
         99998     716.982143
         99999    1562.500000
         Name: EMI_per_Month, Length: 100000, dtype: float64
```

Encoding-

Encoded the below categorical columns to numbers so that the model is able to understand and extract valuable information.

Zone-	Label Encoding
Existing Customer-	Label Encoding
Gender-	Label Encoding
City-	Catboost Encoding
Employment Profile-	Catboost Encoding
Occupation-	Catboost Encoding

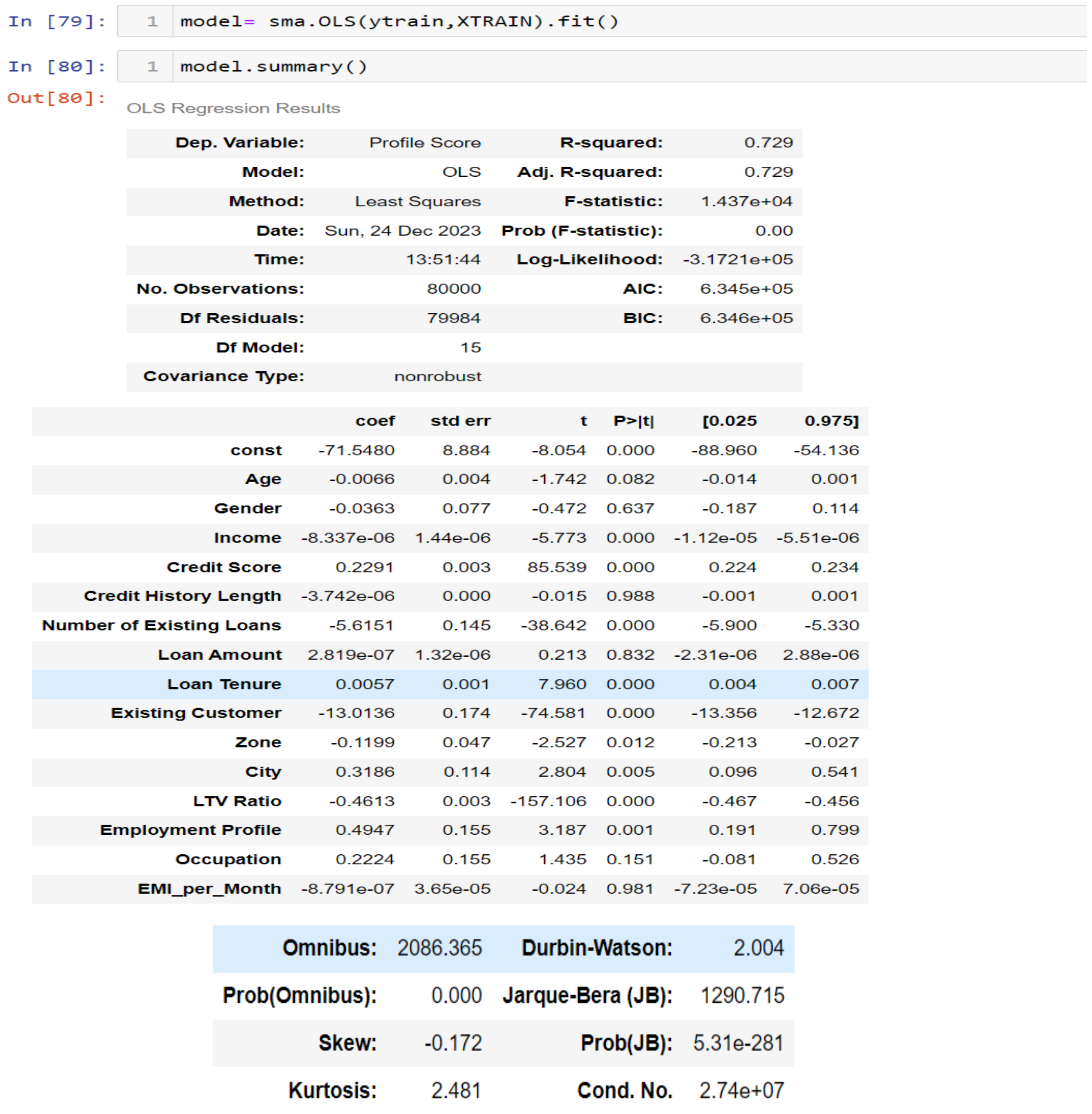
Statistically Significant Features:

We performed Mannwhitneyu test on all numerical and Krushkal wallis test on all categorical columns for identifying if they contribute towards explainingvariation in target variable as the data was not normal and found all variablesas significant except **Gender** and **Zone** on the basis of pValue.

DATA MODELING

We split the data in train and test in the ratio of 80:20.
We tried to fitting our data into various models such as OLS, Linear regression, Stochastic Gradient Descent Regressor, Decision Tree Regressor, Gradient Boosting Regressor.

Base Model: OLS



```
In [79]:  model= sma.OLS(ytrain,XTRAIN).fit()

In [80]:  model.summary()

Out[80]:
```

OLS Regression Results

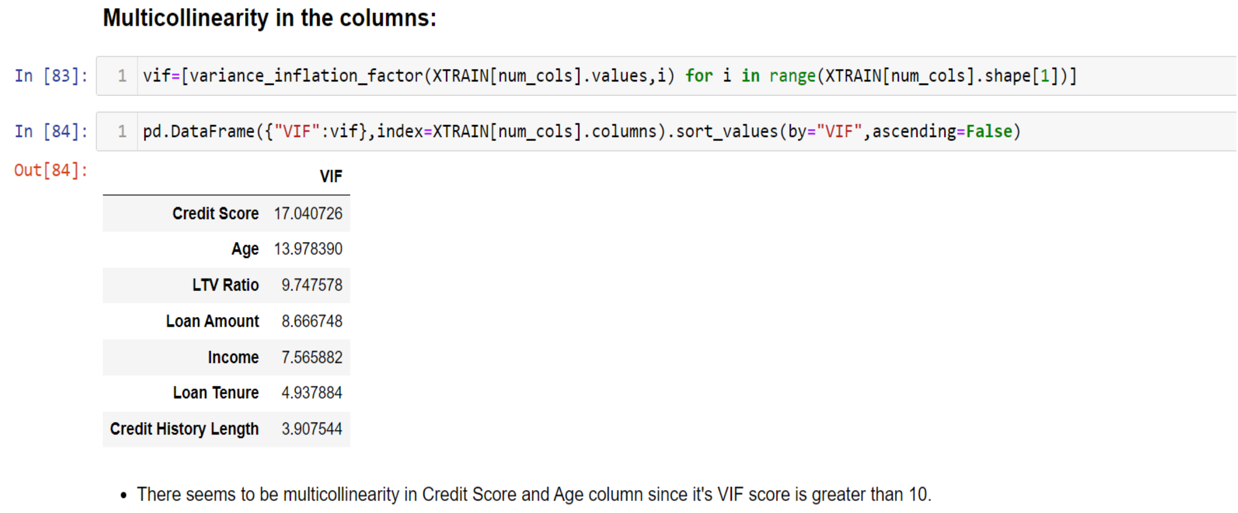
Dep. Variable:	Profile Score	**R-squared:**	0.729
Model:	OLS	**Adj. R-squared:**	0.729
Method:	Least Squares	**F-statistic:**	1.437e+04
Date:	Sun, 24 Dec 2023	**Prob (F-statistic):**	0.00
Time:	13:51:44	**Log-Likelihood:**	-3.1721e+05
No. Observations:	80000	**AIC:**	6.345e+05
Df Residuals:	79984	**BIC:**	6.346e+05
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-71.5480	8.884	-8.054	0.000	-88.960	-54.136
Age	-0.0066	0.004	-1.742	0.082	-0.014	0.001
Gender	-0.0363	0.077	-0.472	0.637	-0.187	0.114
Income	-8.337e-06	1.44e-06	-5.773	0.000	-1.12e-05	-5.51e-06
Credit Score	0.2291	0.003	85.539	0.000	0.224	0.234
Credit History Length	-3.742e-06	0.000	-0.015	0.988	-0.001	0.001
Number of Existing Loans	-5.6151	0.145	-38.642	0.000	-5.900	-5.330
Loan Amount	2.819e-07	1.32e-06	0.213	0.832	-2.31e-06	2.88e-06
Loan Tenure	0.0057	0.001	7.960	0.000	0.004	0.007
Existing Customer	-13.0136	0.174	-74.581	0.000	-13.356	-12.672
Zone	-0.1199	0.047	-2.527	0.012	-0.213	-0.027
City	0.3186	0.114	2.804	0.005	0.096	0.541
LTV Ratio	-0.4613	0.003	-157.106	0.000	-0.467	-0.456
Employment Profile	0.4947	0.155	3.187	0.001	0.191	0.799
Occupation	0.2224	0.155	1.435	0.151	-0.081	0.526
EMI_per_Month	-8.791e-07	3.65e-05	-0.024	0.981	-7.23e-05	7.06e-05

Omnibus:	2086.365	**Durbin-Watson:**	2.004
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1290.715
Skew:	-0.172	**Prob(JB):**	5.31e-281
Kurtosis:	2.481	**Cond. No.**	2.74e+07

Inference from Base Model (OLS Model):

Applied OLS as a base model where R2 Score is approximately 73% however the assumptions of the OLS model that is strong multicollinearity in the data exists which can be seen via condition number mentioned in the model summary which is 2.74e+07.
We have checked multicollinearity in the columns using VIF and seen Credit Score and Age have VIF score is greater than 10.



Multicollinearity in the columns:

```
In [83]: vif=[variance_inflation_factor(XTRAIN[num_cols].values,i) for i in range(XTRAIN[num_cols].shape[1])]
```

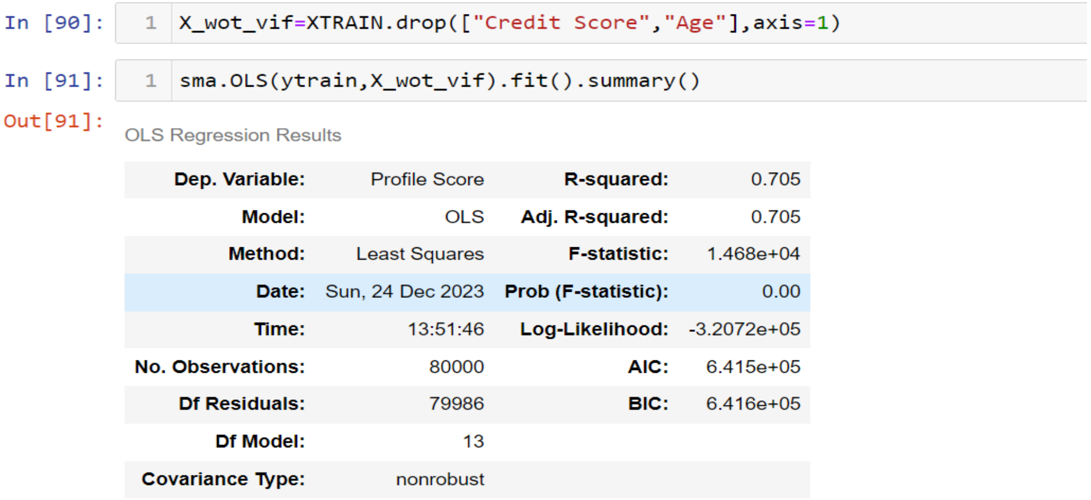
```
In [84]: pd.DataFrame({"VIF":vif},index=XTRAIN[num_cols].columns).sort_values(by="VIF",ascending=False)
```

Out[84]:

	VIF
Credit Score	17.040726
Age	13.978390
LTV Ratio	9.747578
Loan Amount	8.666748
Income	7.565882
Loan Tenure	4.937884
Credit History Length	3.907544

- There seems to be multicollinearity in Credit Score and Age column since it's VIF score is greater than 10.

Therefore, we are dropping Credit Score and Age column and applying the model on the basis of the rest of the predictors to check if the model meets the assumption or not.



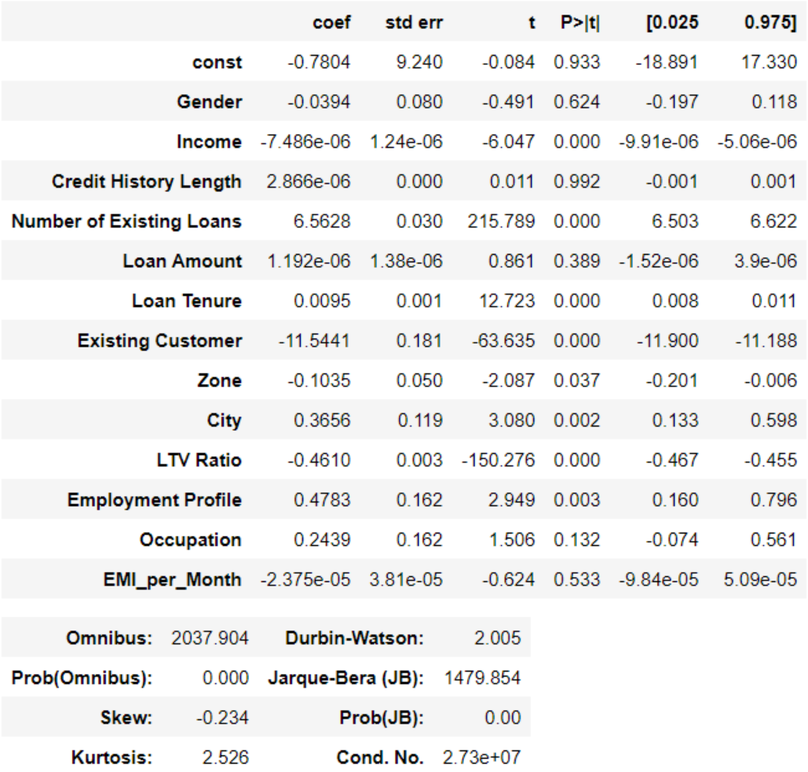
```
In [90]: X_wot_vif=XTRAIN.drop(["Credit Score","Age"],axis=1)
```

```
In [91]: sma.OLS(ytrain,X_wot_vif).fit().summary()
```

Out[91]:

OLS Regression Results

Dep. Variable:	Profile Score	**R-squared:**	0.705
Model:	OLS	**Adj. R-squared:**	0.705
Method:	Least Squares	**F-statistic:**	1.468e+04
Date:	Sun, 24 Dec 2023	**Prob (F-statistic):**	0.00
Time:	13:51:46	**Log-Likelihood:**	-3.2072e+05
No. Observations:	80000	**AIC:**	6.415e+05
Df Residuals:	79986	**BIC:**	6.416e+05
Df Model:	13		
Covariance Type:	nonrobust		



	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.7804	9.240	-0.084	0.933	-18.891	17.330
Gender	-0.0394	0.080	-0.491	0.624	-0.197	0.118
Income	-7.486e-06	1.24e-06	-6.047	0.000	-9.91e-06	-5.06e-06
Credit History Length	2.866e-06	0.000	0.011	0.992	-0.001	0.001
Number of Existing Loans	6.5628	0.030	215.789	0.000	6.503	6.622
Loan Amount	1.192e-06	1.38e-06	0.861	0.389	-1.52e-06	3.9e-06
Loan Tenure	0.0095	0.001	12.723	0.000	0.008	0.011
Existing Customer	-11.5441	0.181	-63.635	0.000	-11.900	-11.188
Zone	-0.1035	0.050	-2.087	0.037	-0.201	-0.006
City	0.3656	0.119	3.080	0.002	0.133	0.598
LTV Ratio	-0.4610	0.003	-150.276	0.000	-0.467	-0.455
Employment Profile	0.4783	0.162	2.949	0.003	0.160	0.796
Occupation	0.2439	0.162	1.506	0.132	-0.074	0.561
EMI_per_Month	-2.375e-05	3.81e-05	-0.624	0.533	-9.84e-05	5.09e-05

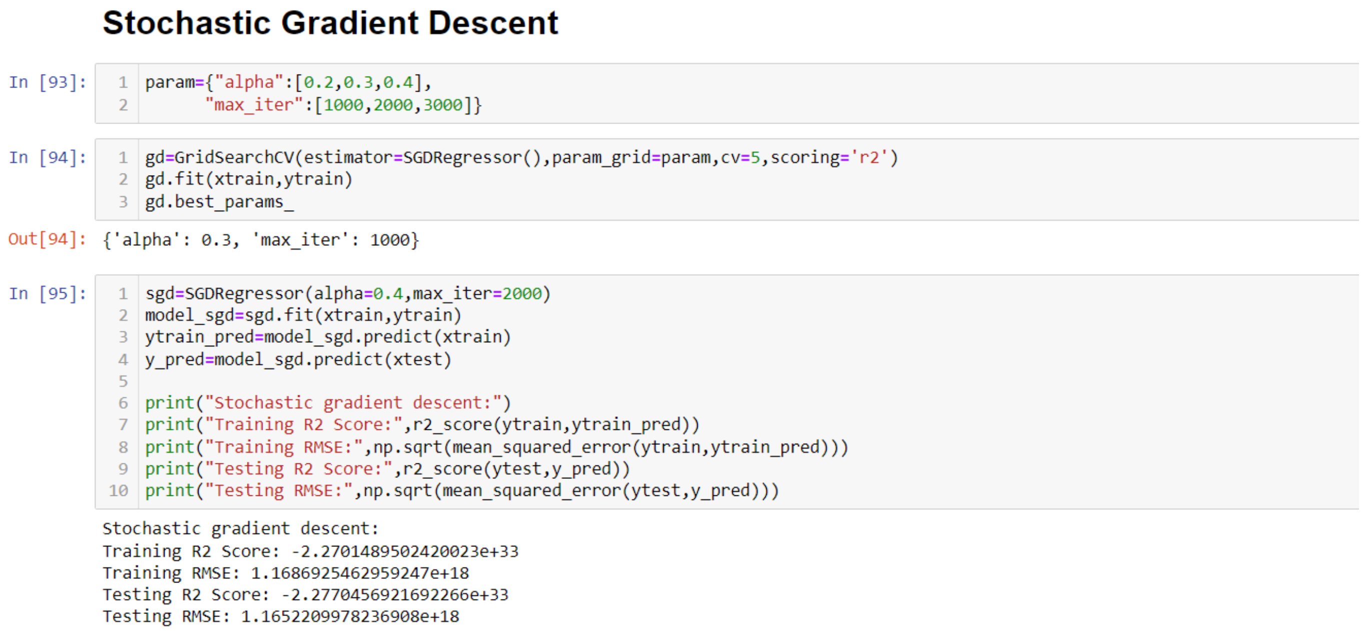
Omnibus:	2037.904	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1479.854
Skew:	-0.234	Prob(JB):	0.00
Kurtosis:	2.526	Cond. No.	2.73e+07

Inference of updated model:

R2 score of the model dropped and the multicollinearity in the data still exists. So, the conclusion came out to be that Linear regression model is not working best on the dataset.

Therefore, exploring other model for better performance.

Stochastic Gradient Descent



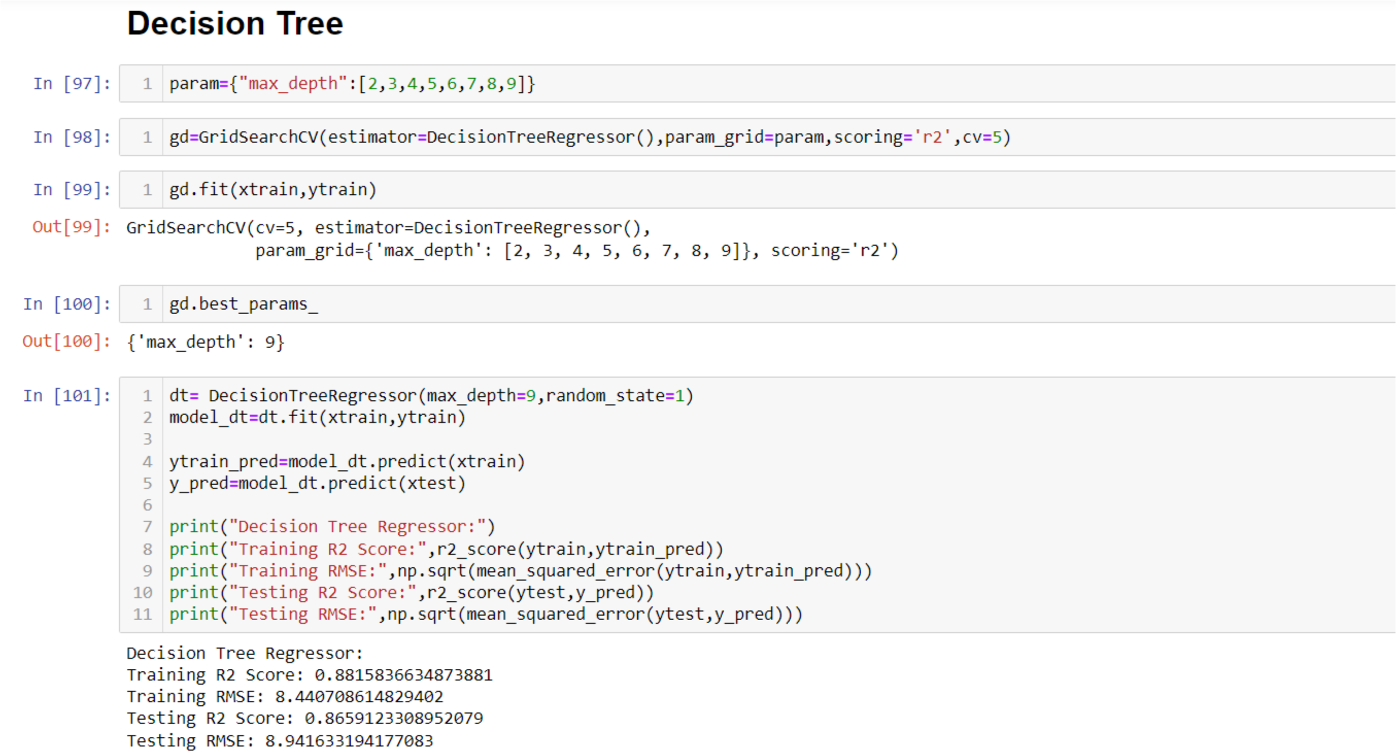
Stochastic Gradient Descent

```
In [93]: param={"alpha":[0.2,0.3,0.4],
               "max_iter":[1000,2000,3000]}

In [94]: gd=GridSearchCV(estimator=SGDRegressor(),param_grid=param,cv=5,scoring='r2')
         gd.fit(xtrain,ytrain)
         gd.best_params_

Out[94]: {'alpha': 0.3, 'max_iter': 1000}

In [95]: sgd=SGDRegressor(alpha=0.4,max_iter=2000)
         model_sgd=sgd.fit(xtrain,ytrain)
         ytrain_pred=model_sgd.predict(xtrain)
         y_pred=model_sgd.predict(xtest)

         print("Stochastic gradient descent:")
         print("Training R2 Score:",r2_score(ytrain,ytrain_pred))
         print("Training RMSE:",np.sqrt(mean_squared_error(ytrain,ytrain_pred)))
         print("Testing R2 Score:",r2_score(ytest,y_pred))
         print("Testing RMSE:",np.sqrt(mean_squared_error(ytest,y_pred)))

         Stochastic gradient descent:
         Training R2 Score: -2.2701489502420023e+33
         Training RMSE: 1.1686925462959247e+18
         Testing R2 Score: -2.2770456921692266e+33
         Testing RMSE: 1.1652209978236908e+18
```

The application of Stochastic Gradient Descent (SGD) on the dataset has resulted in a concerning phenomenon known as Underfitting.

Underfitting is a scenario in machine learning where a model is too simple to capture the underlying patterns in the data. Essentially, the model lacks the complexity or flexibility to represent the relationships between the features (input variables) and the target variable (output variable). As a result, an underfit model performs poorly on both the training data and new, unseen data.

Decision Tree



Decision Tree

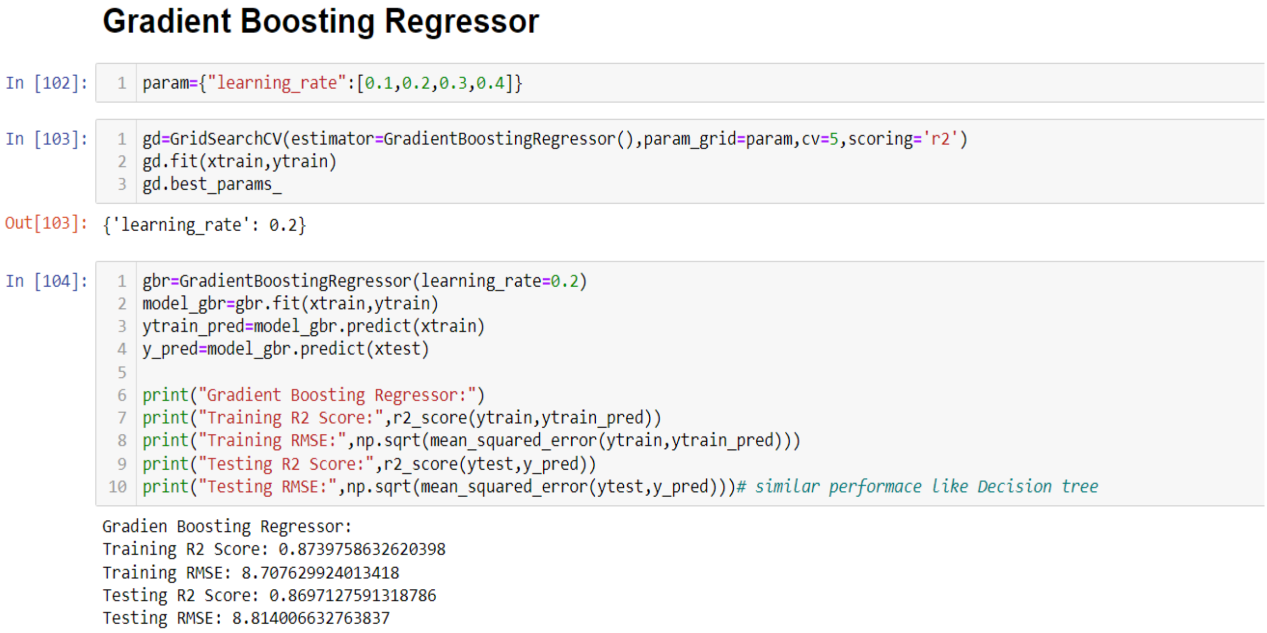
```
In [97]:  param={"max_depth":[2,3,4,5,6,7,8,9]}

In [98]:  gd=GridSearchCV(estimator=DecisionTreeRegressor(),param_grid=param,scoring='r2',cv=5)

In [99]:  gd.fit(xtrain,ytrain)

Out[99]:  GridSearchCV(cv=5, estimator=DecisionTreeRegressor(),
                       param_grid={'max_depth': [2, 3, 4, 5, 6, 7, 8, 9]}, scoring='r2')

In [100]: gd.best_params_

Out[100]: {'max_depth': 9}

In [101]: dt= DecisionTreeRegressor(max_depth=9,random_state=1)
          model_dt=dt.fit(xtrain,ytrain)

          ytrain_pred=model_dt.predict(xtrain)
          y_pred=model_dt.predict(xtest)

          print("Decision Tree Regressor:")
          print("Training R2 Score:",r2_score(ytrain,ytrain_pred))
          print("Training RMSE:",np.sqrt(mean_squared_error(ytrain,ytrain_pred)))
          print("Testing R2 Score:",r2_score(ytest,y_pred))
          print("Testing RMSE:",np.sqrt(mean_squared_error(ytest,y_pred)))

          Decision Tree Regressor:
          Training R2 Score: 0.8815836634873881
          Training RMSE: 8.440708614829402
          Testing R2 Score: 0.8659123308952079
          Testing RMSE: 8.941633194177083
```

The decision tree regressor appears to be a strong model for the given problem statement

The high training and testing R^2 Score is 0.8816 and 0.8659 respectively indicate that the Decision Tree Regressor is capturing a substantial portion of the variance in both the training and testing datasets.

The relatively low RMSE values is 8.4407 for training and 8.9416 for testing suggest that the model’s predictions are, on average, quite close to the actual values in both datasets.

Gradient Boosting Regressor



Gradient Boosting Regressor

```
In [102]:
param={"learning_rate":[0.1,0.2,0.3,0.4]}
```

```
In [103]:
gd=GridSearchCV(estimator=GradientBoostingRegressor(),param_grid=param,cv=5,scoring='r2')
gd.fit(xtrain,ytrain)
gd.best_params_
```

```
Out[103]: {'learning_rate': 0.2}
```

```
In [104]:
gbr=GradientBoostingRegressor(learning_rate=0.2)
model_gbr=gbr.fit(xtrain,ytrain)
ytrain_pred=model_gbr.predict(xtrain)
y_pred=model_gbr.predict(xtest)

print("Gradient Boosting Regressor:")
print("Training R2 Score:",r2_score(ytrain,ytrain_pred))
print("Training RMSE:",np.sqrt(mean_squared_error(ytrain,ytrain_pred)))
print("Testing R2 Score:",r2_score(ytest,y_pred))
print("Testing RMSE:",np.sqrt(mean_squared_error(ytest,y_pred)))# similar performace like Decision tree
```

```
Gradien Boosting Regressor:
Training R2 Score: 0.8739758632620398
Training RMSE: 8.707629924013418
Testing R2 Score: 0.8697127591318786
Testing RMSE: 8.814006632763837
```

Overall Inference:

Both the training and testing R² scores are high, suggesting that the Gradient Boosting Regressor and Decision Tree Regressor is effective at explaining the variance in both the training and testing datasets.

The RMSE values for both training and testing sets are relatively low, indicating that the model's predictions are, on average, close to the actual values in both datasets.

Considerations:

The R² scores and RMSE values are quite comparable between the training and testing sets, suggesting good generalization.

In summary, the Gradient Boosting Regressor and Decision Tree Regressor appears to be a robust model for the given problem, demonstrating high performance on both the training and testing datasets.

Evaluation

We evaluated our models using r squared, root mean squared error and found **Decision Tree Regressor** to be the best model giving highest r squared value and lowest root mean squared error.

Model	**Training R2**	**Training RMSE**	**Testing R2**	**Testing RMSE**
Stochastic Gradient Descent	-2.2701489502420023e+33	1.1686925462959247e+18	-2.2770456921692266e+33	1.1652209978236908e+18
Decision Tree	0.8815836634873881	8.440708614829402	0.8659123308952079	8.941633194177083
Gradient Boosting Regressor	0.8739758632620398	8.707629924013418	0.8697127591318786	8.814006632763837

Notes For Project Team

Sample Reference for Datasets (to be filled by team and mentor)

Original owner of data	Kaggle
Data set information	Details of the applicant who applied for two wheelers loan
Previous relevant journals used the data set	Not yet discussed
Link to web page	https://www.kaggle.com/datasets/yashkmd/credit-profile-two-wheeler-loan-dataset