Table of Contents

Introduction

In the ever-evolving financial landscape, there has been a growing demand from lending institutions for more sophisticated and advanced tools to determine the ideal loan amounts to offer prospective borrowers. To address this challenge, we have developed a precise loan sanction prediction model that can efficiently assess the loan amount to be granted to prospective borrowers and employ data-driven insights to aid lending institutions in making well-informed lending decisions.

This report is centered on predicting the Loan Sanction Amount in USD through the use of an Ordinary Least Squares (OLS) regression model. Our primary objective is to craft the most accurate model capable of estimating loan sanction amounts, a pivotal element for assessing loan applicants' creditworthiness that is predicted based on a comprehensive analysis of a dataset that was collected containing various features.

The dataset at our disposal consists of a multitude of features, and after meticulous feature selection and thorough analysis, we have identified nine key features that significantly impact the prediction of loan amounts. These features exhibit strong correlations with the target variable, which is the loan sanction amounts, thus making them indispensable for our predictive model. Here is an overview of the selected features, along with brief descriptions explaining their respective roles within the dataset:

1. Loan Amount Request (USD): This feature represents the amount requested by loan applicants.
2. m_Dependents: This flag indicates whether the Dependents feature had missing values, which were imputed by taking the mode of the available data.
3. m_Credit Score: This flag indicates missing values in the Credit Score feature, which were imputed by taking the mean of the available data.
4. imp_Credit Score: This feature contains imputed data, representing the mean values of the Credit Score feature for the instances with missing values.
5. m_Property Age: This flag indicates missing values in the Property Age feature, which were imputed by taking the median of the available data.
6. Income Stability_Low: This binned categorical feature represents loan applicants with relatively low-income sources. It categorizes data from the Income Stability feature, specifically focusing on "Low" income values.
7. loanAmountRequestBin_(200000, 300000]: This binned feature represents loan applicants whose requested loan amounts fall within the range of $200,000 (inclusive) to $300,000 (exclusive), referencing the data from the Loan Amount Request feature, specifically focusing on the ranged values.

8. creditScoreBin_(700, 800]: This binned feature represents loan applicants with credit scores ranging from 700 (inclusive) to 800 (exclusive), referencing the data from the Credit Score, specifically focusing on the ranged values.
9. creditScoreBin_(800, 900]: Similar to the previous feature, this bin specifies the range from 800 (inclusive) to 900 (exclusive).
10. Co-ApplicantAdjusted: This feature indicates whether the loan applicant is applying with a co-applicant. We adjusted this feature by addressing negative values, which were considered data entry errors, and clipping them to 0 to enhance the model's data robustness.

Our rigorous testing revealed that these features possess significant potential to influence loan sanction amount predictions, with each feature playing a unique role in enhancing the model's performance. As previously mentioned, we addressed missing values through imputation, binned some features into specific categories to simplify the modelling process, and adjusted certain features to handle outliers containing extreme or irrational data. These actions collectively contribute to the robustness of our predictive model.

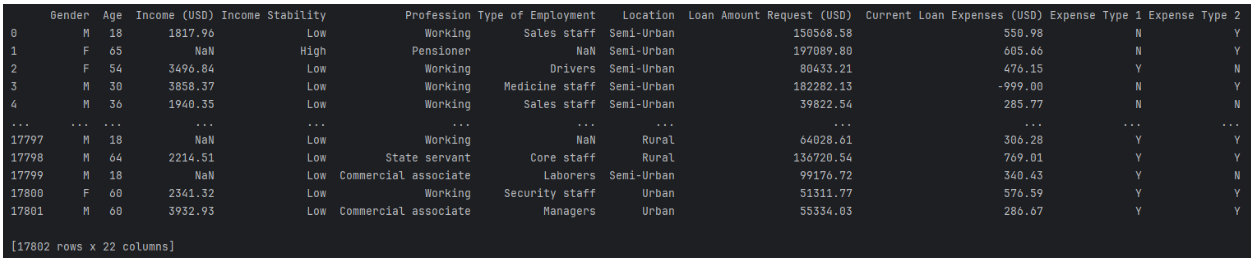
In the subsequent sections of this report, we will delve into a comprehensive evaluation of the model, offering valuable insights and discussing the practical implications for loan assessment.

Exploratory Data Analysis

In this section, we will provide an overview of the dataset while focusing on the key variables that significantly impact our predictive model. We aim to present a concise yet informative summary of the data while ensuring that we prioritize relevant variables and statistics.

Summary of the Data

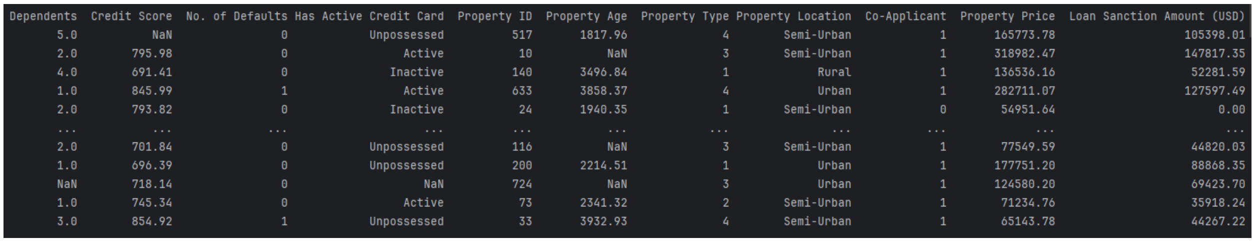
Figure 1 and Figure 2 below provides a comprehensive view of the dataset, displaying all 22 features and showcasing the vastness of our data, comprising a total of 17,802 rows. This extensive dataset is the foundation of our predictive model.



```
      Gender  Age  Income (USD) Income Stability            Profession Type of Employment    Location  Loan Amount Request (USD)  Current Loan Expenses (USD) Expense Type 1 Expense Type 2
0          M   18       1817.96              Low               Working        Sales staff  Semi-Urban                  150568.58                       550.98              N              Y
1          F   65           NaN             High             Pensioner                NaN  Semi-Urban                  197089.80                       605.66              N              Y
2          F   54       3496.84              Low               Working            Drivers  Semi-Urban                   80433.21                       476.15              Y              N
3          M   30       3858.37              Low               Working     Medicine staff  Semi-Urban                  182282.13                      -999.00              N              Y
4          M   36       1940.35              Low               Working        Sales staff  Semi-Urban                   39822.54                       285.77              N              N
...      ...  ...           ...              ...                   ...                ...         ...                        ...                          ...            ...            ...
17797      M   18           NaN              Low               Working                NaN       Rural                   64028.61                       306.28              Y              Y
17798      M   64       2214.51              Low         State servant         Core staff       Rural                  136720.54                       769.01              Y              Y
17799      M   18           NaN              Low  Commercial associate           Laborers  Semi-Urban                   99176.72                       340.43              Y              N
17800      F   60       2341.32              Low               Working     Security staff       Urban                   51311.77                       576.59              Y              Y
17801      M   60       3932.93              Low  Commercial associate           Managers       Urban                   55334.03                       286.67              Y              Y

[17802 rows x 22 columns]
```

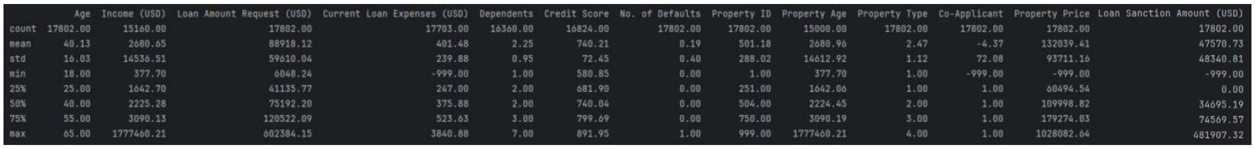
Figure 1 — The first half of the original dataset features and values



```
Dependents  Credit Score  No. of Defaults Has Active Credit Card  Property ID  Property Age  Property Type Property Location  Co-Applicant  Property Price  Loan Sanction Amount (USD)
       5.0           NaN                0            Unpossessed          517       1817.96              4        Semi-Urban             1       165773.78                   105398.01
       2.0        795.98                0                 Active           10           NaN              3        Semi-Urban             1       318982.47                   147817.35
       4.0        691.41                0               Inactive          140       3496.84              1             Rural             1       136536.16                    52281.59
       1.0        845.99                1                 Active          633       3858.37              4             Urban             1       282711.07                   127597.49
       2.0        793.82                0               Inactive           24       1940.35              1        Semi-Urban             0        54951.64                        0.00
       ...           ...              ...                    ...          ...           ...            ...               ...           ...             ...                         ...
       2.0        701.84                0            Unpossessed          116           NaN              3        Semi-Urban             1        77549.59                    44820.03
       1.0        696.39                0            Unpossessed          200       2214.51              1             Urban             1       177751.20                    88868.35
       NaN        718.14                0                    NaN          724           NaN              3             Urban             1       124580.20                    69423.70
       1.0        745.34                0                 Active           73       2341.32              2        Semi-Urban             1        71234.76                    35918.24
       3.0        854.92                1            Unpossessed           33       3932.93              4        Semi-Urban             1        65143.78                    44267.22
```

Figure 2 — The second half of the original dataset features and values

Upon initial inspection, we identify noteworthy aspects. First, the presence of NaN values, specifically in the Property Age column, indicates missing data, which requires further attention. Additionally, some data entries raise red flags and possible outliers, such as a seemingly unreasonable value of -999 in the Current Loan Expenses (USD) column. To confirm and address these issues, we can turn to the statistical summary of each feature, as presented in Figure 3 below. The "Count" section in this summary reveals the total number of data entries for each feature. Any count less than the total number of rows, which is 17,802, signifies the presence of missing values, necessitating a strategy for handling these missing entries. Moreover, by examining the minimum and maximum values in comparison to the mean value for each feature, we can identify extreme values that warrant further investigation and potential data preprocessing.



```
            Age  Income (USD)  Loan Amount Request (USD)  Current Loan Expenses (USD)  Dependents  Credit Score  No. of Defaults  Property ID  Property Age  Property Type  Co-Applicant  Property Price  Loan Sanction Amount (USD)
count  17802.00      15160.00                   17802.00                     17703.00    16360.00      16824.00         17802.00     17802.00      15000.00       17802.00      17802.00        17802.00                    17802.00
mean      40.13       2680.65                   88918.12                       401.48        2.25        740.21             0.19       501.18       2680.96           2.47         -4.37       132039.41                    47570.73
std       16.03      14536.51                   59610.04                       239.88        0.95         72.45             0.40       288.02      14612.92           1.12         72.08        93711.16                    48340.81
min       18.00        377.70                    6048.24                      -999.00        1.00        580.85             0.00         1.00        377.70           1.00       -999.00         -999.00                     -999.00
25%       25.00       1642.70                   41135.77                       247.00        2.00        681.90             0.00       251.00       1642.06           1.00          1.00        60494.54                        0.00
50%       40.00       2225.28                   75192.20                       375.88        2.00        740.04             0.00       504.00       2224.45           2.00          1.00       109998.82                    34695.19
75%       55.00       3090.13                  120522.09                       523.63        3.00        799.69             0.00       750.00       3090.19           3.00          1.00       179274.03                    74569.57
max       65.00    1777460.21                  602384.15                      3840.88        7.00        891.95             1.00       999.00    1777460.21           4.00          1.00      1028082.64                   481907.32
```

Figure 3 — Statistical summary for each feature in the dataset

Furthermore, Figure 4 below illustrates the distributions of the variables, offering visual insights into the data's characteristics. Notably, Credit Score appears to follow a normal-like distribution, while Loan Amount Request, Dependents, Property Price, and Loan Sanction Amount exhibit right-skewed distributions. Property Age and Income are clustered, likely due to data values represented as decimals, making identical values highly improbable. The Co-Applicant feature presents an outlier with an extreme value close to -1000, with the majority of the data concentrated around 0 and 1, resulting in an unusual distribution.

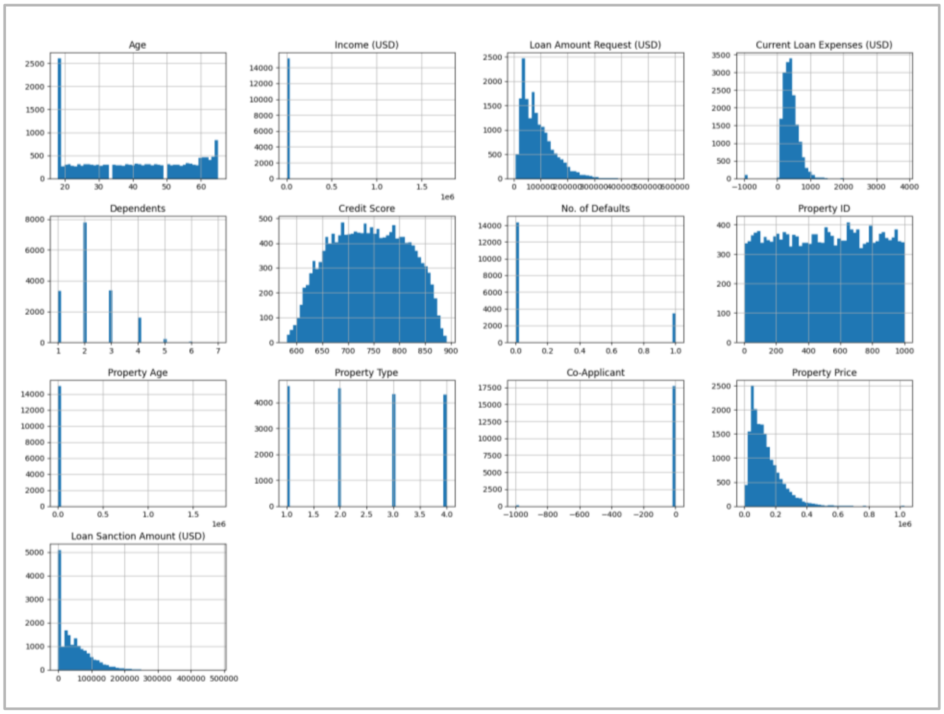


Figure 4 —Distribution for each feature in the dataset

This initial data exploration provides a solid foundation for our subsequent analysis, allowing us to focus our efforts on addressing missing values and outliers while acknowledging the specific distribution patterns of our key variables.

Correlation

In this section, we delve into the interplay between our key variables and the Loan Sanction Amount, aiming to identify potential predictor variables that impact our model positively or negatively.

To start, we investigate the correlations between each pair of feature columns, enabling us to discern the variables with the most significant relationships with the Loan Sanction Amount, our target variable. The heatmap shown in Figure 5 below provides a visual representation of these correlations. Notably, Loan Amount Request, Property Price, and Current Loan Expenses exhibit high positive correlation scores with Loan Sanction Amount. These correlations imply that as these features increase, the Loan Sanction Amount also tends to rise. Conversely, Dependents and Property Age display low to negligible correlations, while the Co-Applicant feature negatively correlates with Loan Sanction Amount most likely due to the outliers that exhibit negative values, indicating a decrease in Loan Sanction Amount when this feature is present.

A different perspective on these correlations can be gleaned from the scatter matrix displayed in Figure 6 below. This matrix visually represents the bivariate relationships between feature combinations and provides insights into the spread and distribution of data points. Once again, the Property Price, Current Loan Expenses, and Loan Amount Request features emerge as positively correlated with Loan Sanction Amount. On the other hand, Dependents, Credit Score, Property Age, and Co-Applicant features do not show any correlation with our target variable.

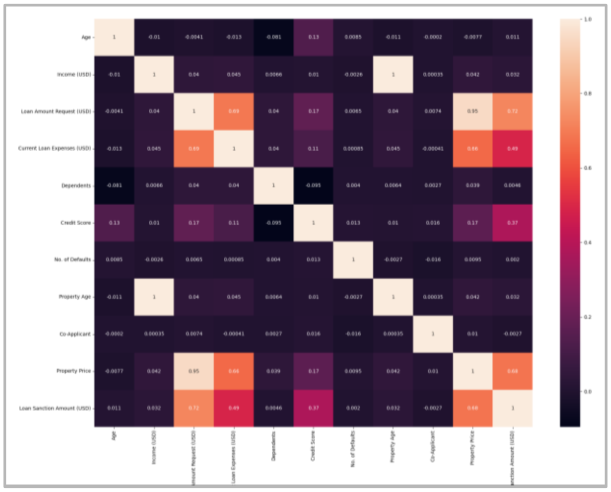


Figure 5 — Correlation heatmap for each feature

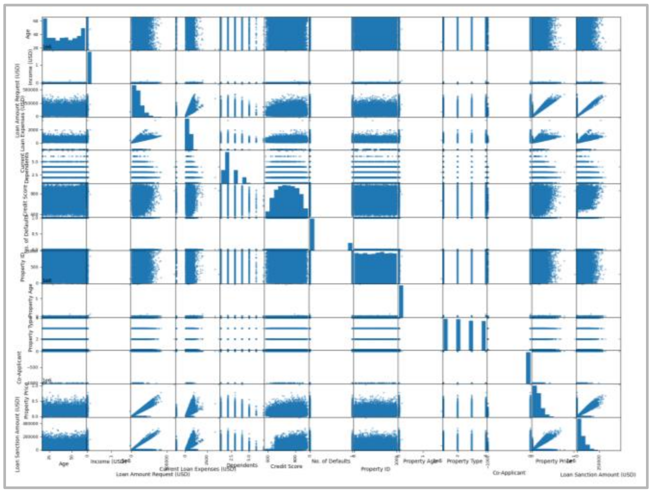


Figure 6 — Scatter matrix for each feature

In summary, our examination highlights that Loan Amount Request, Property Price, and Current Loan Expenses are influential variables that positively impact Loan Sanction Amount. These features show a strong potential to serve as predictor variables in our model. Conversely, the remaining variables exhibit minimal to no correlation with Loan Sanction Amount, suggesting that they may not significantly influence our predictions.

Model Analysis Breakdown

In this section, we provide a detailed analysis of the models developed for predicting Loan Sanction Amounts, with a focus on model comparisons and selection.

Model Comparisons and Selection

Having gained insights into the correlations between variables and our target variable, we constructed three distinct multiple regression models. In Table 1 below, we present a comprehensive statistical comparison of these models, highlighting the best-selected features, along with the average Root Mean Squared Error (RMSE), R^2 (Coefficient of Determination), Adjusted R^2, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) values.

Model	A	B	C
Variables	Age, Loan Amount Request (USD), m_Dependents, m_Credit Score, imp_Credit Score, m_Property Age	Loan Amount Request (USD), m_Dependents, m_Credit Score, imp_Credit Score, m_Property Age, Income Stability_High, Income Stability_Low, Type of Employment_Managers, loanAmountRequestBin_(200000, 300000]	Loan Amount Request (USD), m_Dependents, m_Credit Score, imp_Credit Score, m_Property Age, Income Stability_Low, loanAmountRequestBin_(200000, 300000], creditScoreBin_(700, 800], creditScoreBin_(800, 900], Co-ApplicantAdjusted
#Vars	6	9	10
Average R^2	0.5916	0.6	0.6468
Average R^2Adj	0.591	0.5998	0.6466
Average AIC	334,900	334,620	332,840
Average BIC	334,960	334,720	332,920
Average RMSE	30922.14383617876	30602.039837390847	28739.067277410155

Table 1 — Comparison table for three multiple regression models

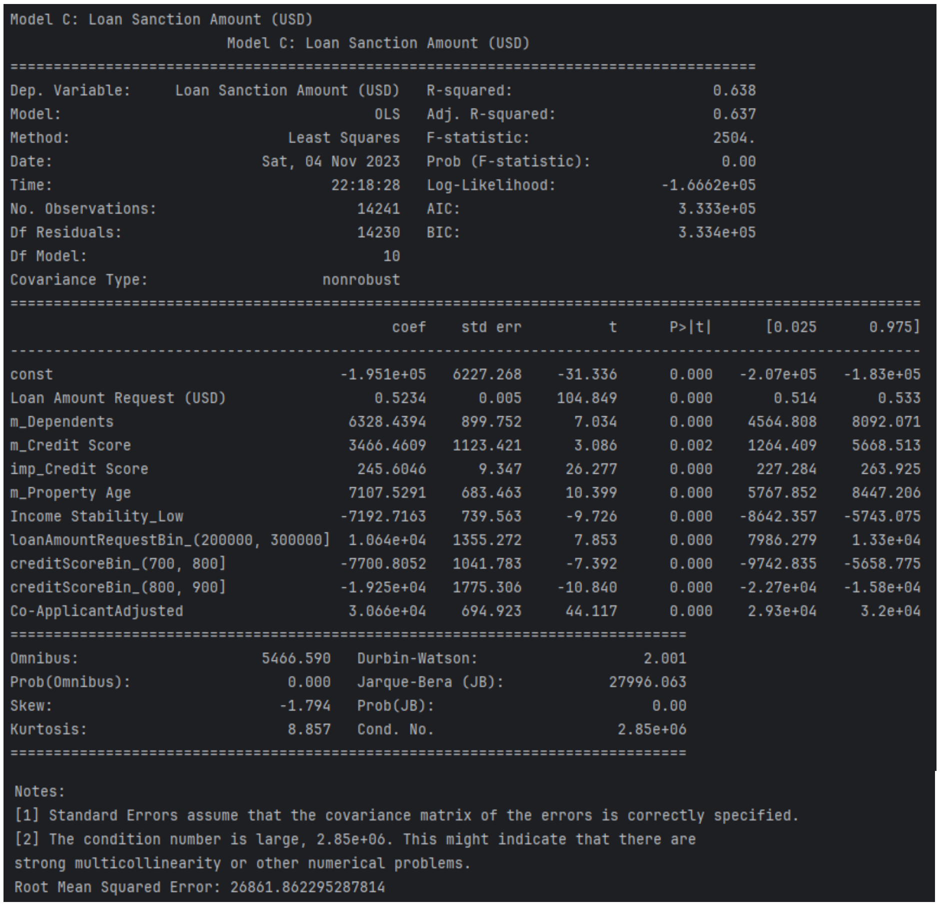
Model C stands out as the superior choice among the three models for several compelling reasons. Examining the statistics, we begin with the Coefficient of Determination (R^2), which quantifies the proportion of variance in the target variable explained by the predictor variables. Model C boasts the highest R^2 value, indicating that it is a strong fit for the data and that our chosen predictor variables effectively predict the target variable. Furthermore, Model C also achieves the highest adjusted R^2, which takes into account the number of predictor variables while penalizing excessive variables, providing a more accurate measure of the model's goodness of fit. Additionally, when we consider the square root of the R^2 value, measuring the correlation between the model and the data, Model C once again emerges as the leader with a robust correlation coefficient of 0.804, signifying its strong predictive power.

Turning to the AIC (Akaike Information Criterion), which evaluates how well a model fits the data while penalizing larger numbers of predictor variables, Model C claims the lowest value, a desirable trait. The Bayesian Information Criterion (BIC), which shares similarities with AIC but employs a distinct penalty for the number of parameters, also positions Model C with the lowest value, an indication of its superior performance.

In conclusion, Model C consistently outperforms the other two models in all measured categories, solidifying its position as the optimal choice for predicting Loan Sanction Amount. Its superior R^2 and adjusted R^2 values, impressive correlation, and favorable AIC and BIC scores collectively affirm Model C as the most favorable and effective model for our predictive task.

Model Evaluation

In this section, we focus on the evaluation of Model C, which was selected as the optimal choice in the previous section. By examining a sample model summary from one of the test runs, we can gain further insights into its performance. The summary, as displayed in Figure 7 below, provides various significant statistics.



```
Model C: Loan Sanction Amount (USD)
                         Model C: Loan Sanction Amount (USD)
======================================================================================
Dep. Variable:     Loan Sanction Amount (USD)   R-squared:                       0.638
Model:                                    OLS   Adj. R-squared:                  0.637
Method:                         Least Squares   F-statistic:                     2504.
Date:                        Sat, 04 Nov 2023   Prob (F-statistic):               0.00
Time:                                22:18:28   Log-Likelihood:            -1.6662e+05
No. Observations:                       14241   AIC:                         3.333e+05
Df Residuals:                           14230   BIC:                         3.334e+05
Df Model:                                  10
Covariance Type:                    nonrobust
=========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                 -1.951e+05   6227.268    -31.336      0.000   -2.07e+05   -1.83e+05
Loan Amount Request (USD)                 0.5234      0.005    104.849      0.000       0.514       0.533
m_Dependents                           6328.4394    899.752      7.034      0.000    4564.808    8092.071
m_Credit Score                         3466.4609   1123.421      3.086      0.002    1264.409    5668.513
imp_Credit Score                        245.6046      9.347     26.277      0.000     227.284     263.925
m_Property Age                         7107.5291    683.463     10.399      0.000    5767.852    8447.206
Income Stability_Low                  -7192.7163    739.563     -9.726      0.000   -8642.357   -5743.075
loanAmountRequestBin_(200000, 300000]  1.064e+04   1355.272      7.853      0.000    7986.279    1.33e+04
creditScoreBin_(700, 800]             -7700.8052   1041.783     -7.392      0.000   -9742.835   -5658.775
creditScoreBin_(800, 900]             -1.925e+04   1775.306    -10.840      0.000   -2.27e+04   -1.58e+04
Co-ApplicantAdjusted                   3.066e+04    694.923     44.117      0.000    2.93e+04     3.2e+04
==============================================================================
Omnibus:                     5466.590   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            27996.063
Skew:                          -1.794   Prob(JB):                         0.00
Kurtosis:                       8.857   Cond. No.                     2.85e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.85e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
Root Mean Squared Error: 26861.862295287814
```

Figure 7 — Statistical model summary of Model C

One critical metric to consider is the Root Mean Squared Error (RMSE), which in this particular iteration stands impressively low at 26861.86. This low RMSE value indicates the model's remarkable accuracy, highlighting how closely our predictions align with the actual values. Such precision is vital to ensuring the reliability of our model's forecasts.

The F-statistic is another valuable parameter to examine. It assesses whether the chosen predictor variables, that is, our selected features, are statistically significant in explaining the variance in the target variable. In this instance, the F-statistic is statistically significant, as evidenced by a probability value of 0. This means that we can confidently reject the null hypothesis, which is essentially an assessment of the validity of our predictive model or the significance of our predictor variables. The rejection of the null hypothesis signifies that our model and predictor variables have a statistically significant effect on the outcomes of our predictions, enhancing the model's credibility.

The log-likelihood, on the other hand, describes the cumulative sum of probability for all sample values under the best-fitting normal distribution and provides insights into the goodness of fit of our model. For Model C, the log-likelihood score is -166,620, indicating a relatively poor fit for the data. While this suggests room for improvement in terms of fit, it doesn't diminish the model's overall effectiveness in predicting Loan Sanction Amount.

Furthermore, the Durbin-Watson test examines whether the error residuals are independent of each other. Model C attains a value of 2, which is considered ideal, indicating that the error residuals are indeed independent.

In the decision-making process for selecting the best features for our models, we conducted a thorough evaluation of the coefficient p-values. These p-values, derived from hypothesis tests for the model coefficients, indicate the statistical significance of each feature. A variable is deemed statistically significant if its p-value is less than 0.05. In Figure 7 above, every selected feature displays a p-value less than 0.05, affirming their significance as predictors.

In summary, Model C incorporates the most predictor variables among the three models we constructed. Despite this, the model showcases robust correlations and relationships between variables, demonstrating their strong positive impact on our predictions. This underlines the effectiveness of Model C in accurately predicting Loan Sanction Amounts.

Multicollinearity

In this section, we address the issue of multicollinearity within our model. Multicollinearity occurs when one independent variable in a regression model exhibits a linear correlation with another independent variable. This phenomenon is evident in Figure 7 above, showcasing strong correlations among the m_Credit Score, imp_Credit Score, and Loan Amount Request features.

The consequences of these strong correlations can be observed when any of these correlated features are removed from the model. Notably, this leads to a substantial increase in the RMSE value, significantly degrading the model's overall performance. As a result, it becomes evident that these features are vital for the model's predictive accuracy.

The Conditional Number, which in this case stands at a remarkably high value of 2,850,000, is an indicator of the extent of multicollinearity. Such a high value underscores the strong correlation among the mentioned features. However, it is important to note that in this particular scenario, given the nature of the lending industry, the presence of collinearity may be acceptable and, in fact, reflective of real-world lending conditions.

In conclusion, while the presence of multicollinearity is acknowledged in our model, it is considered acceptable and even potentially reflective of the lending industry's intricacies. These correlated features are essential for achieving the model's desired performance in accurately predicting Loan Sanction Amounts.

Visualizing Actual, Predicted, and Error Residuals

In this section, we employ visual representations to enhance our understanding of the model's performance. Figure 8 below provides a comprehensive visual overview of our predictions and their relationship with actual Loan Sanction Amounts, as well as an analysis of error residuals.

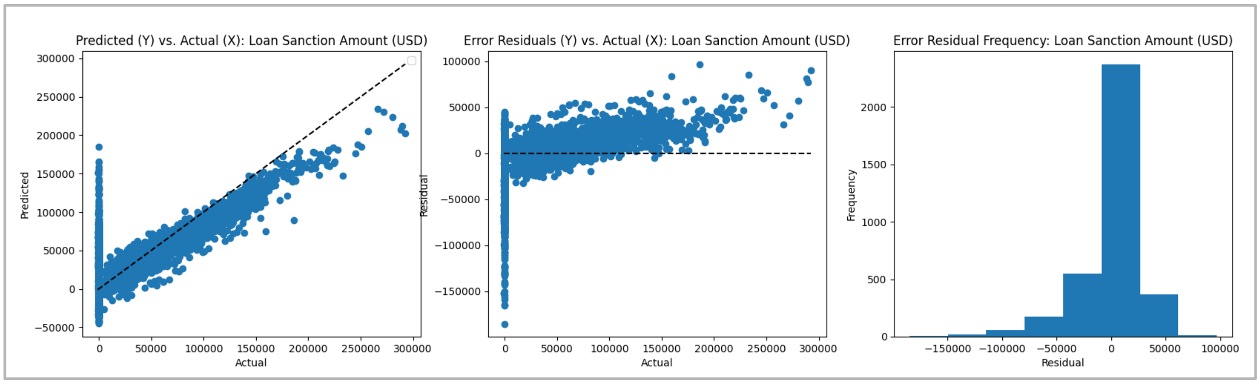


Figure 8 — Plots of Predicted vs. Actual Results, Error Residuals and Error Residual Frequency vs. Actual Results

The first plot in Figure 8 displays the model's predicted values versus the actual results for Loan Sanction Amount. Our predictions align quite closely with the actual values, maintaining accuracy for the majority of the data. However, we observe a slight deviation in our predictions from the actual results, particularly when actual Loan Sanction Amounts surpass the 200,000 mark.

The second plot showcases the error residuals versus the actual results. This plot also exhibits a commendable alignment between the error residuals and actual results for the most part. However, it is worth noting that deviations become more noticeable when actual Loan Sanction Amounts exceed the 175,000 mark. The presence of outliers, represented by values significantly deviating from the predictions, is evident in both of these plots.

The third and final plot is a frequency histogram distribution of the error residuals. This distribution skews to the left, indicating that the majority of error residuals are concentrated on the negative side, with some extreme deviations. This skew suggests that our model is better at underestimating Loan Sanction Amounts.

In summary, our model performs well in predicting Loan Sanction Amount values ranging from 50,000 to 175,000. It maintains high accuracy in most instances, with slight deviations noted for higher Loan Sanction Amounts. The presence of outliers and the leftward skew in the error residuals distribution serve as valuable insights into our model's predictive performance.

Model Equation and Interpretation

In this section, we translate Model C into a comprehensive equation by extracting the coefficient values for each feature from the model summary in Figure 7. The resulting equation takes the form:

$$\begin{aligned}
&y_{\text{Loan Sanction Amount (USD)}} = -195100 + 0.5234 * x_{\text{Loan Amount Request (USD)}} + 6328.4394 * x_{\text{m\_Dependents}} \\
&+ 3466.4609 * x_{\text{m\_Credit Score}} + 245.6046 * x_{\text{imp\_Credit Score}} + 7107.5291 * x_{\text{m\_Property Age}} \\
&- 7192.7163 * x_{\text{Income Stability\_Low}} + 10640 * x_{\text{loanAmountRequestBin\_(200000, 300000]}} \\
&- 7700.8052 * x_{\text{creditScoreBin\_(700, 800]}} - 19250 * x_{\text{creditScoreBin\_(800, 900]}} + 30660 * x_{\text{Co-ApplicantAdjusted}}
\end{aligned}$$

Interpreting this model equation in plain language, we can derive valuable insights into the factors influencing Loan Sanction Amount (USD).

The model starts with a base Loan Sanction Amount of -195,100 USD. Loan Amount Request (in USD) positively impacts the Loan Sanction Amount, meaning that as the requested loan amount increases, the predicted Loan Sanction Amount also rises. The presence of Dependents, Credit Score, and Property Age in the dataset contributes positively to the Loan Sanction Amount. Having a low-stability income source negatively affects the Loan Sanction Amount, reducing the predicted value. If the Credit Score falls within the range of 700 (inclusive) to 800 (exclusive), this results in a negative impact on the Loan Sanction Amount. Conversely, if the Credit Score falls within the range of 800 (inclusive) to 900 (exclusive), it leads to a decrease in the Loan Sanction Amount. A Loan Amount Request within the range of 200,000 to 300,000 USD positively influences the predicted Loan Sanction Amount. Applying for a loan with a co-applicant also positively impacts the Loan Sanction Amount.

In summary, the Loan Sanction Amount in USD is determined by the interplay of these selected features, as derived from the applicant's data. This model equation allows us to gain a deeper understanding of how various factors affect the Loan Sanction Amount, facilitating more informed lending decisions.

Future Considerations

In evaluating our model's performance, it becomes evident that it demonstrates a commendable ability to predict Loan Sanction Amounts, particularly within the range of 50,000 to 175,000 USD. However, several areas offer opportunities for future improvements and refinements.

First and foremost, the availability of more extensive and complete datasets with minimal missing data holds the potential to significantly enhance the predictive capabilities of our model. Addressing missing data would enable more accurate and robust predictions, especially for individuals falling outside the aforementioned Loan Sanction Amount range.

Another area for consideration is the presence of outliers, which may have influenced the model's performance to some extent. Implementing alternative outlier treatment methods or working with datasets free from significant outliers could contribute to a more stable and precise model.

Additionally, the application of data transformation techniques, such as scaling, could yield improved results. Given the substantial amount of continuous data in our dataset, proper scaling techniques may help in achieving better model performance and predictive accuracy.

Furthermore, expanding the dataset to incorporate related properties or additional relevant features would also be advantageous. The inclusion of such variables could play a vital role in further enhancing the model's performance by capturing additional insights into the loan sanctioning process.

In conclusion, the model exhibits promise and competency, but future considerations should focus on data quality, outlier handling, feature engineering, and data transformation to achieve even more accurate and reliable predictions. These enhancements could pave the way for more informed and precise lending decisions.