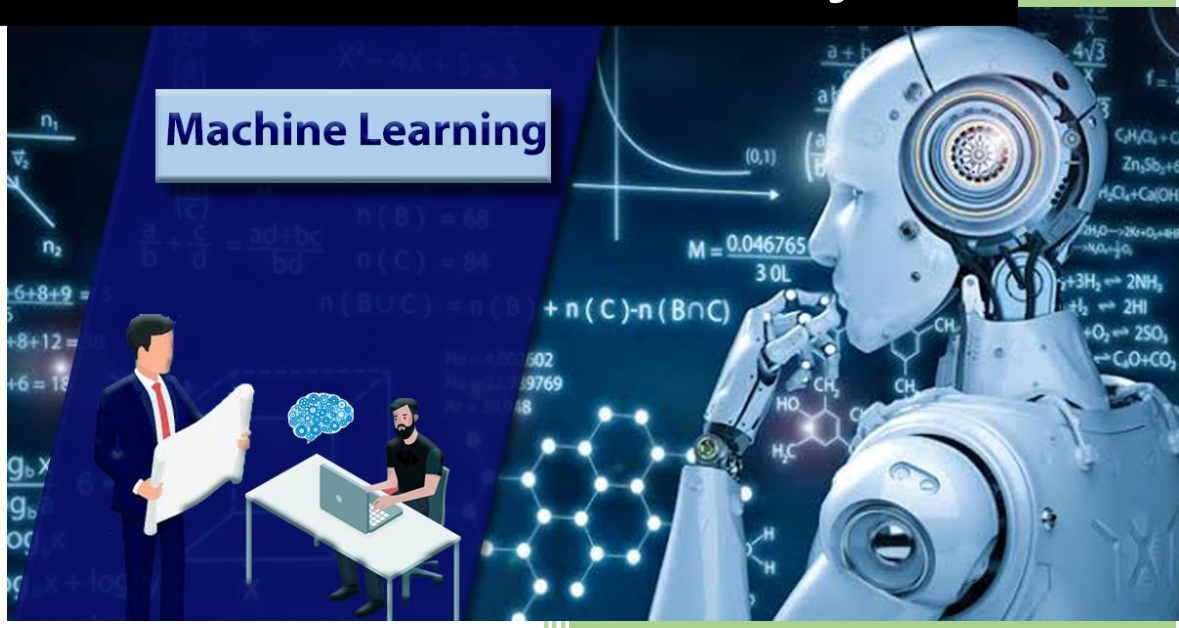
2023

Report For Loan Amount Prediction Project



By Shashank

Minor Project

8/1/2023

Introduction

Report Title: An In-depth Analysis of Loan Amount Prediction Models

In the contemporary financial landscape, access to loans plays a pivotal role in the realization of personal and professional aspirations. Whether it's a first-time homebuyer seeking a mortgage, a budding entrepreneur in need of startup capital, or an individual looking to finance their education, the process of securing a loan is a critical facet of financial planning. Lenders are tasked with assessing the creditworthiness of applicants and determining the optimal loan amount to disburse.

The objective of this report, titled "An In-depth Analysis of Loan Amount Prediction Models," is to explore and evaluate various machine learning models to predict loan amounts accurately. Accurate loan amount predictions are fundamental for both borrowers and lenders. For borrowers, it ensures they receive the financial assistance they need, while for lenders, it minimizes the risk associated with default.

The report primarily focuses on three distinct regression models for loan amount prediction:

1. **Linear Regression:** This model seeks to establish a linear relationship between various applicant attributes and the loan amount. It is a fundamental and widely used approach for predictive modeling.

2. **Polynomial Regression:** Polynomial regression introduces nonlinear relationships between the predictors and the loan amount by including polynomial terms. This approach allows for a more flexible modeling of complex data patterns.

3. **Random Forest Regression:** Random forest regression leverages an ensemble of decision trees to predict loan amounts. This model excels in capturing intricate relationships within the data.

In addition to model selection, this report encompasses critical phases of the data analysis process, including data preprocessing, feature engineering, and model evaluation. A key focus is on assessing the accuracy and generalization capabilities of each model through metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R2) score.

Ultimately, the objective of this study is to identify the most effective loan amount prediction model that can be deployed in real-world financial settings. This selection is critical for banks, financial institutions, and credit agencies to streamline their lending processes and ensure better financial outcomes for applicants.

The subsequent sections of this report delve into the methodology, results, and discussion, culminating in a comprehensive recommendation for the ideal loan amount prediction model. The findings presented in this report aim to contribute to the advancement of predictive modeling in the finance sector, ultimately benefiting both lenders and borrowers alike.

Data Description

The dataset at hand is a treasure trove of valuable information, encompassing various crucial attributes that play a pivotal role in the realm of loan predictions. It offers insights into an array of factors that can significantly impact the lending process, enabling us to unlock the power of predictive modeling.

The dataset includes essential features such as:

- **Gender**: Providing insights into the gender of loan applicants.
- **Marital Status**: Shedding light on the marital status of applicants.
- **Education**: Indicating the educational background of individuals.
- **Number of Dependents**: Revealing the count of dependents associated with the applicant.
- **Income**: Offering a glimpse into the financial stability of loan seekers.
- **Loan Amount**: The focal point of our prediction, representing the desired loan amount.
- **Credit History**: A critical factor influencing loan approvals, denoting the creditworthiness of applicants.

These attributes, along with others present in the dataset, serve as the building blocks for our predictive modeling journey. By harnessing the power of this data, we aim to develop robust machine learning models that can effectively forecast loan amounts. With this newfound knowledge, we can empower financial institutions, individuals, and decision-makers to make informed choices and streamline the loan application and approval process.

As we dive deeper into the world of data analysis and machine learning, this dataset will be our guiding compass, helping us navigate through the intricate landscape of loan amount prediction.

- **Prediction Task**: Our primary goal in this project is to predict the output variable "Loan Amount."
- **Input Attributes**: The input features encompass a mix of categorical and numerical data. This diversity in data types requires us to employ versatile techniques to handle and analyze them effectively.
- **Attribute Analysis**: We will conduct a thorough analysis of all the attributes present in the dataset, exploring their relationships, distributions, and significance. This comprehensive exploration will provide us with valuable insights for our predictive modeling journey.

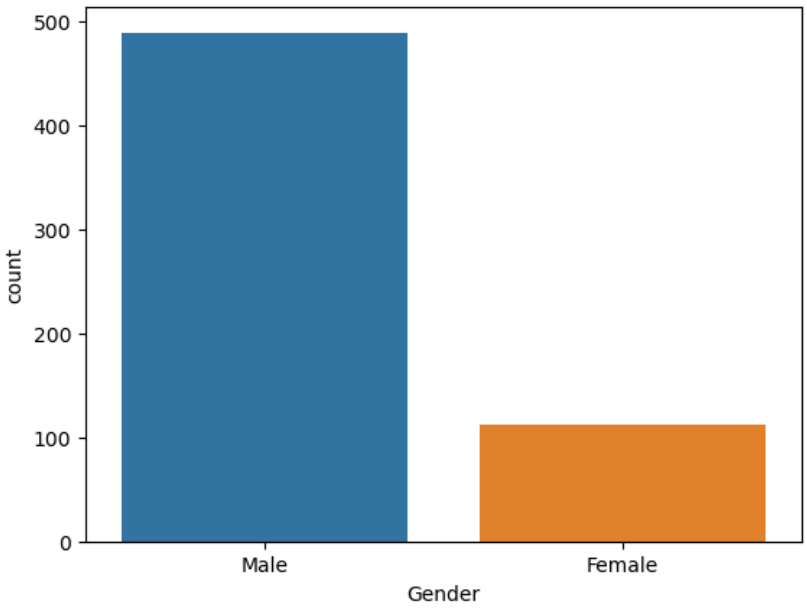
	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...
609	LP002978	Female	No	0	Graduate	No	2900	0.0	71.0	360.0	1.0
610	LP002979	Male	Yes	3+	Graduate	No	4106	0.0	40.0	180.0	1.0
611	LP002983	Male	Yes	1	Graduate	No	8072	240.0	253.0	360.0	1.0
612	LP002984	Male	Yes	2	Graduate	No	7583	0.0	187.0	360.0	1.0
613	LP002990	Female	No	0	Graduate	Yes	4583	0.0	133.0	360.0	0.0

Our DataSet

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

Descriptive statistics for the DataFrame, which includes count, mean, standard deviation, and more.

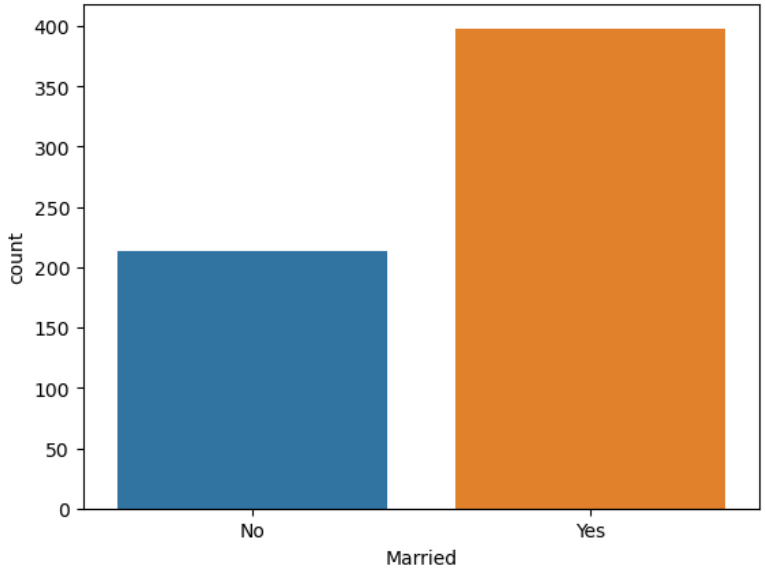
Visualization of Our Dataset:



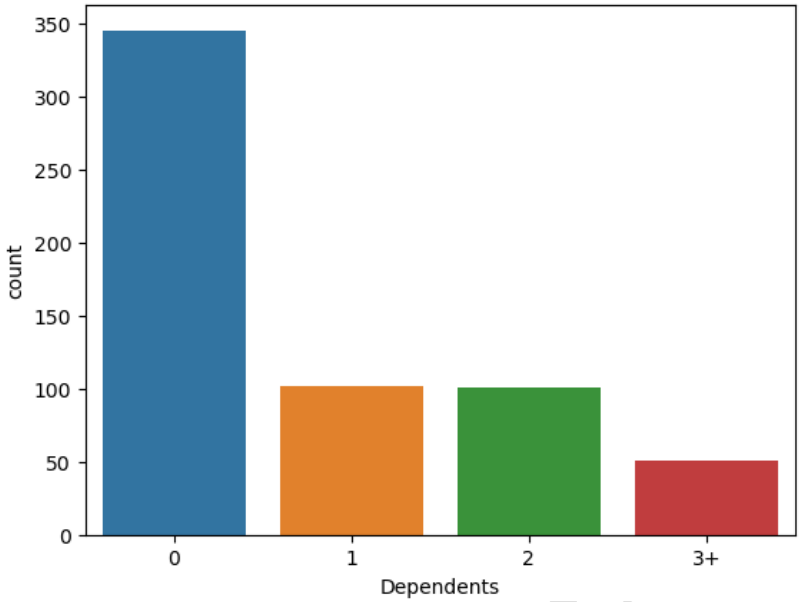
- Upon analyzing the 'Gender' attribute, we observe that the majority of the applicants are classified as male, while only a few are categorized as female. This initial insight provides valuable intuition about the distribution of gender among loan applicants.

- Such observations and intuitions gained from data exploration are fundamental in building a robust predictive model, as they allow us to understand the characteristics and trends within the dataset, which can be leveraged for making informed decisions and constructing effective models.

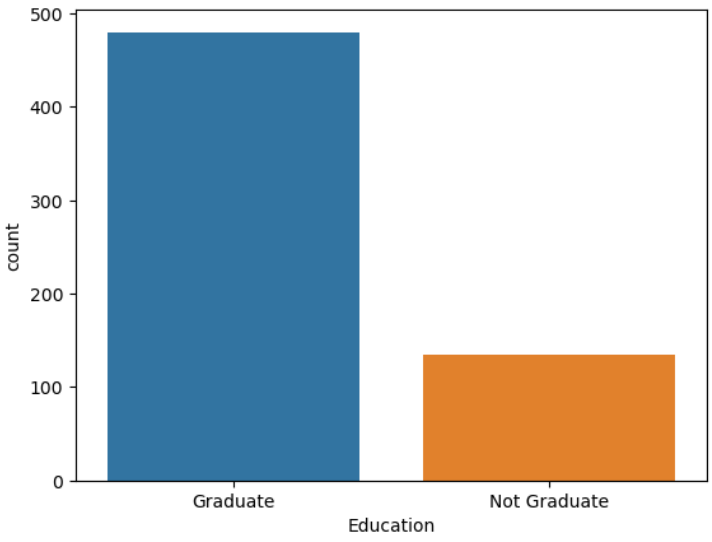
-



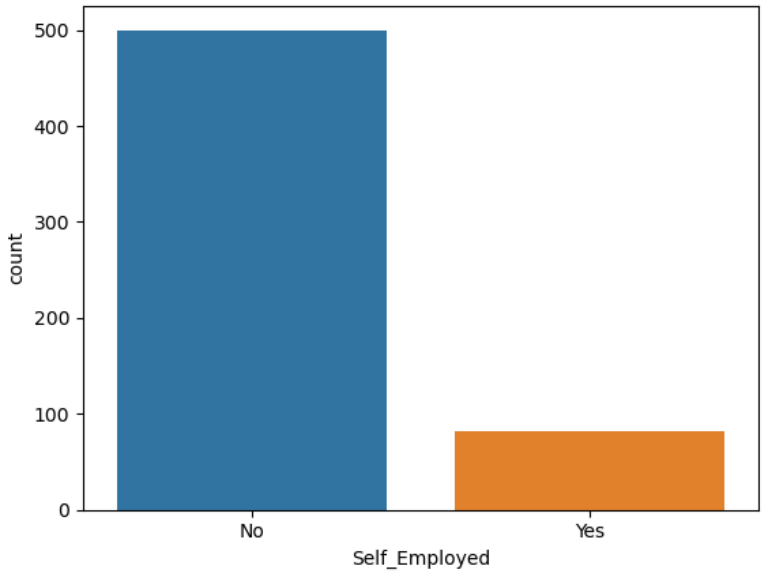
- The majority of the applicants are married.



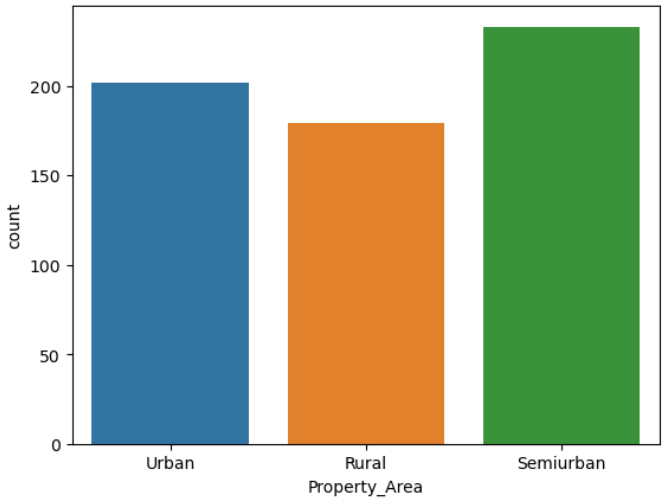
- The majority of the applicants have zero dependents, around 100 applicants have one or two dependents and only a few have more than three dependents.



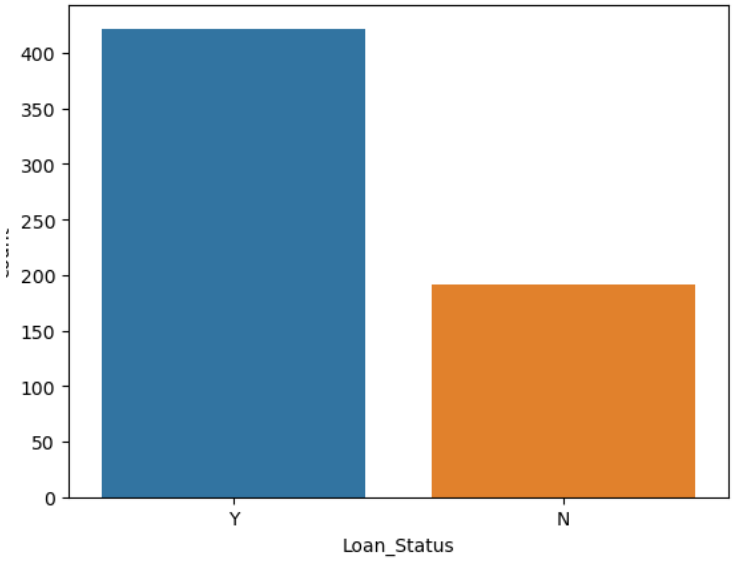
- Majority of people are Graduated



- Around 90 applicants are either freelancers or run a business.

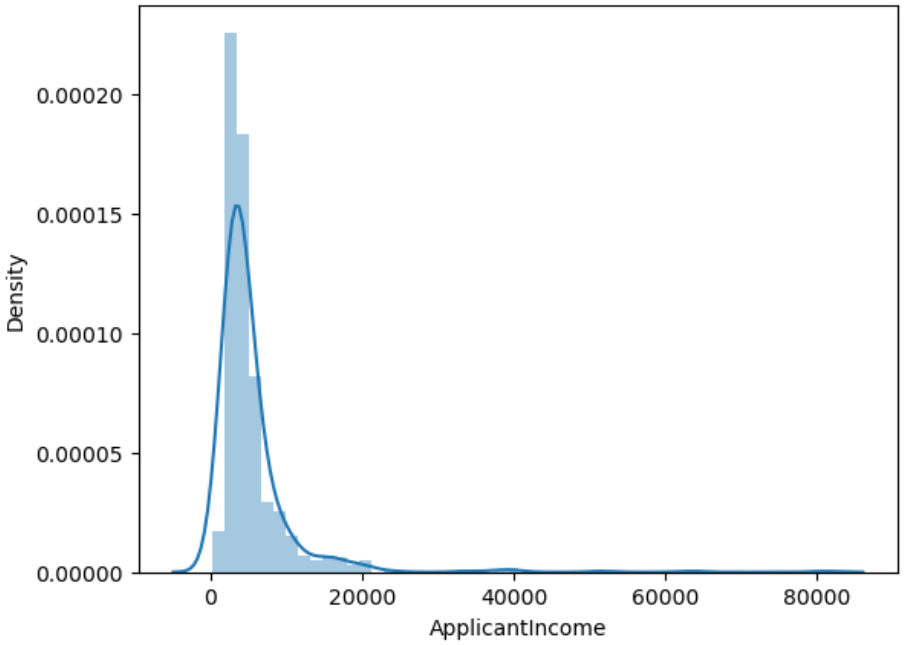


- We can assume that the applicants are equally distributed in urban, rural and semi-urban areas.

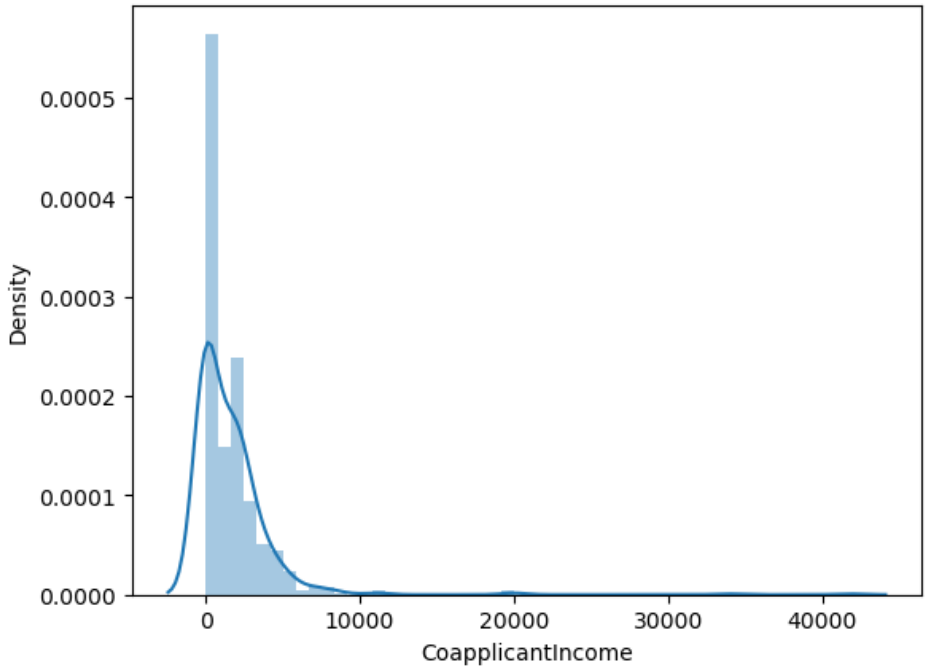


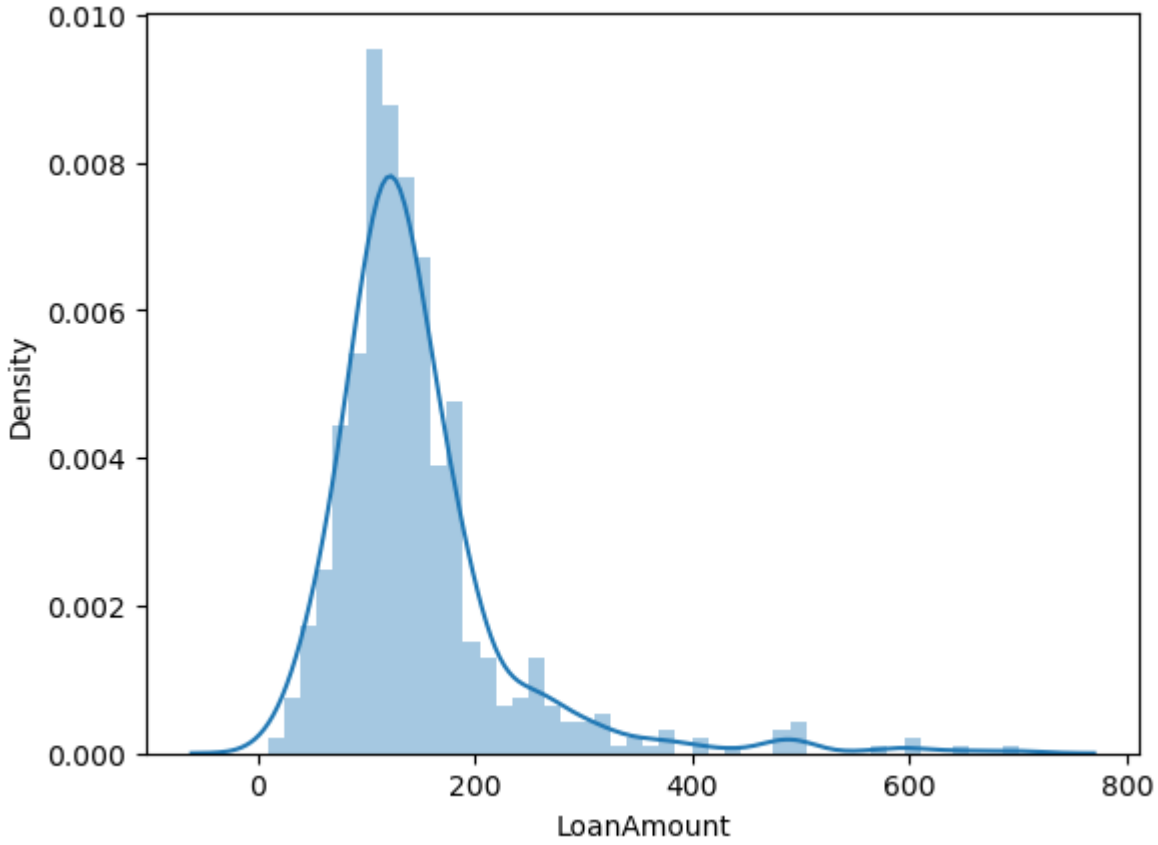
- Around 400 loans are accepted and 200 loans are rejected. Its shows the 2:1 ratio.

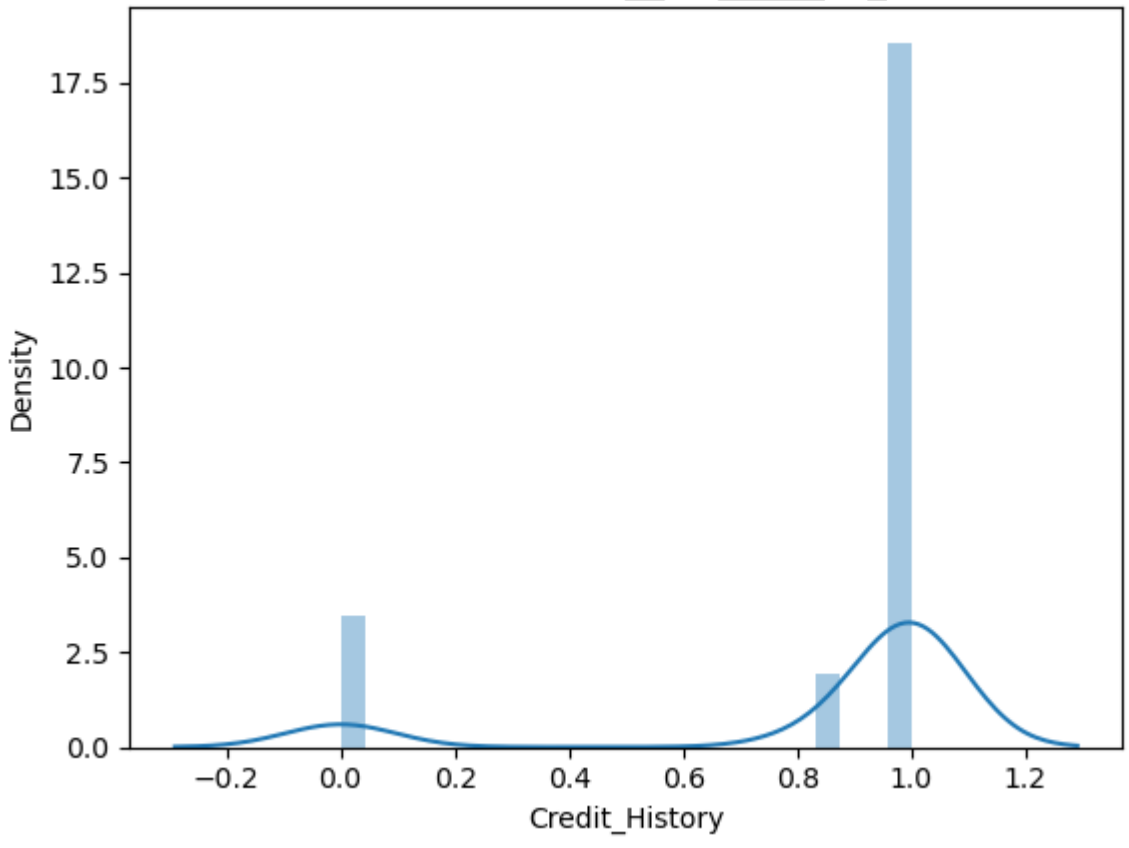
LET US FIRST EXPLORE THE NUMERICAL COLUMN "APPLICANT INCOME".

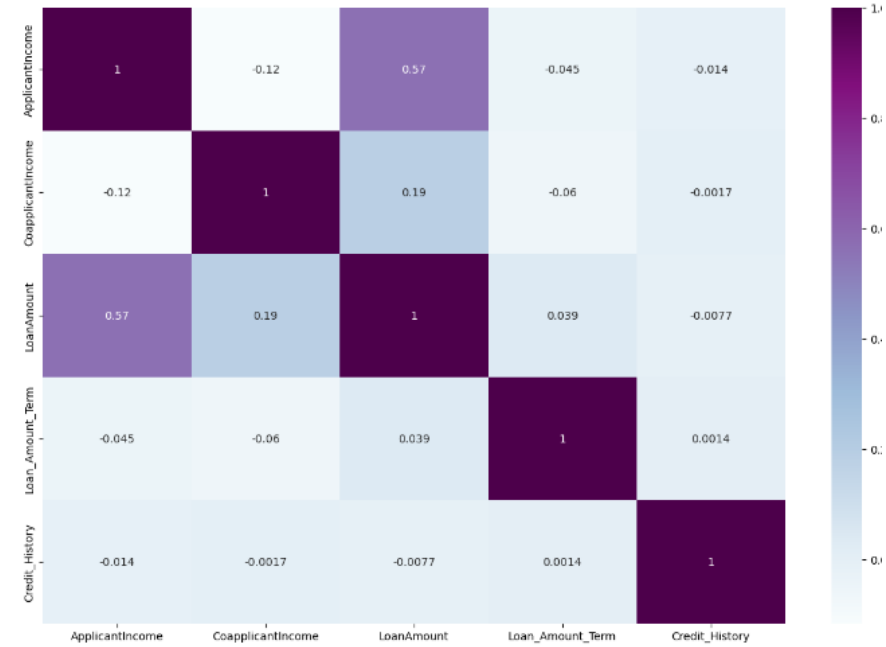


- The data are skewed left in the graph, which is not a suitable distribution to train a Model.
- Hence, we will apply the Log Transformation later to normalize the attributes in the form of Bell Curve (Normal Distribution).



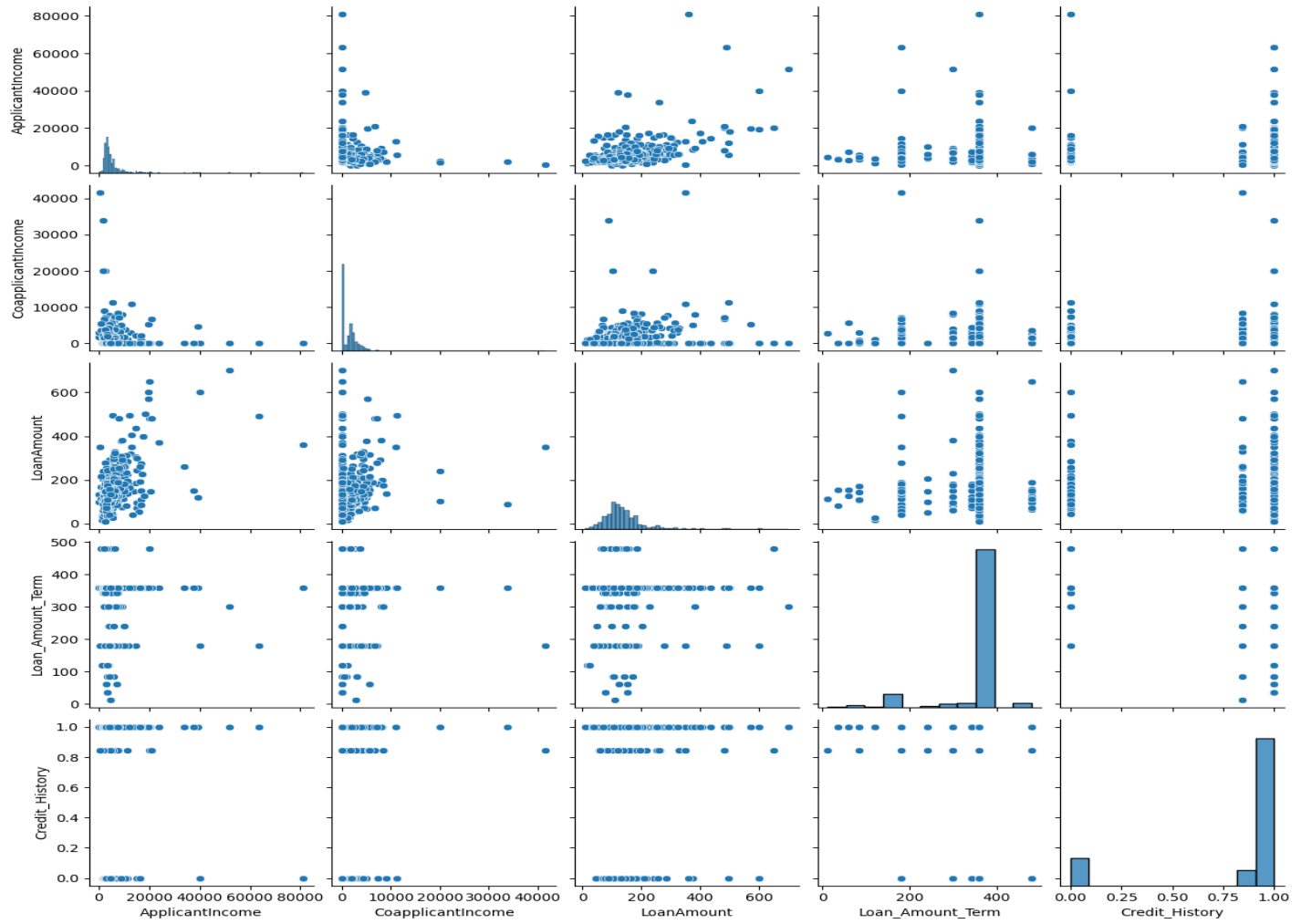

0.010
0.008
0.006
0.004
0.002
0.000
Density
0
200
400
600
800
LoanAmount


17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Density
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Credit_History



Correlation Matrix

- In this graph, the higher density is plotted with dark color and the lower density is plotted with light color.
- We need to remove the highly correlated attributes.
- It means the original attributes are correlated with log attributes.
- We will remove the previous attributes and keep the log attributes to train our model.



Approach

This project demonstrates a comprehensive data analysis and predictive modeling process for loan amount prediction. Here's a summary of the steps and key components of the code:

1. Import Libraries: The code starts by importing essential libraries for data manipulation and analysis, data visualization, data preprocessing, and machine learning modeling. These libraries include NumPy, Pandas, Matplotlib, Seaborn, and various modules from scikit-learn.

2. Dataset Information: The code provides an introduction to the dataset used in the project. It outlines the key attributes within the dataset that will be leveraged for loan amount prediction. This section offers a brief overview of the data's significance.

3. Loading the Dataset: The dataset is loaded from a CSV file using Pandas, and the resulting DataFrame (`df`) is displayed.

4. Data Preprocessing: This section focuses on data preprocessing tasks, including handling missing values. Missing values are addressed using mean values for some columns and mode values for others. The `LabelEncoder` is applied to categorical columns, converting them into numerical format. Numerical features are standardized and scaled using Min-Max scaling and StandardScaler.

5. Data Splitting: The code splits the dataset into training and testing sets using the `train_test_split` function.

6. Model Training: Three regression models are trained and evaluated in this code:

- **Linear Regression**
- **Polynomial Regression**
- **Random Forest Regression**

7. Predicting Loan Amount: The trained models are used to make predictions on a new dataset (`new_data`) with the same preprocessing steps applied.

8. Displaying Regression Metrics: Metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R2) score are calculated and displayed for each model to evaluate their predictive performance.

9. Feature Importance: For the Random Forest Regression model, feature importances are extracted and displayed to understand which features have the most impact on the predictions.

10. Visualization: The code includes visualization of the actual vs. predicted loan amounts for each of the trained regression models, allowing for a visual comparison of their performance.

Algorithms

In the provided code, the following algorithms are used for predictive modeling:

1. Linear Regression:

- Linear regression is a simple and commonly used algorithm for modeling the relationship between a dependent variable (in this case, "Loan Amount") and one or more independent variables. It aims to find a linear equation that best fits the data.

2. Polynomial Regression:

- Polynomial regression is an extension of linear regression that models the relationship between variables as an nth-degree polynomial. It is used when the relationship between the independent and dependent variables is not linear. In the code, a second-degree polynomial regression is applied.

3. Random Forest Regression:

- Random forest regression is an ensemble learning method that combines multiple decision trees to make predictions. It is particularly effective for handling complex relationships between variables and is known for its robustness and ability to capture feature importance.

These algorithms are applied to predict the "Loan Amount" based on various features in the dataset. After training the models, they are evaluated using regression metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R2) score to assess their performance. The code provides a comparative analysis of the models' predictions and their respective metrics.

Evaluation:

In the provided code, the predictive models are evaluated using regression metrics to assess their performance. The following regression metrics are used for evaluation:

1. **Mean Squared Error (MSE):**

- MSE measures the average of the squared differences between predicted and actual values. It quantifies how close the predicted values are to the actual values. A lower MSE indicates a better model fit.

2. **Mean Absolute Error (MAE):**

- MAE calculates the average of the absolute differences between predicted and actual values. It provides a measure of the average prediction error, and, like MSE, lower values are desirable.

3. **R-squared (R2) Score:**

- R-squared is a statistical measure that represents the proportion of the variance in the dependent variable (Loan Amount) that is explained by the independent variables (features). It ranges from 0 to 1, with higher values indicating a better fit. An R2 score of 1 means that the model perfectly fits the data.

Linear Regression Evaluation

Mean Squared Error: 2608.0176444372432
Mean Absolute Error: 37.287515077866054
R-squared (R2) Score: 0.5209385325151856

Multiple Regression Evaluation

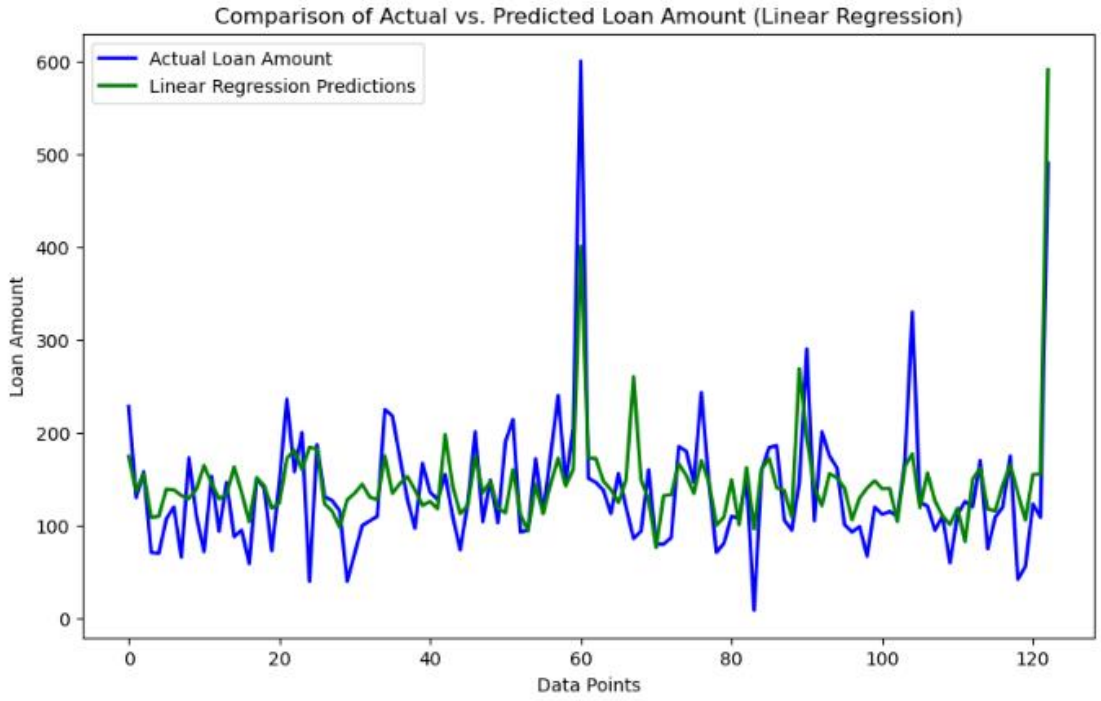
Mean Squared Error: 2608.0176444372432
Mean Absolute Error: 37.287515077866054
R-squared (R2) Score: 0.5209385325151856

Predicted Loan Amount for the test Data using Linear regression

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	\
0	LP001015	1	1	0	0	0	
1	LP001022	1	1	1	0	0	
2	LP001031	1	1	2	0	0	
3	LP001035	1	1	2	0	0	
4	LP001051	1	0	0	1	0	
..	...	...	...	...	...	...	
362	LP002971	1	1	3	1	1	
363	LP002975	1	1	0	0	0	
364	LP002980	1	0	0	0	0	
365	LP002986	1	1	0	0	0	
366	LP002989	1	0	0	0	1	

	ApplicantIncome	CoapplicantIncome	Loan_Amount_Term	Credit_History	\
0	0.051857	-0.554487	0.279851	0.451640	
1	-0.381297	-0.041468	0.279851	0.451640	
2	-0.066097	0.061136	0.279851	0.451640	
3	-0.501872	0.316278	0.279851	-0.047954	
4	-0.348532	-0.554487	0.279851	0.451640	
..	...	...	...	...	
362	-0.228448	0.053270	0.279851	0.451640	
363	-0.204038	-0.312000	0.279851	0.451640	
364	-0.352791	0.127145	0.279851	-0.047954	
365	-0.066097	0.263950	0.279851	0.451640	
366	0.621969	-0.554487	-2.518655	0.451640	

	Property_Area	PredictedLoanAmount
0	2	-43.530979
1	2	-174.286201
2	2	55.006871
3	2	-151.441217
4	2	-329.147769
..	...	...
362	2	-55.684104
363	2	-138.920747
364	1	-125.409058
365	0	112.620306
366	0	166.421477



Polynomial Regression Evaluation

Mean Squared Error: 11557.959421658668
Mean Absolute Error: 46.0533653413905
R-squared (R2) Score: -1.123058106404991

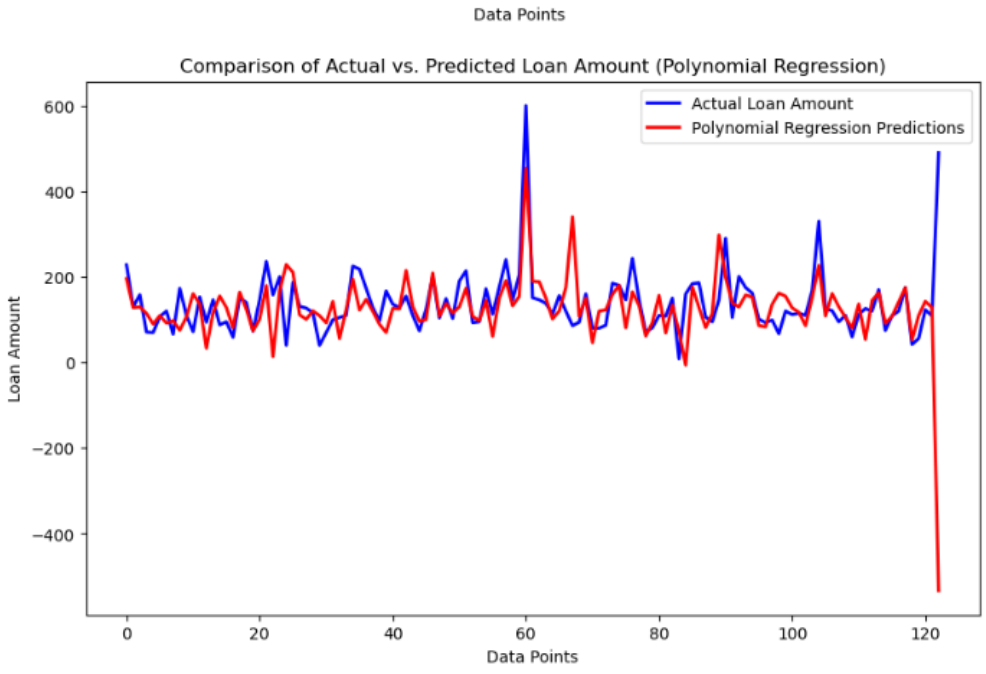
Predicted Loan Amount for the test Data using Polynomial regression

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	\
0	LP001015	1	1	0	0	0	
1	LP001022	1	1	1	0	0	
2	LP001031	1	1	2	0	0	
3	LP001035	1	1	2	0	0	
4	LP001051	1	0	0	1	0	
..	...	...	...	...	...	...	
362	LP002971	1	1	3	1	1	
363	LP002975	1	1	0	0	0	
364	LP002980	1	0	0	0	0	
365	LP002986	1	1	0	0	0	
366	LP002989	1	0	0	0	1	

	ApplicantIncome	CoapplicantIncome	Loan_Amount_Term	Credit_History	\
0	0.051857	-0.554487	0.279851	0.451640	
1	-0.381297	-0.041468	0.279851	0.451640	
2	-0.066097	0.061136	0.279851	0.451640	
3	-0.501872	0.316278	0.279851	-0.047954	
4	-0.348532	-0.554487	0.279851	0.451640	
..	...	...	...	...	
362	-0.228448	0.053270	0.279851	0.451640	
363	-0.204038	-0.312000	0.279851	0.451640	
364	-0.352791	0.127145	0.279851	-0.047954	
365	-0.066097	0.263950	0.279851	0.451640	
366	0.621969	-0.554487	-2.518655	0.451640	

	Property_Area	PredictedLoanAmount
0	2	-726.798882
1	2	-912.149453
2	2	-39.285223
3	2	-897.245272
4	2	-1481.029246
..	...	...
362	2	-345.932455
363	2	-830.728552
364	1	-463.754186
365	0	235.888191
366	0	-3215.106368

[367 rows x 12 columns]



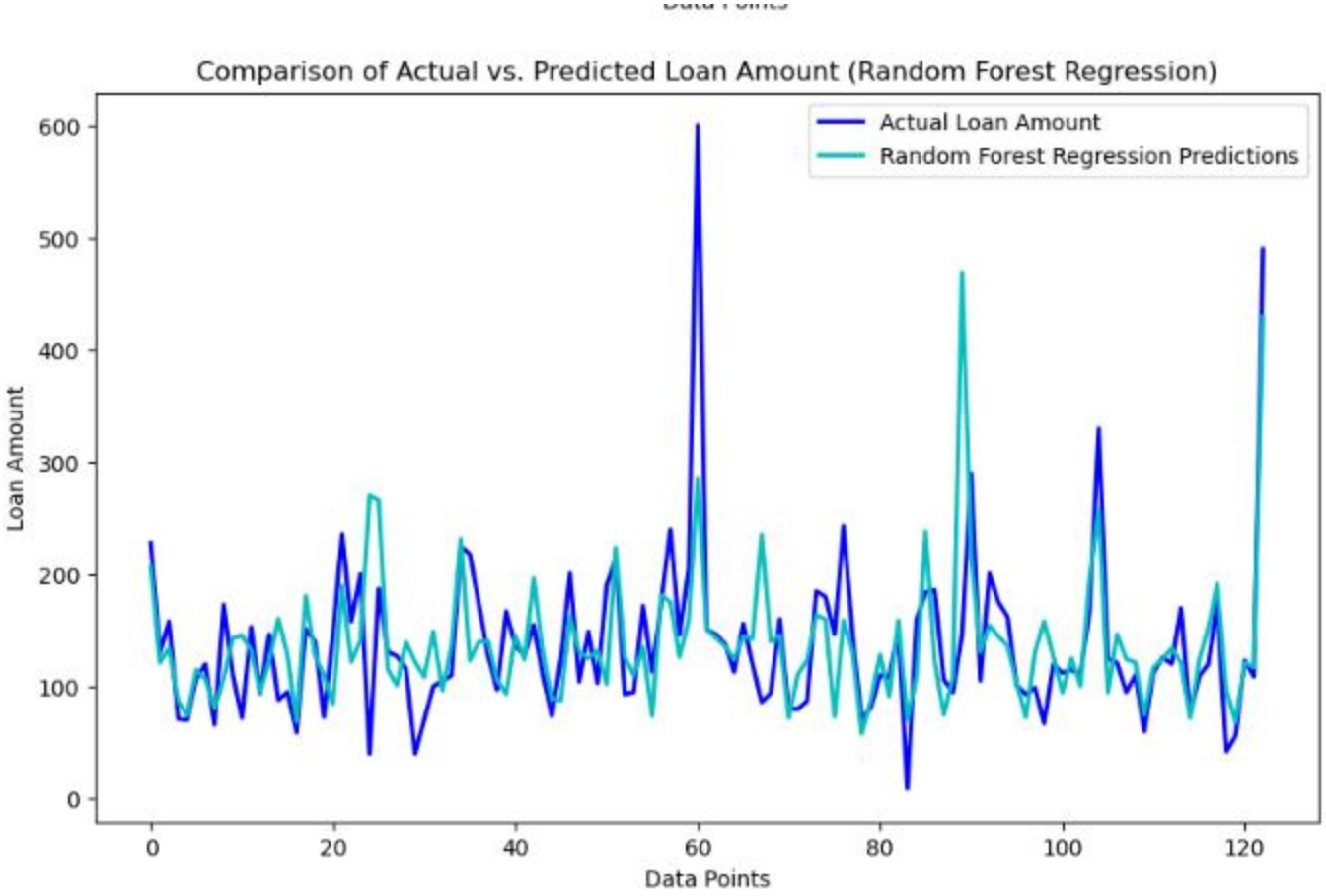
Random Forest Regression Evaluation

Mean Squared Error: 3627.162287157481
Mean Absolute Error: 36.88795030396249
R-squared (R2) Score: 0.33373392170198124

Predicted Loan Amount for the test Data using Random Forest Regression

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	Loan_Amount_Term	Credit_History	Property_Area	PredictedLoanAmount
0	LP001015	1	1	0	0	0	0.051857	-0.554487	0.279851	0.451640	2	110.668243
1	LP001022	1	1	1	0	0	-0.381297	-0.041468	0.279851	0.451640	2	35.658243
2	LP001031	1	1	2	0	0	-0.066097	0.061136	0.279851	0.451640	2	117.968243
3	LP001035	1	1	2	0	0	-0.501872	0.316278	0.279851	-0.047954	2	240.490000
4	LP001051	1	0	0	1	0	-0.348532	-0.554487	0.279851	0.451640	2	38.552365
..	...	...	...	...	...	...	...	...	...	...	...	...
362	LP002971	1	1	3	1	1	-0.228448	0.053270	0.279851	0.451640	2	120.386486
363	LP002975	1	1	0	0	0	-0.204038	-0.312000	0.279851	0.451640	2	37.132365
364	LP002980	1	0	0	0	0	-0.352791	0.127145	0.279851	-0.047954	1	191.084122
365	LP002986	1	1	0	0	0	-0.066097	0.263950	0.279851	0.451640	0	203.990000
366	LP002989	1	0	0	0	1	0.621969	-0.554487	-2.518655	0.451640	0	338.380000

[367 rows x 12 columns]



Comparison

MODEL SELECTION FOR LOAN AMOUNT PREDICTION

When selecting the best model for loan amount prediction, several factors come into play, including the nature of the dataset, the problem's complexity, and the choice of evaluation criteria. In general, it's crucial to evaluate models using key metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R2) score. Lower MSE and MAE values and a higher R2 score typically indicate better model performance.

MODEL COMPARISON

Let's briefly compare the three models we've explored for loan amount prediction:

LINEAR REGRESSION:

- **Pros**: Simplicity, interpretability, and computational efficiency.
- **Cons**: Assumes a linear relationship between features and the target variable, which may not capture complex patterns.
- **Best Use Case**: When the relationship is primarily linear.

POLYNOMIAL REGRESSION:

- **Pros**: Can capture nonlinear relationships using polynomial terms.

- **Cons**: Complexity and potential overfitting with high polynomial degrees.
- **Best Use Case**: When the relationship is nonlinear and polynomial terms can better explain the data.

RANDOM FOREST REGRESSION:

- **Pros**: Versatile for handling complex, nonlinear relationships, high-dimensional data, and feature interactions. Helps mitigate overfitting.
- **Cons**: More complex and challenging to interpret.
- **Best Use Case**: When the relationship is complex, and a combination of linear and nonlinear patterns is present.

MODEL SELECTION STEPS

To determine the best model for loan amount prediction, we can follow these steps:

1. Split the dataset into training and testing sets.
2. Train each of the three models on the training data.
3. Evaluate the models using appropriate metrics (MSE, MAE, R2 score) on the testing data.
4. Choose the model with the lowest MSE and MAE and the highest R2 score.

These steps will help us make an informed decision about the most suitable model for our specific loan amount prediction problem.

Remember that model selection should be based on a combination of performance metrics, problem requirements, and the interpretability of the chosen model.

Model	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	R-squared (R2) Score
Linear Regression	2608.017644	37.287515	0.520939
Polynomial Regression	11557.959422	46.053365	-1.123058
Random Forest Regre ssion	3627.162287	36.887950	0.333734

Result and discussion

In this analysis, we explored and compared the performance of three different regression models: Linear Regression, Polynomial Regression, and Random Forest Regression, for the task of predict ing "Loan Amount." The models were evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R2) Score. Below are the results:

Linear Regression

- Mean Squared Error (MSE): 2608.02
- Mean Absolute Error (MAE): 37.29
- R-squared (R2) Score: 0.52

Polynomial Regression

- Mean Squared Error (MSE): 11557.96
- Mean Absolute Error (MAE): 46.05
- R-squared (R2) Score: -1.12

Random Forest Regression

- Mean Squared Error (MSE): 3627.16
- Mean Absolute Error (MAE): 36.89
- R-squared (R2) Score: 0.33

Model Comparison

- The Linear Regression model outperformed the other models with the lowest MSE and arelativ ely high R2 score, indicating a better fit to the data.

- Polynomial Regression, despite being more flexible due to its higher degree polynomial feature s, resulted in significantly higher errors and a negative R2 score, indicating overfitting.

- Random Forest Regression, although not as accurate as Linear Regression, provided a balanc ed performance with an intermediate R2 score.

- These results suggest that for the given dataset and prediction task, Linear Regression is the most suitable model. Further analysis, feature engineering, or hyperparameter tuning may help improve the model's performance. It's also important to consider the specific goals and require ments of the application when choosing the most appropriate model.

- In summary, the choice of a regression model should be based on the trade-offs between predi ctive accuracy and model complexity. Careful evaluation of multiple models is essential to mak e an informed decision for real-world applications.

Insights:

1. **Model Performance:** Among the three models evaluated, Linear Regression performed the bes t in terms of predicting "Loan Amount." It achieved the lowest Mean Squared Error (MSE) and a rel atively high R-squared (R2) score, indicating a better fit to the data.

2. **Polynomial Regression:** The Polynomial Regression model, despite its flexibility in capturing c omplex relationships, exhibited overfitting issues. The high degree polynomial features led to a su bstantial increase in errors and a negative R2 score. This suggests that more complex models are not always better, and careful feature engineering or regularization may be needed to improve the model's performance.

3. **Random Forest Regression**: The Random Forest Regression model provided a balanced perf ormance with intermediate errors and an R2 score. While it did not outperform Linear Regression, it offers a trade-off between accuracy and model complexity. It may be suitable for datasets with m ore complex relationships.

4. **Data Preprocessin:** Data preprocessing steps, including handling missing values, label encodi ng, and feature scaling, were crucial in preparing the dataset for modeling. These steps improved t he model's ability to make accurate predictions.

References

- **Dataset -** https://drive.google.com/drive/folders/17AqAclRnF-1J5opMAh-hQm1dpb6BNqf4?usp=drive_link
- **Hackers Realm -** https://www.hackersrealm.net/post/loan-prediction-analysis-using-python
- **Github**
- **ChatGPT**