Case Study - Income Estimation

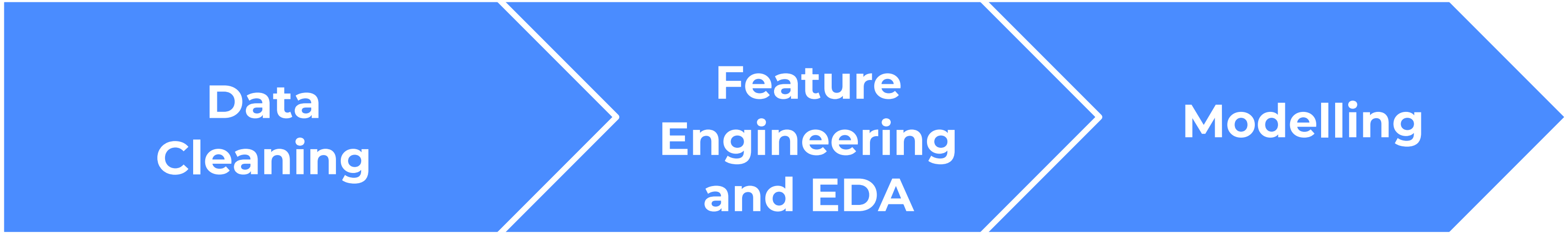
By- Het Upadhyay, MSc. Data Science

Table of Contents

1. Process Flow
2. Problems in the data
3. Missing values and inconsistency
4. EDA
5. Feature Generation
6. Feature Importance
7. Modelling

Building an Income Estimation Model Using Credit History

Process Flow



Data Cleaning

Feature Engineering and EDA

Modelling

- Missing Values imputation
- Consistency Check and handling

- Feature Creation
- Univariate and Bivariate Analysis
- Checking correlation and relationship with target variable
- Feature transformation(log)

- Used Xgboost, Random Forest, Lasso to find feature importance
- Found best parameters using Grid search to achieve RMSE improvement of 40%. R2 - 0.20

01

About the Data

About the Credit Data

Shape of the data

6.4 lakhs rows and 15 columns

Loans per ID

14 loans per ID on average

No. of ID

46k unique ID

Problems in the data

Multiple credit data per ID

Solution - Requires aggregating the Data

Large no. of Null values

Solution - Require correctly imputing and filling missing values

Absence of relevant predictors of income

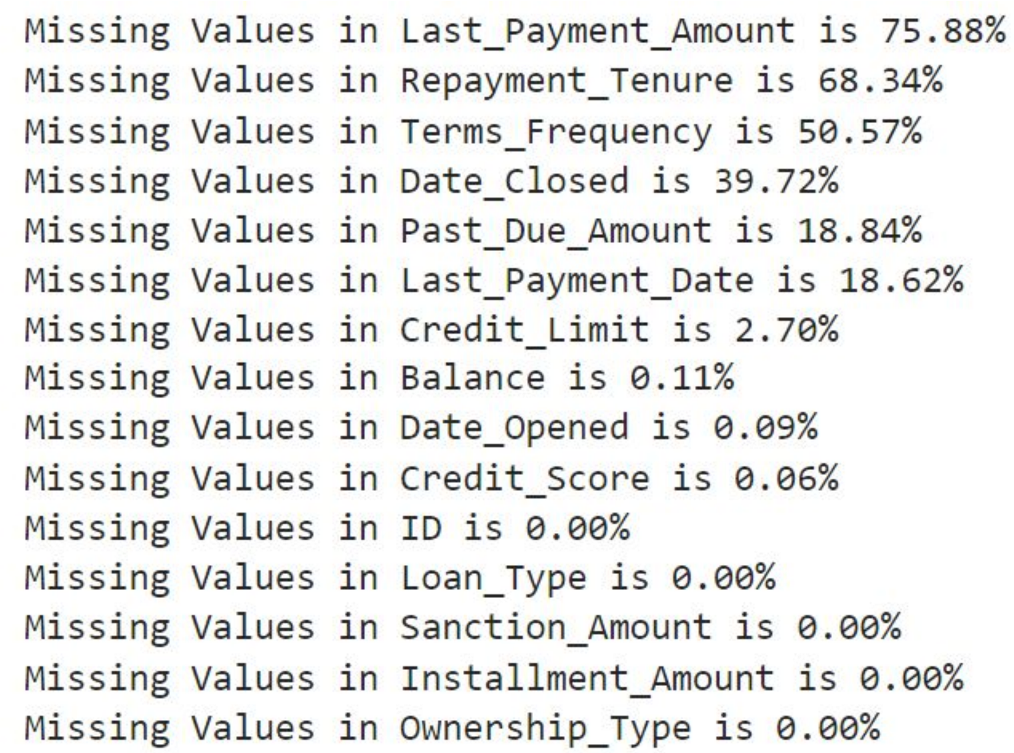
Solution - Feature creation and engineering

Presence of Large Outliers

Solution - Applying appropriate Transformations

Note - We cant directly remove the outliers because this is what provide the variability to the dataset. And it is absolutely necessary for business outcome to equally predict High Income Individuals

Missing data



```
Missing Values in Last_Payment_Amount is 75.88%
Missing Values in Repayment_Tenure is 68.34%
Missing Values in Terms_Frequency is 50.57%
Missing Values in Date_Closed is 39.72%
Missing Values in Past_Due_Amount is 18.84%
Missing Values in Last_Payment_Date is 18.62%
Missing Values in Credit_Limit is 2.70%
Missing Values in Balance is 0.11%
Missing Values in Date_Opened is 0.09%
Missing Values in Credit_Score is 0.06%
Missing Values in ID is 0.00%
Missing Values in Loan_Type is 0.00%
Missing Values in Sanction_Amount is 0.00%
Missing Values in Installment_Amount is 0.00%
Missing Values in Ownership_Type is 0.00%
```

Missing Values Filling

Last Payment Amount

Despite **75% missing** last payment amounts, around **36k unique IDs** have **at least one non-null value**, which can be used for aggregation. Therefore, we **fill the null values with 0.**

Terms Frequency

The Terms Frequency contingency table with Loan Type shows dominant values of TF for each loan type, so we'll fill missing values with the **mode of term frequency for each loan type.**

Date Closed

If the date closed is null, it means the **loan is still active**, so we leave it as is. We will still use this to calculate the **days to close** (using date difference) for **aggregation**

Credit Limit

Credit limits **mainly apply to credit card loans**, and **most null values are from non-credit card loans.** We address this in aggregation

Consistency Check

Date	There are 247 rows where the Date Closed is earlier than the Date Opened, so we drop them as they represent a very small number compared to the 6 lakh total rows
Installment Amount	184 rows where Installment amount was negative, which is not possible. So we drop those rows.
Balance	Around 18k rows where Balance amount > Sanction Amount. This maybe due to accumulated interest charges, late fees etc.

02

Exploratory Data Analysis

Some Findings -

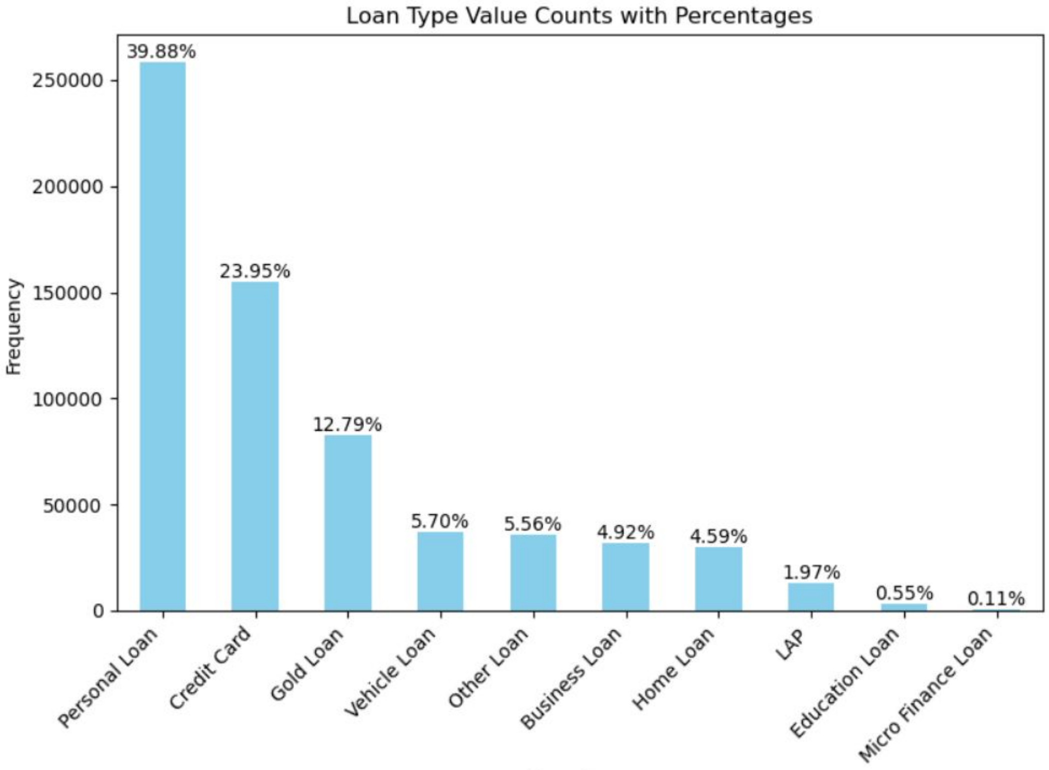
- **Balance** - *75% of values of balance are* ***below 8820***.

- **Sanction Amount** - *It seems that* ***majority of values*** *are around* ***50k to 2 lakh rs*** *(50%tile - 51,000 and 75%tile - 2 lakh)*

- **Dates**- *The date range is from* ***1988 to 2024***

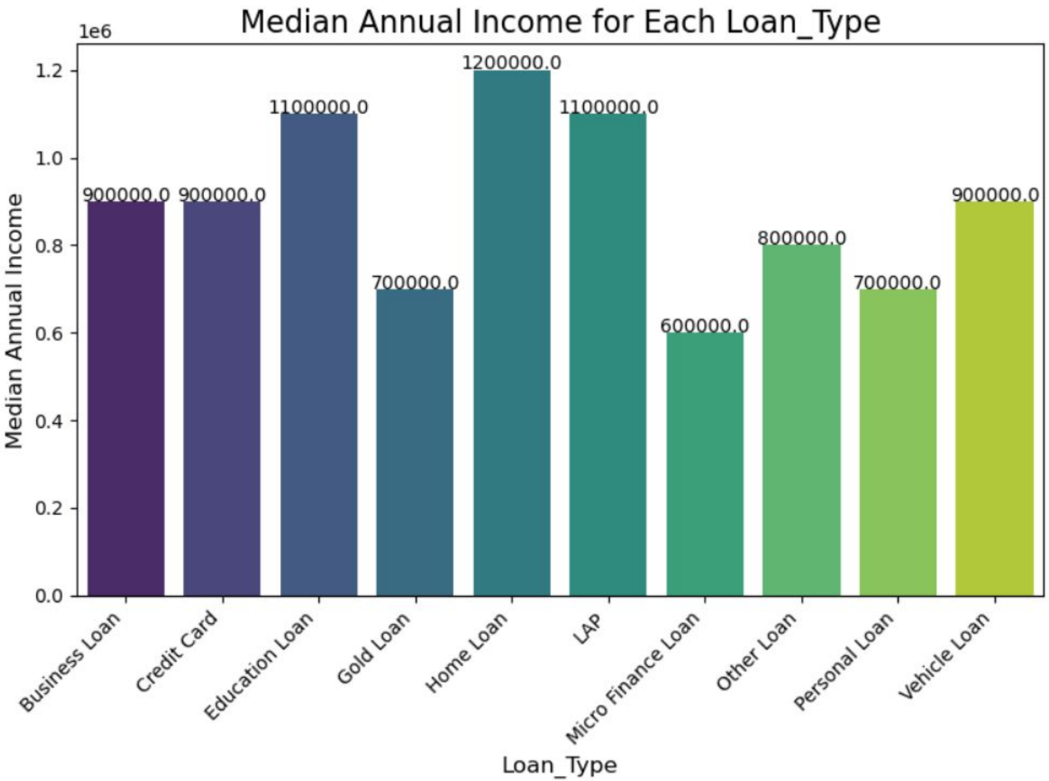
- **Repayement** Tenure- *It seems that majority of loans(75%tile) are* ***under 5 years tenure***

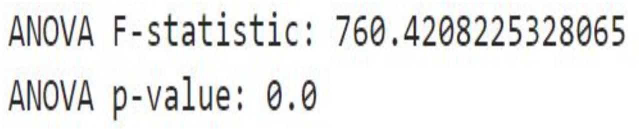
Loan Type Distribution



- Personal loan is the highest taken loan type followed by Credit Card and Gold Loan.
- Education Loan and Microfinance Loan are the lowest taken loan types.

Does loan type have any relationship with Annual Income ?

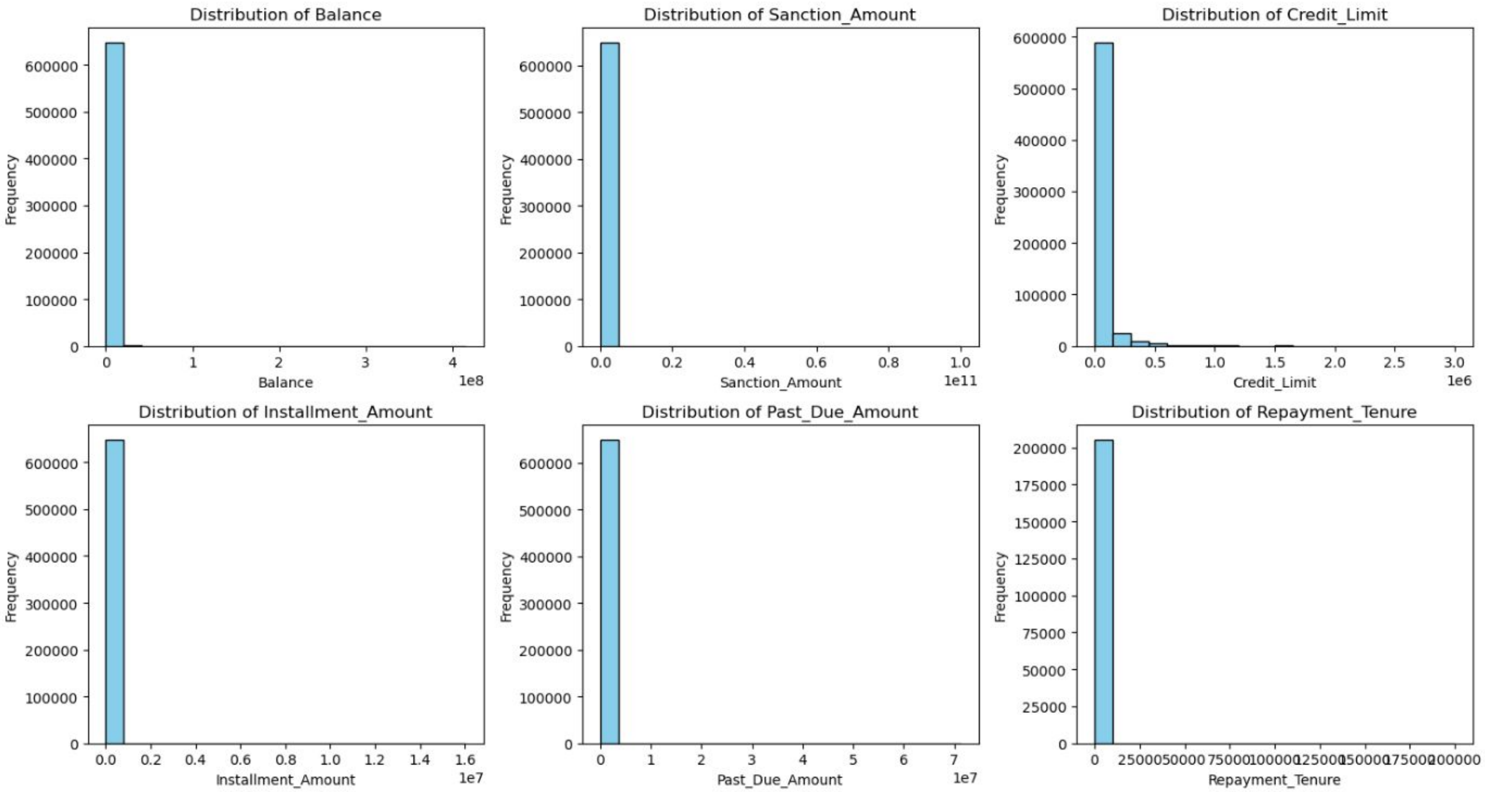




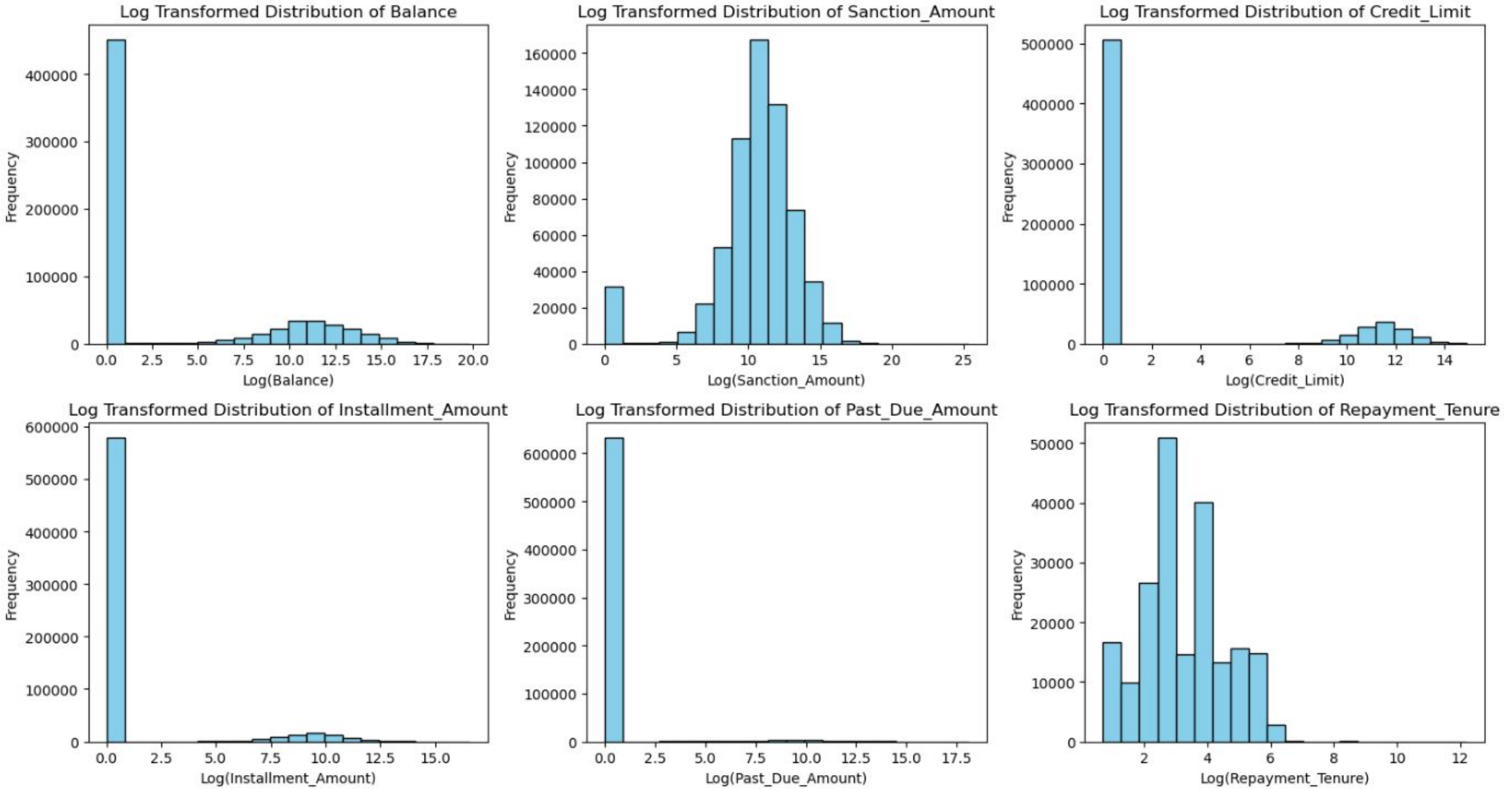
```
ANOVA F-statistic: 760.4208225328065
ANOVA p-value: 0.0
```

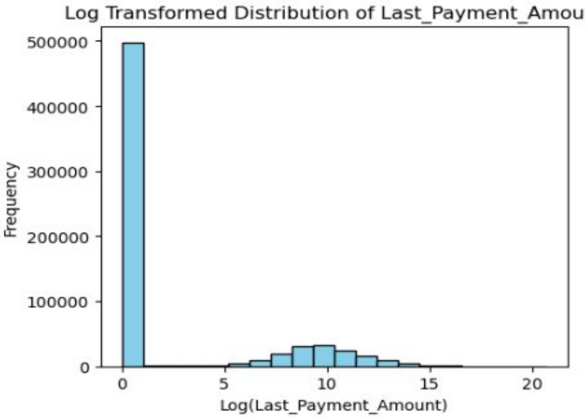
As p-value is near zero, One of the loan type has significant relationship with Annual Income.

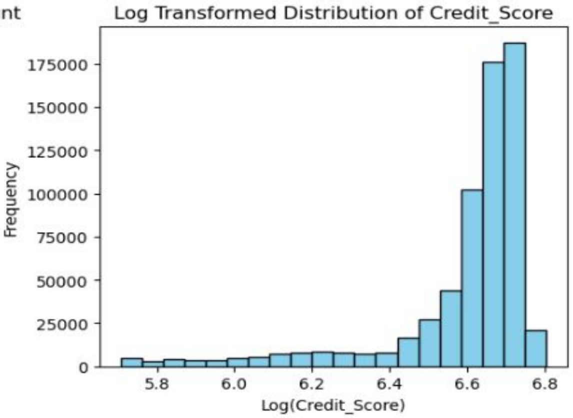
Continuous columns in the dataframe have large outliers

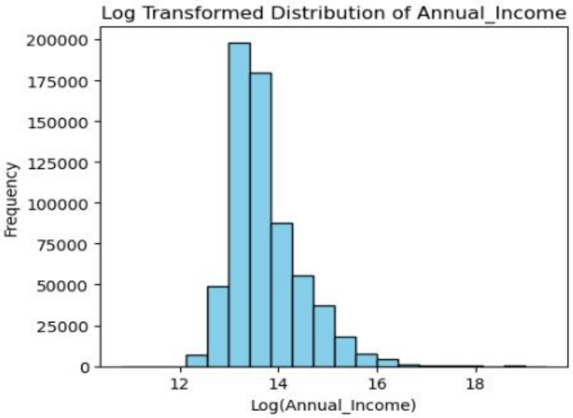


We apply log transformation to reduce the effect of outliers.









Findings:

Balance: A large number of zero values indicate that the **majority of loans in the dataset are inactive**.

Sanction Amount: Appears normally distributed after transformation, but there are also **many zero values, likely due to credit card or other loan types where a sanction is not required.**

Credit Limit: High occurrence of zero values, as **credit limits apply only to specific loan types (credit card).**

Installment Amount: The majority of values are zero.

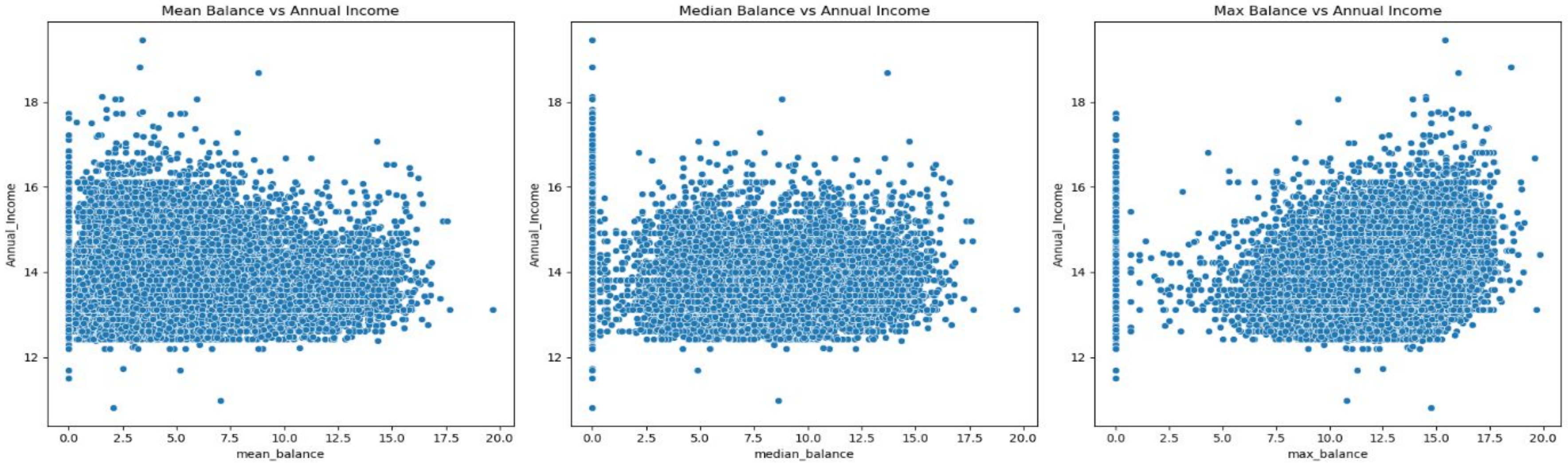
Repayment Tenure: Does not follow a normal distribution and is left-skewed. **Indicating less duration of repayment tenure.**

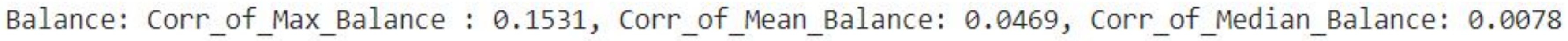
Credit Score: Right-skewed distribution. Indicating **sample with high credit score in the data**.

.

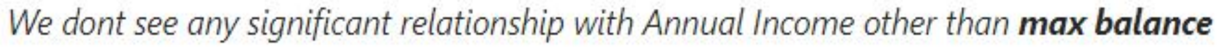
Annual Income: Nearly follows a normal distribution after transformation.

Do Balance have any relationship with Annual Income ?





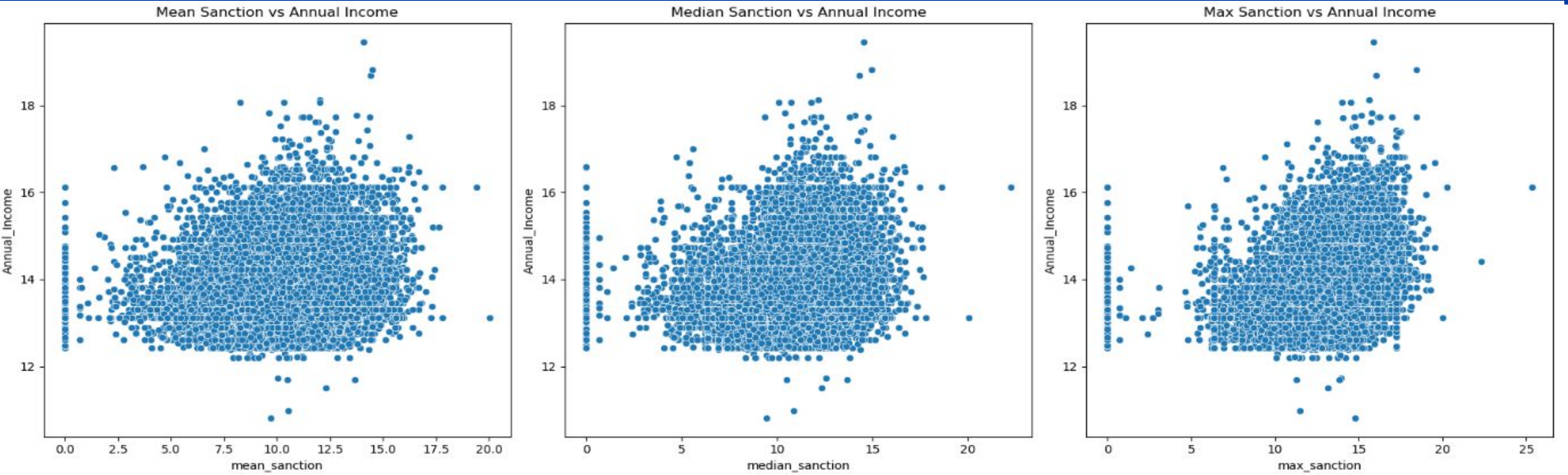
```
Balance: Corr_of_Max_Balance : 0.1531, Corr_of_Mean_Balance: 0.0469, Corr_of_Median_Balance: 0.0078
```



We dont see any significant relationship with Annual Income other than ***max balance***

Max balance(per id) is slightly correlated with Annual Income

Do Sanction Amount have any relationship with Annual Income ?

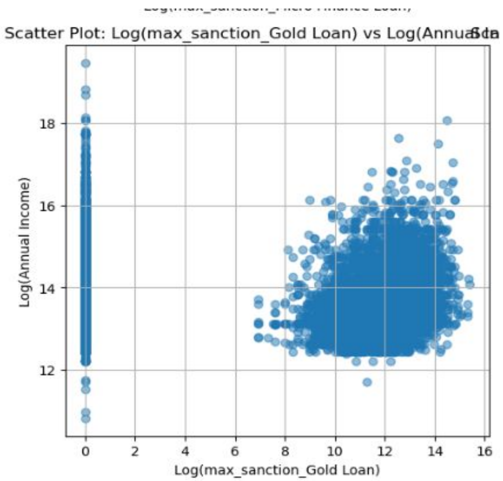


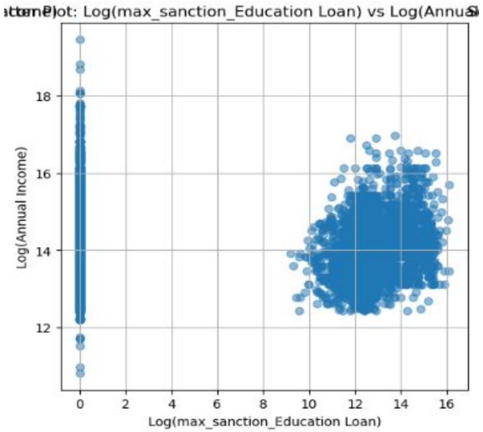
```
Sanction_amount: Corr_of_Max_Sanction_Amount : 0.0193, Corr_of_Mean_Sanction_Amount: 0.0190, Corr_of_Median_Sanction_Amount: 0.0207
```

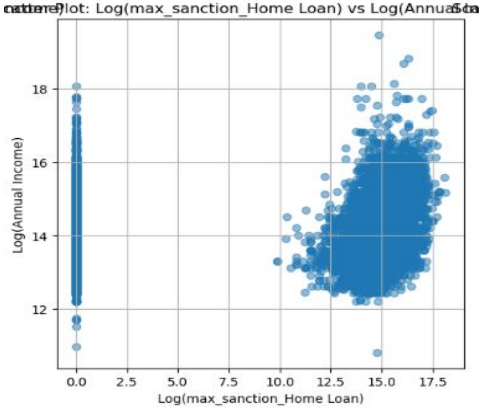
Sanction Amount does not have relationship with annual_Income

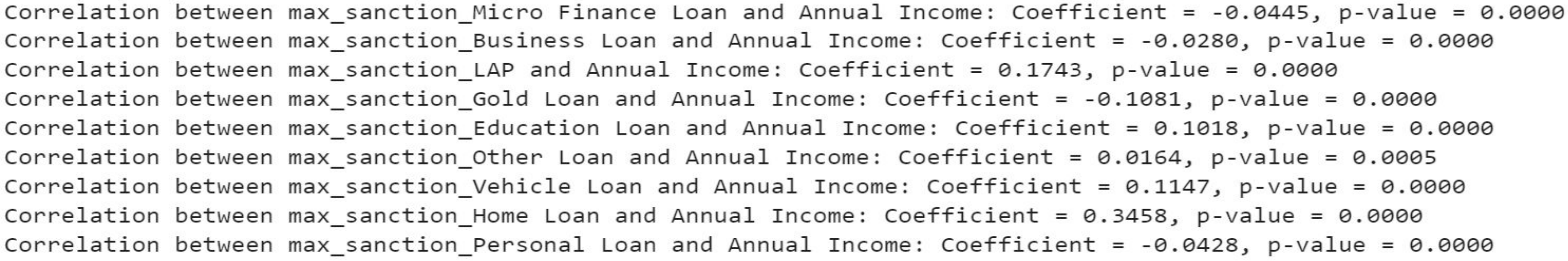
Do aggregate sanction amount by loan type can have some relationship with Annual Income?

Do Sanction Amount have any relationship with Annual Income ?









```
Correlation between max_sanction_Micro Finance Loan and Annual Income: Coefficient = -0.0445, p-value = 0.0000
Correlation between max_sanction_Business Loan and Annual Income: Coefficient = -0.0280, p-value = 0.0000
Correlation between max_sanction_LAP and Annual Income: Coefficient = 0.1743, p-value = 0.0000
Correlation between max_sanction_Gold Loan and Annual Income: Coefficient = -0.1081, p-value = 0.0000
Correlation between max_sanction_Education Loan and Annual Income: Coefficient = 0.1018, p-value = 0.0000
Correlation between max_sanction_Other Loan and Annual Income: Coefficient = 0.0164, p-value = 0.0005
Correlation between max_sanction_Vehicle Loan and Annual Income: Coefficient = 0.1147, p-value = 0.0000
Correlation between max_sanction_Home Loan and Annual Income: Coefficient = 0.3458, p-value = 0.0000
Correlation between max_sanction_Personal Loan and Annual Income: Coefficient = -0.0428, p-value = 0.0000
```

Max Sanction Home loan, Max sanction Gold Loan and Max Sanction Educational Loan has slight correlation with Annual Income

Correlation

Credit Limit Aggregation

Max Credit Agg have slight correlation with Annual Income. **Corr_of_Max_Credit_Limit : 0.2180,** Corr_of_Mean_Credit_Limit: 0.1623, Corr_of_Median_Credit_Limit: 0.0732.

Date Diff aggregation

Max Date/Mean Date diff does seem to have a very sligh correlation with annual income.

Installment_Amount

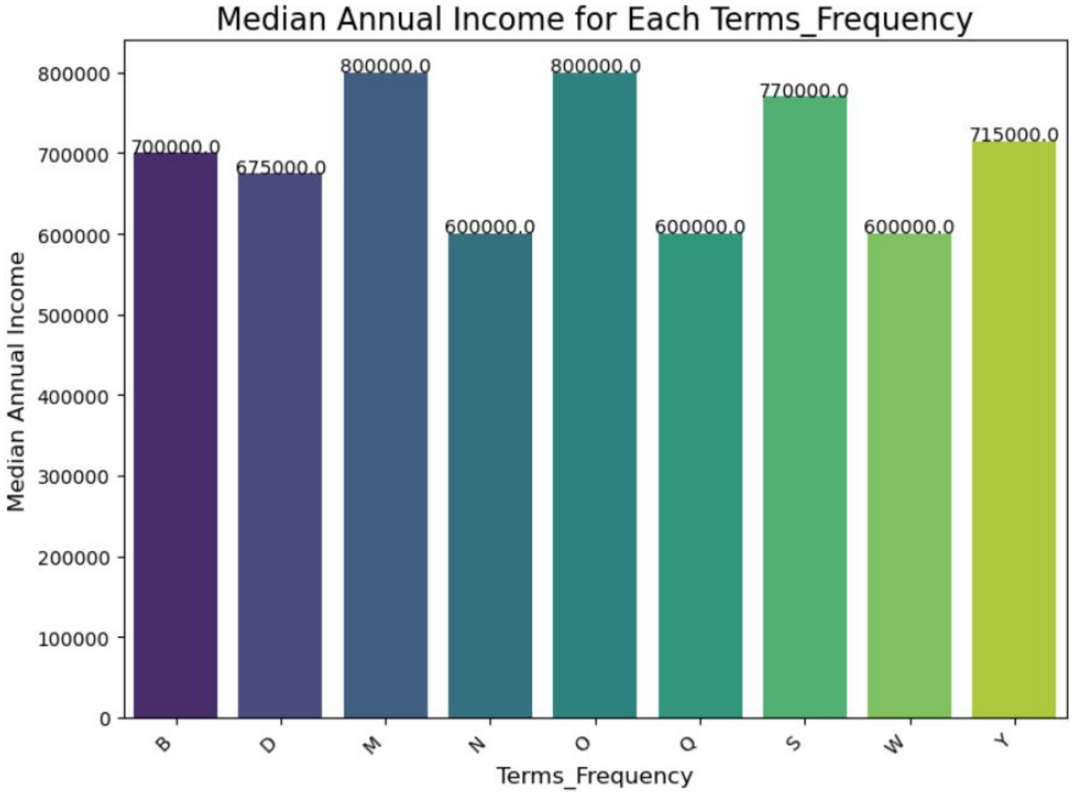
Past_Due_Amount

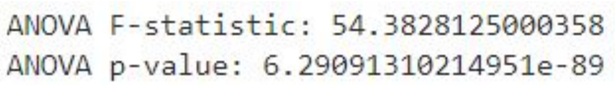
Repayment_Tenure

LastPaymentAmount

These aggregate of these variables do not have any significant correlation with Annual Income

Does Term Frequency have any relationship with Annual Income ?

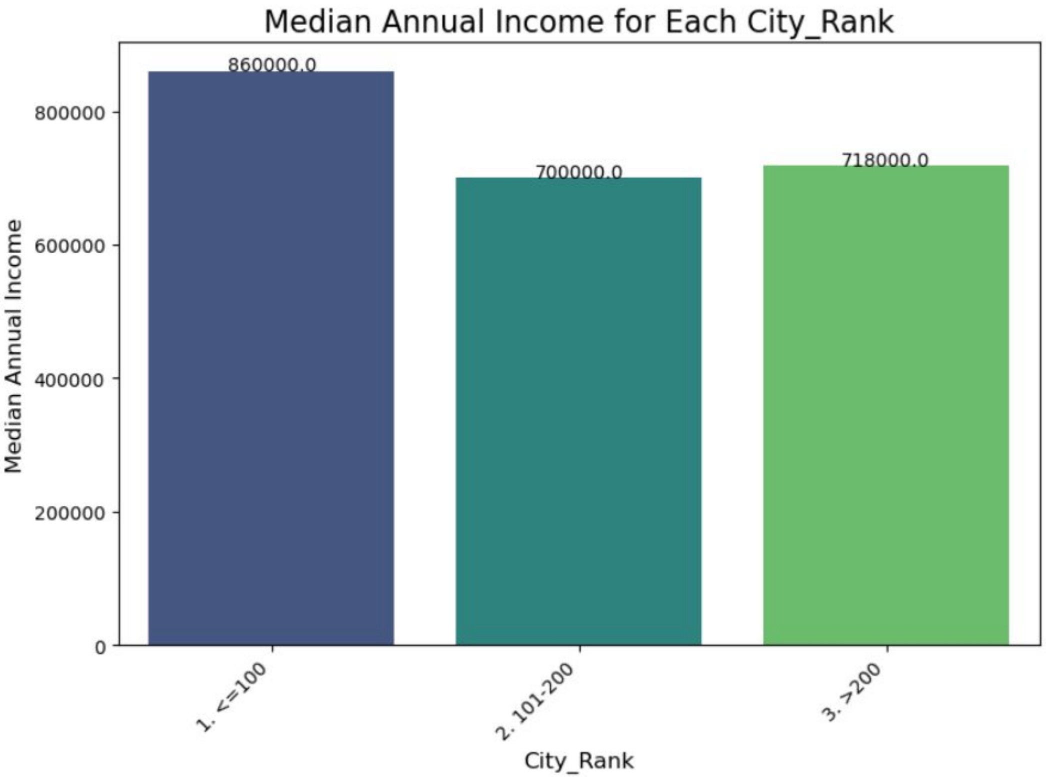


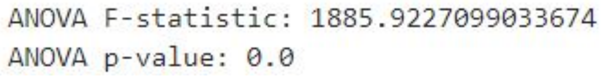


```
ANOVA F-statistic: 54.3828125000358
ANOVA p-value: 6.29091310214951e-89
```

As p-value is near zero, One of the Term Frequency has significant relationship with Annual Income.

Does City Rank have any relationship with Annual Income ?





```
ANOVA F-statistic: 1885.9227099033674
ANOVA p-value: 0.0
```

As p-value is near zero, One of the City Rank has significant relationship with Annual Income.

03

Feature Generation

Feature creation was done on the basis of EDA and intuitively.

Active Loan Feature Creation

Active Loan Flag

- If the loan is an active loan or not

Sum Active Loans

- This feature represents sum of all the active loans per ID

% Active loans

- This feature represents how many loans are active out of all the loans taken by an ID

Loan Type combined with other feature aggregation

Num Loan_type

These variable represents total number of loans a person has taken of that loan_type. For Ex. Num_Home Loan etc.

Max Sanction Loan Type

This feature represent max balance of that loan type per ID. For Ex. Max balance Home Loan

Max Days to Close Loan Type

This feature represents Max days a person takes to close the for that particular loan type.

Max Balance Loan Type

This feature represents Max balance left on loan for that particular loan per ID.

Credit Card Feature Generation

Credit Card Flag

If the person have a credit card loan or not

Credit limit aggregated

If the person has a credit card then the max credit limit of that person is taken but if credit limit is null(even if the person has credit card), then max of sanction amount is taken in place of credit limit.

Note - Credit limit values are null for other loan types but some null values are present even if the person has credit card loan. In this case, we tend to consider max_sanction as sanction amount may be according to the income in some cases.

Installment Amount Feature Generation

Installment Flag

If the person have a non null installment amount value in any row or not.

Max Installment Amount

This feature calculates the max installment amount a person has paid.

Note - Installment Amount intuitively signify a person capacity to pay emi each month. So if we consider max installment amount of a person, we may get an idea of his/her income.

Demographic Data Feature Engineering

City Rank

Label Encoding of City rank.
As higher the city rank better the facility it may indicate a better Annual Income.

Occupation

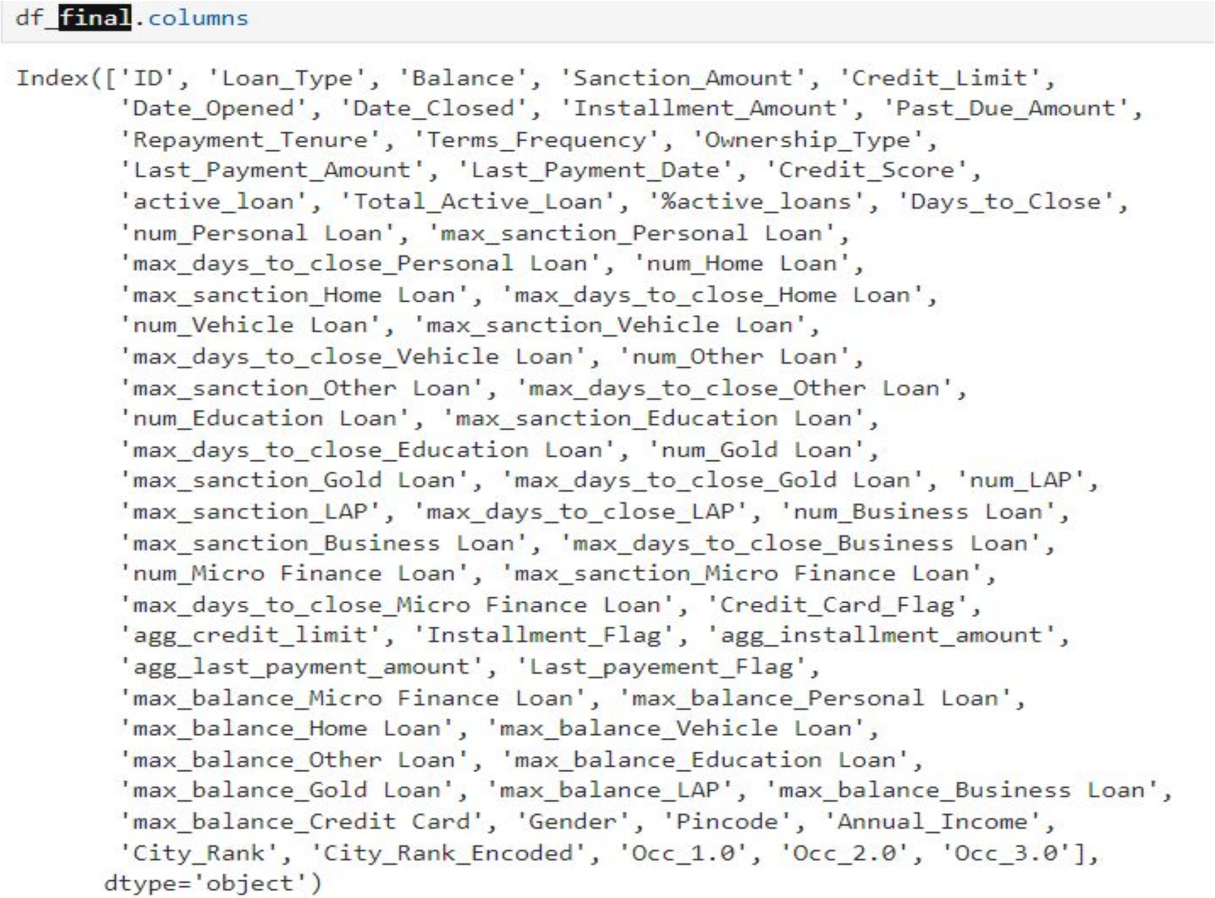
One hot Encoding for Occupation feature.

Note - We performed one hot encoding and not label encoding because , model would take higher values as better profession with is untrue.

Feature Importance

We use Xgboost and lasso feature importance

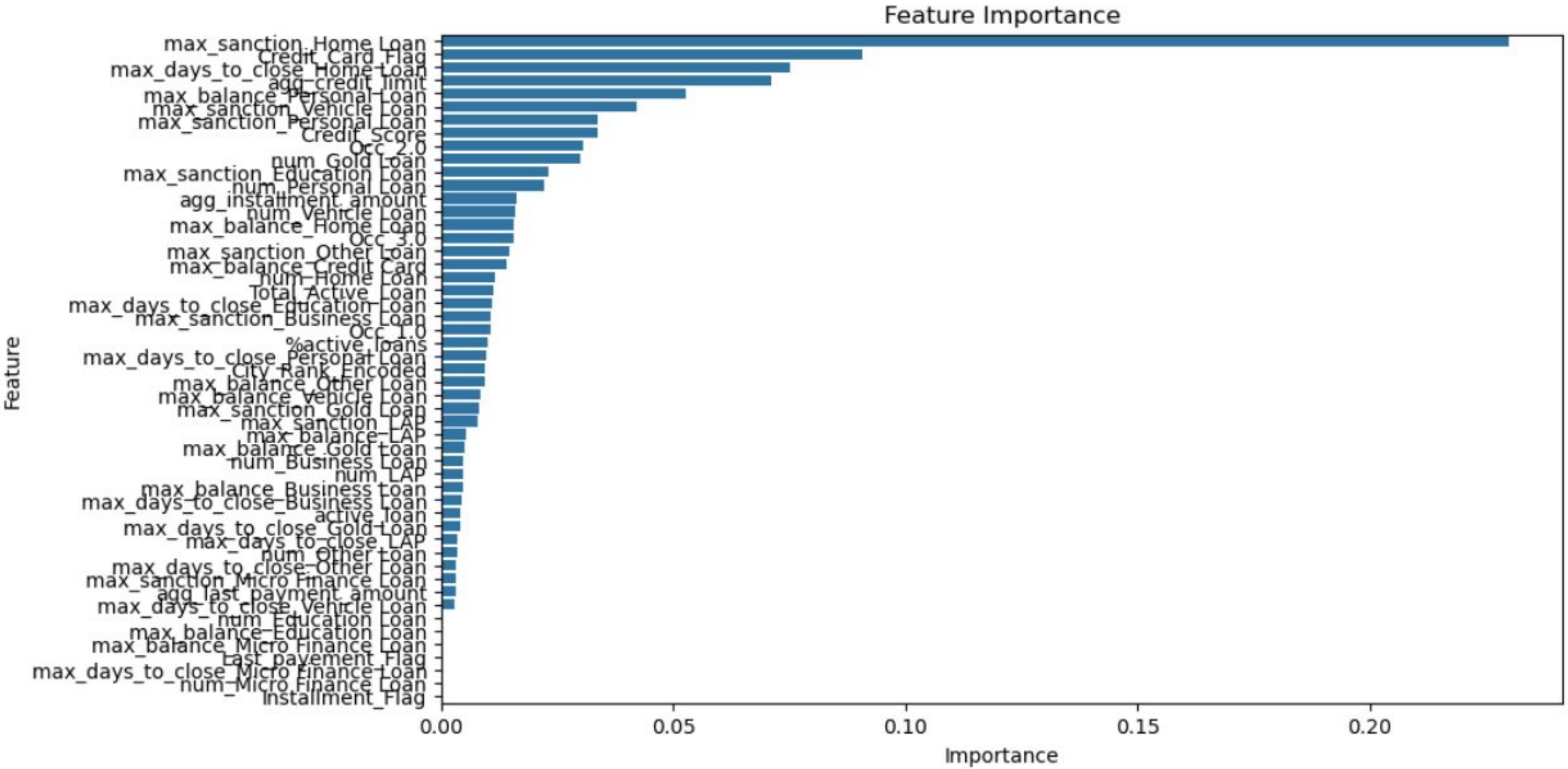
Final Total Columns



```
df_final.columns
```

```
Index(['ID', 'Loan_Type', 'Balance', 'Sanction_Amount', 'Credit_Limit',
       'Date_Opened', 'Date_Closed', 'Installment_Amount', 'Past_Due_Amount',
       'Repayment_Tenure', 'Terms_Frequency', 'Ownership_Type',
       'Last_Payment_Amount', 'Last_Payment_Date', 'Credit_Score',
       'active_loan', 'Total_Active_Loan', '%active_loans', 'Days_to_Close',
       'num_Personal Loan', 'max_sanction_Personal Loan',
       'max_days_to_close_Personal Loan', 'num_Home Loan',
       'max_sanction_Home Loan', 'max_days_to_close_Home Loan',
       'num_Vehicle Loan', 'max_sanction_Vehicle Loan',
       'max_days_to_close_Vehicle Loan', 'num_Other Loan',
       'max_sanction_Other Loan', 'max_days_to_close_Other Loan',
       'num_Education Loan', 'max_sanction_Education Loan',
       'max_days_to_close_Education Loan', 'num_Gold Loan',
       'max_sanction_Gold Loan', 'max_days_to_close_Gold Loan', 'num_LAP',
       'max_sanction_LAP', 'max_days_to_close_LAP', 'num_Business Loan',
       'max_sanction_Business Loan', 'max_days_to_close_Business Loan',
       'num_Micro Finance Loan', 'max_sanction_Micro Finance Loan',
       'max_days_to_close_Micro Finance Loan', 'Credit_Card_Flag',
       'agg_credit_limit', 'Installment_Flag', 'agg_installment_amount',
       'agg_last_payment_amount', 'Last_payement_Flag',
       'max_balance_Micro Finance Loan', 'max_balance_Personal Loan',
       'max_balance_Home Loan', 'max_balance_Vehicle Loan',
       'max_balance_Other Loan', 'max_balance_Education Loan',
       'max_balance_Gold Loan', 'max_balance_LAP', 'max_balance_Business Loan',
       'max_balance_Credit Card', 'Gender', 'Pincode', 'Annual_Income',
       'City_Rank', 'City_Rank_Encoded', 'Occ_1.0', 'Occ_2.0', 'Occ_3.0'],
      dtype='object')
```

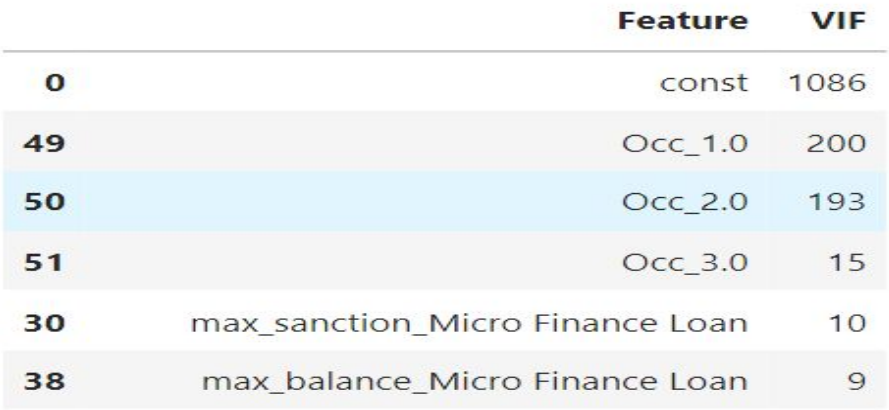
Feature Reduction using Xgboost Feature Importance



Feature Reduction using Xgboost Feature Importance

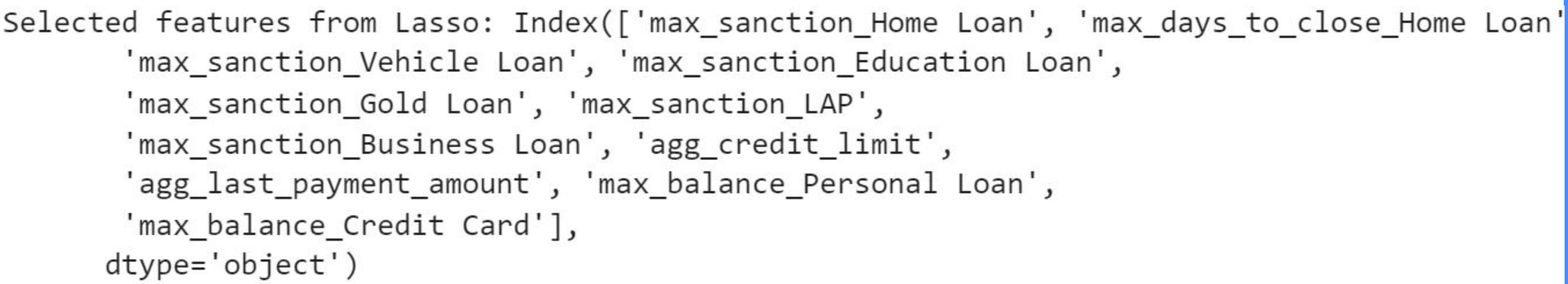
Top 10 features : 'max_sanction_Home Loan', 'Credit_Card_Flag', 'max_days_to_close_Home Loan', 'agg_credit_limit', 'max_balance_Personal Loan', 'max_sanction_Vehicle Loan', 'max_sanction_Personal Loan', 'Credit_Score', 'Occ_2.0', 'num_Gold Loan'

VIF - To check multicolinearity



	Feature	VIF
0	const	1086
49	Occ_1.0	200
50	Occ_2.0	193
51	Occ_3.0	15
30	max_sanction_Micro Finance Loan	10
38	max_balance_Micro Finance Loan	9

Feature Reduction using Lasso



```
Selected features from Lasso: Index(['max_sanction_Home Loan', 'max_days_to_close_Home Loan'
       'max_sanction_Vehicle Loan', 'max_sanction_Education Loan',
       'max_sanction_Gold Loan', 'max_sanction_LAP',
       'max_sanction_Business Loan', 'agg_credit_limit',
       'agg_last_payment_amount', 'max_balance_Personal Loan',
       'max_balance_Credit Card'],
      dtype='object')
```

Note - We see that across all the feature importance - max sanction home loan is the most important feature.

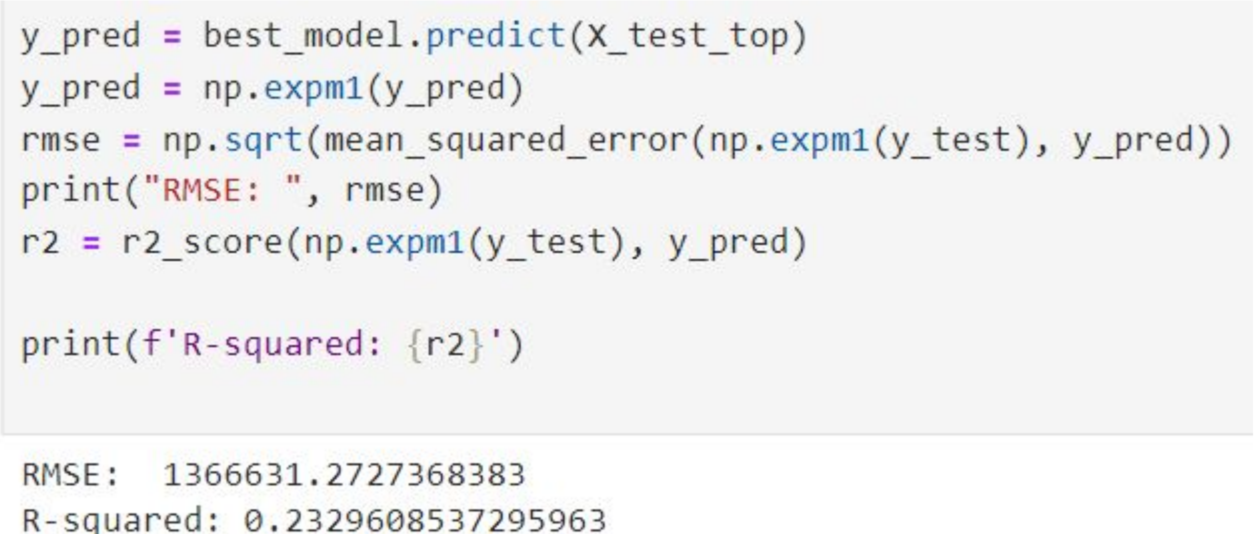
Modelling

We use Xgboost, Random Forest and MLR

XgBoost Hyperparameter Tuning

- Best Hyperparameter after tuning - {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200, 'subsample': 0.8}

- After Hyperparameter tuning using grid search and selected top 20 features we get an RMSE : 13 lakh(80:20 train test split)



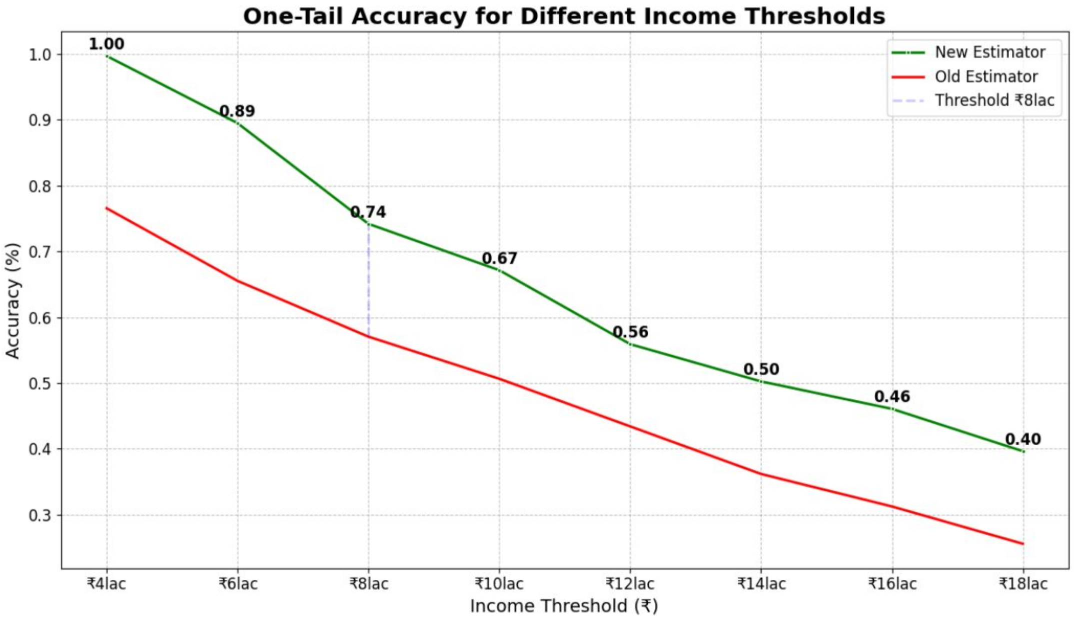
```
y_pred = best_model.predict(X_test_top)
y_pred = np.expm1(y_pred)
rmse = np.sqrt(mean_squared_error(np.expm1(y_test), y_pred))
print("RMSE: ", rmse)
r2 = r2_score(np.expm1(y_test), y_pred)

print(f'R-squared: {r2}')
```

```
RMSE:  1366631.2727368383
R-squared: 0.2329608537295963
```

Note-

- RMSE changes considerably if the split changes. It goes to 19 lakh from 13 lakh.
- R-squared 0.2 means model explain 20% variability in the data

One Tailed Accuracy



If the model estimates a person’s income is > 8 lakhs, it is 74% accurate that his/her true income is also >8 lakhs

This measures how accurate the model estimates a consumer’s income higher than $x.

Thanks

het.upadhyay14@gmail.com
+91 6352723227
LinkedIn - linkedin.com/in/het-upadhyay

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**