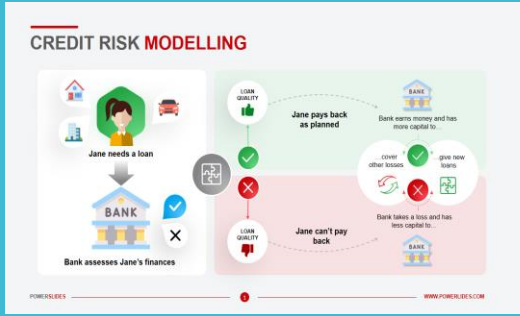


Home Credit Risk Analysis

Kanishka Balwani

Problem Statement

As a business analyst for Home Credit, need to develop a credit scoring mechanism using applicant and bureau data. The goal is to assist Home Credit in making informed decisions on loan approvals based on past applicant behavior and application information. This involves cleaning the data, aggregating trade-level bureau information to the applicant level, creating manual features, and building a classification model to differentiate between approved and rejected applications. Key Questions involved:

- How can trade-level information from credit bureaus be aggregated to the applicant level to capture payment behavior?
- What application or payment behavior factors significantly influence a borrower's behavior on a new loan?
- How can these factors be leveraged to build a model for decision-making?
- Once the model is built, how can its output be translated into strategies and business insights for the bank?

Analysis Strategies

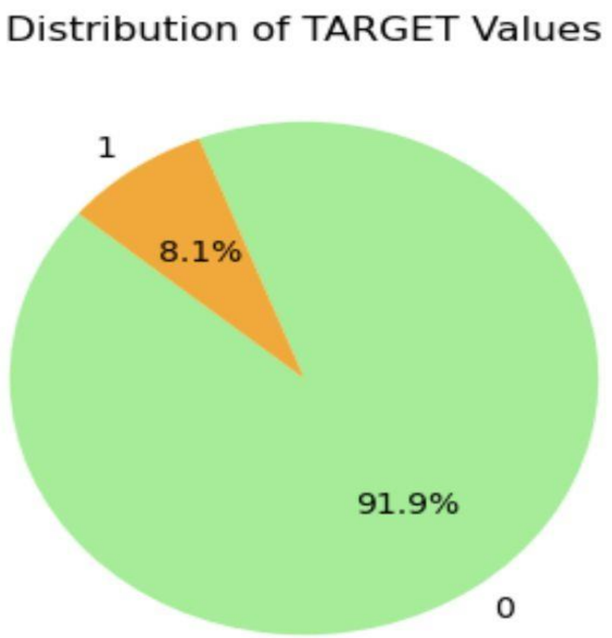
We will follow the following Major Steps in solving this problem :

1. **Data Exploration of Application Data**
2. **Data Quality Checks and Corrections: Null, Duplicate, missing values and imputation or drop**
3. **EDA - Application Data Features (Univariate / Bivariate - Numerical /Categorical data analysis)**
4. **Data Exploration of Bureau Data & Data Quality**
5. **Merging of Application data and Bureau data**
6. **Feature Engineering - Bureau Data (Extracting Manual Features)**
7. **Feature Selection using ANOVA F value**
8. **Data Pre-processing for Machine Learning Model Building (Label encoding, train_test split, StandardScaler)**
9. **Classification Model Building and Model Evaluation alongside :**
   - **LogisticRegression**
   - **Logistic Regression with Hyper Parameter Tuning**
   - **RandomForestClassifier for Binary Classification with Hyper Parameter Tuning**
   - **Light Gradient Descent Boosting with Hyperparameter tuning**
10. **Conclusion / Recommendation as a Business Analysis to Finance Institutions based on application and Bureau trade level information.**

Exploratory data analysis on application data

We performed various data quality checks and ran EDA on application data. One of the important aspects is the Distribution of the TARGET variable

Inference: The data is highly imbalanced, where 91.9% of the 'TARGET' variable is 0 (other cases/able to pay the loan) and 8.1% of the 'TARGET' variable is 1(difficulty in repaying the loan).

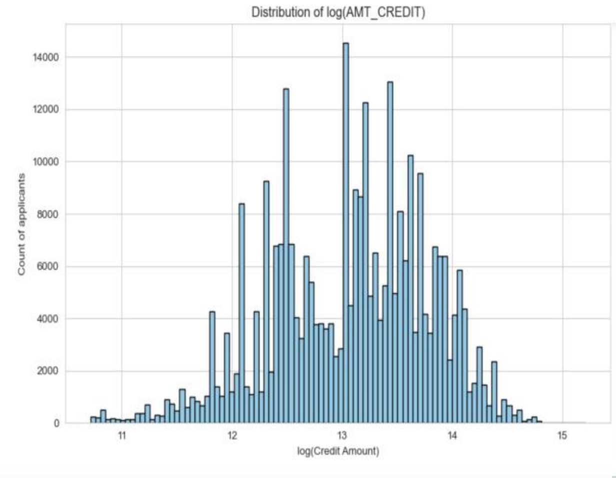


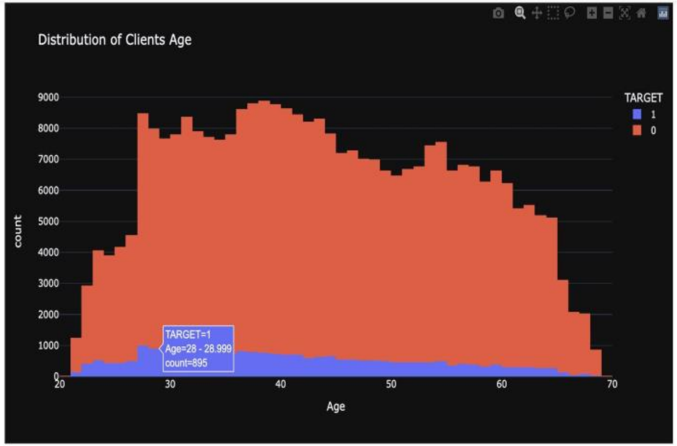
Exploratory data analysis on application data

We performed univariate and bivariate analysis on various numerical and categorical fields Amt Credit and Age Groups seem to be an important factor for borrowers' behaviour.

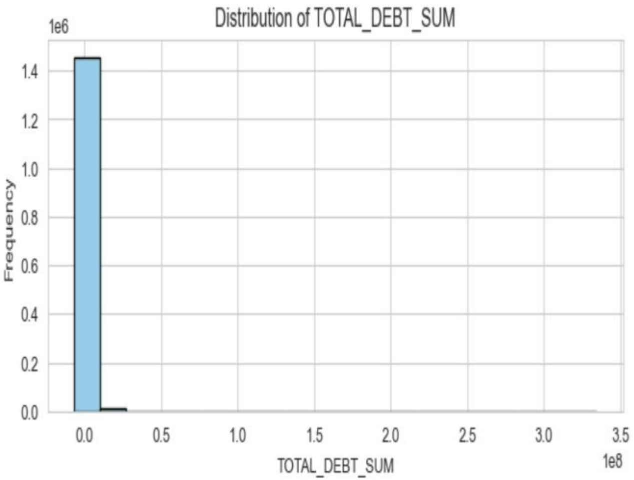
Inferences: People taking credit for large amounts are very likely to repay the loan.

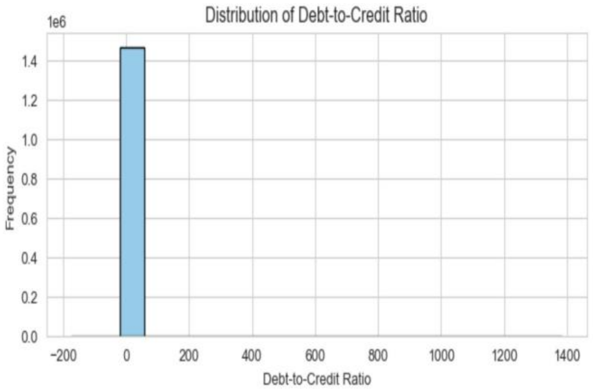
Inference: From the above plot when we compare the two target types, we see that the clients who have difficulty in payment are relatively younger and most of them lie at around 30's.





Feature engineering on bureau data

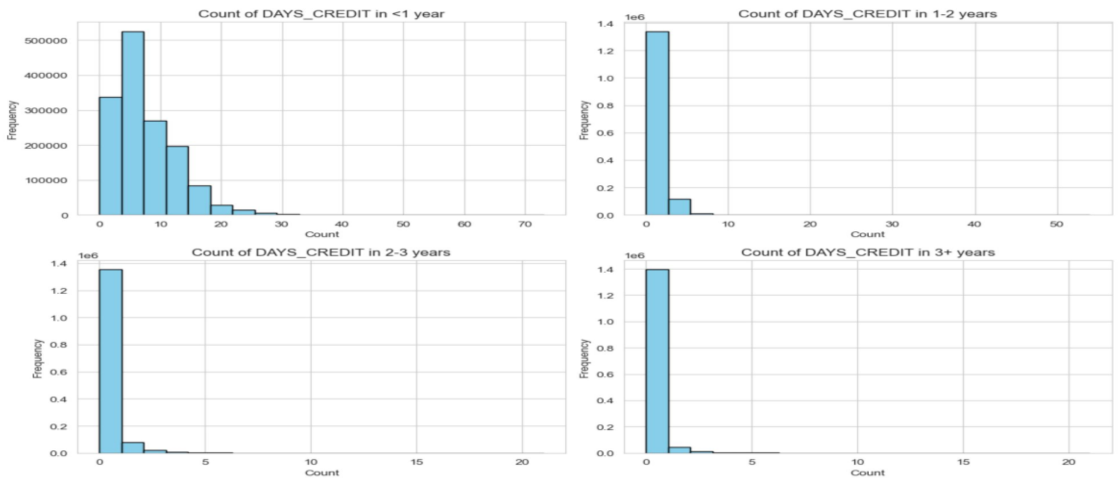




Let's derive the feature name, TOTAL_DEBT_SUM, and DEBT_TO_CREDIT_RATIO

Inferences: High frequency at the lower end of the scale in both the feature indicating, individuals have no or very low outstanding debt and recent credit updated or absence of the data set. It will be immensely useful for better understanding the borrowers' credit behavior.

Feature engineering on bureau data



Let's derive the feature from the days _credit_interval for <1 year, 1-2 years,2-3 years and 3+years

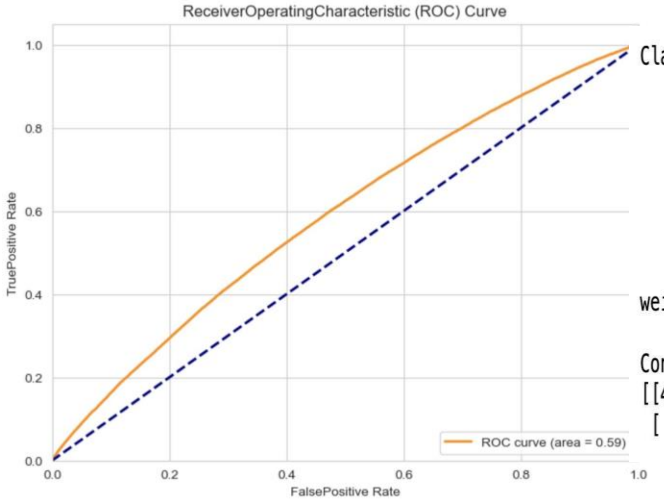
Inferences: A high frequency of credit intervals in <1 year indicates ongoing credit inquiries where, whereas a high frequency at the lower end of other time intervals indicates credit inquiries in those intervals are concentrated more towards the recent past than further back in time.

Feature Selection using ANOVA F value

```
Selected Features : Index(['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'CNT_CHILDREN',
       'AMT_CREDIT', 'AMT_GOODS_PRICE', 'NAME_INCOME_TYPE',
       'NAME_EDUCATION_TYPE', 'NAME_HOUSING_TYPE',
       'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH', 'DAYS_EMPLOYED',
       'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'OCCUPATION_TYPE',
       'CNT_FAM_MEMBERS', 'REGION_RATING_CLIENT',
       'REGION_RATING_CLIENT_W_CITY', 'HOUR_APPR_PROCESS_START',
       'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY',
       'LIVE_CITY_NOT_WORK_CITY', 'ORGANIZATION_TYPE',
       'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE',
       'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE',
       'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_YEAR', 'AGE_Group',
       'ANNUITY_INCOME_PERCENTAGE', 'CREDIT_TERM', 'DAYS_EMPLOYED_PERCENTAGE',
       'CREDIT_ACTIVE', 'DAYS_CREDIT', 'DAYS_CREDIT_ENDDATE',
       'DAYS_ENDDATE_FACT', 'AMT_CREDIT_SUM', 'CREDIT_TYPE',
       'DAYS_CREDIT_UPDATE', 'BUREAU_LOAN_COUNT', 'BUREAU_LOAN_TYPE',
       'DEBT_CREDIT_RATIO', 'DAYS_CREDIT_interval_<1 year',
       'DAYS_CREDIT_interval_1-2 years', 'DAYS_CREDIT_interval_2-3 years',
       'DAYS_CREDIT_interval_3+ years', 'CREDIT_OVERDUE_COUNT',
       'TOTAL_DEBT_SUM', 'TOTAL_OVERDUE_SUM'],
```

Essentials for enhancing the model accuracy and efficiency. It measures the differences between the mean between groups. The high value of ANOVA F value indicates the significant predictors.
Using Feature selection in merged - Application and Bureau data, we selected 50 features
from 76 features.

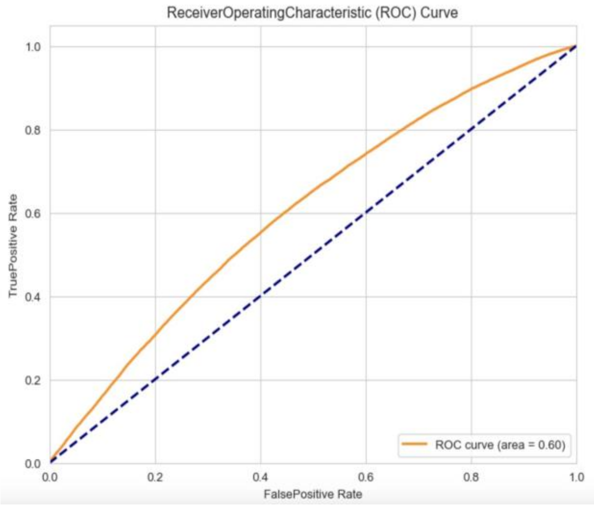
ML Model Building and Model Validation - Logistic Regression




```
Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96    405305
           1       0.25      0.00      0.00     34293

    accuracy                           0.92    439598
   macro avg       0.59      0.50      0.48    439598
weighted avg       0.87      0.92      0.88    439598

Confusion Matrix:
[[405293     12]
 [ 34289      4]]
```

Inference: **Looking at the classification report and confusion matrix, we can see that the model's performance is skewed towards the majority class (class 0). The model is unable to correctly identify most of the positive instances. Hence let's build a Binary classification logistic regression model with a Random Over Sampler.**

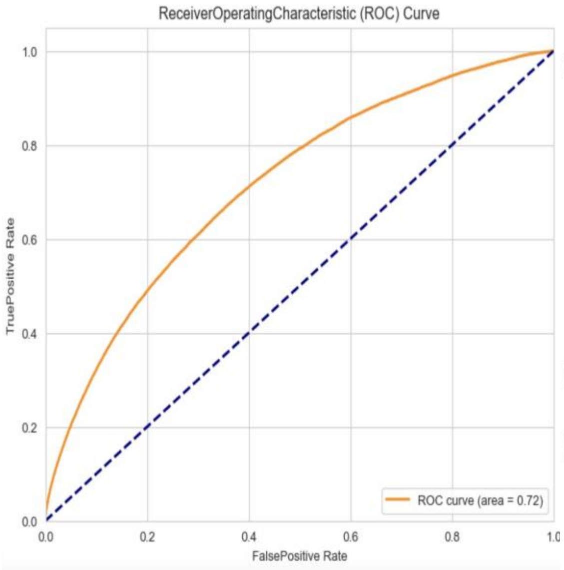
ML Model Building and Model Validation - Logistic Regression with OverSampler




```
assification Report:
              precision    recall  f1-score   support

           0       0.94      0.71      0.81    405305
           1       0.11      0.43      0.18     34293

    accuracy                           0.69    439598
   macro avg       0.52      0.57      0.49    439598
ighted avg       0.87      0.69      0.76    439598

nfusion Matrix:
286466 118839]
 19572  14721]]
```

Inference: ***We can still see that model performance is still poor and Precision for Class-1 is 11%. Though RoC is better i.e. 0.60.*****Hence, let's build a binary classification model with RandomForestClassifier HyperParameters tuning**

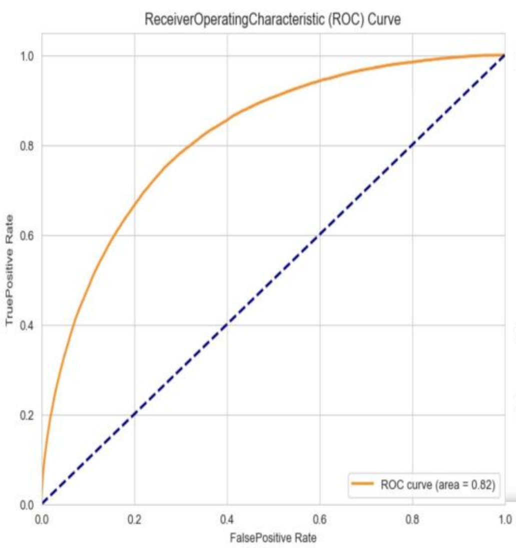
ML Model Building and Model Validation – Random Forest Classifier




```
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.67      0.79    405305
           1       0.14      0.65      0.23     34293

    accuracy                           0.66    439598
   macro avg       0.55      0.66      0.51    439598
weighted avg       0.89      0.66      0.74    439598

Confusion Matrix:
[[269993 135312]
 [ 12116  22177]]
```

Inference: High Precision for Loan Approvals: The model exhibits a high precision rate of 96% for loan approvals, minimizing the risk of granting loans to unqualified applicants. Balanced F1-score for Loan Approvals: With an F1-score of 74%, the model achieves a good balance between precision and recall for approved loans, ensuring reliable decision-making. The ROC curve area of 0.72 demonstrates the model's moderate ability to discriminate between loan approvals and rejections, providing valuable insights for risk assessment.

ML Model Building and Model Validation - Light Gradient Boosting




```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.77      0.86    405305
           1       0.21      0.70      0.32     34293

    accuracy                           0.77    439598
   macro avg       0.59      0.74      0.59    439598
weighted avg       0.91      0.77      0.82    439598

Confusion Matrix:
[[312809  92496]
 [ 10164  24129]]
```

Inference: **Better than the previous model as Light Gradient Boosting model achieves a precision of 97% for loan approvals, minimizing the risk of granting loans to unqualified applicants. With a weighted average F1-score of 82%, the model maintains a good balance between precision and recall, ensuring reliable decision-making in approving loans. The ROC curve area(It's a good metric for imbalanced classes) of 82% demonstrates the model's ability to discriminate between loan approvals and rejections, providing valuable insights for risk assessment. Overall, the model's high precision, balanced F1-score, effective discrimination, and stable predictions make it a valuable tool for the bank in identifying qualified borrowers while minimizing risks associated with loan approvals. However, there is still further scope to optimize the model, especially for Class -1. We can further do Hyperparameter tuning for Gradient Boosting.**

Conclusion and Recommendations

Accuracy: The Final model achieved an accuracy of 76%, indicating its ability to correctly classify loan applications as approved or rejected. With an ROC curve area of 0.82, the model demonstrates a good ability to distinguish between loan approvals and rejections.

Top 10 Factors Influencing Borrower Behavior:

- 'CREDIT_TERM'
- 'DEBT_CREDIT_RATIO'
- 'DAYS_EMPLOYED'
- 'DAYS_BIRTH'
- 'TOTAL_DEBT_SUM'
- 'AMT_GOODS_PRICE'
- 'AMT_CREDIT'
- 'DAYS_CREDIT_interval_1-2 years'
- 'DAYS_CREDIT_interval_3+ years'
- 'DAYS_CREDIT_interval_<1 year'

By aggregating trade-level information to the applicant level, financial institutions can gain insights into borrowers' payment behavior. Features like 'DEBT_CREDIT_RATIO', ' 'DAYS_CREDIT_interval_*' and 'TOTAL_DEBT_SUM' from the credit bureau data contribute significantly to understanding borrowers' historical credit usage and payment patterns.

Factors such as 'CREDIT_TERM', 'AMT_CREDIT', 'AMT_GOODS_PRICE', and DAYS_EMPLOYED significantly influence borrower behavior. These features highlight aspects like loan terms, goods price affordability ratios, Employment duration/days, and debt management practices, providing insights into borrowers' financial health, age group, and risk propensity. By incorporating influential factors identified from borrower behavior analysis, financial institutions can enhance their risk assessment capabilities and make informed decisions regarding loan approvals.