greatlearning
Learning for Life

Project Summary

Batch details	**PGP-DSE FT Apr'23 Gurgaon Batch**
Team members	**Mr. Samarth** **Mr. Nishchay Chauhan** **Mr. Pranshu Agarwal** **Mr. Rahul Rana** **Mr. Vikas Nehra** **Mr. Shivam Singh**
Domain of Project	**Finance**
Proposed project title	**Home Credit Default Risk**
Group Number	**3**
Team Leader	**Mr. Samarth**
Mentor Name	**Mr. Mohit Sahu**

TABLE OF CONTENTS

Project Details

OVERVIEW:

The purpose of this model is to develop and design an effective and efficient model for HomeCredit Default Risk prediction in finance industry. In order to deliver a positive loan experience, Home Credit makes use of a variety of alternativedata like transactional information--to predict their clients' repayment abilities.
We will be using various statistical and machine learning methods to make these predictions.

Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their business to be successful. This necessitates the use of techniques like vintage analysis to establish these definitions based on historical data.

Home credit specializes in providing loans to customers with limited or no credit history, making accurate risk assessment a crucial aspect of their business operations.

This capstone project focuses on leveraging advanced data analytics and machine learning techniques to enhance the accuracy of credit risk predictions , thereby improving the overall efficiency and effectiveness of Home Credit's lending processes.

Business problem statement (GOALS):

1. Business Problem Understanding:

Credit Default Risk is a critical aspect of the financial industry, affecting lending decisions and risk management. A successful machine learning model can improve credit assessments, reduce defaults, and optimize

credit allocation. However, it also raises the challenge of maintaining transparency and regulatory compliance, given the increasing use of complex machine learning models in credit risk assessment.

The goal is to develop an accurate and reliable predictive model to assist financial institutions in making informed decisions regarding loan approvals.

2. Business Objective:

The task involves developing a machine learning model to assess the creditworthiness of Home Loan applicants. This is crucial for banks to decide whether to approve or reject Home loan applications.

The task involves developing a machine learning model to assess the credit worthiness of the applicant. This is crucial for banks to decide whether to approve or reject the applications.

3. Objective :

1. **Risk Prediction Improvement:** The primary goal of this capstone project is to enhance the accuracy of credit default risk predictions. By leveraging historical data, machine learning algorithms, and predictive modeling, the project aims to develop a robust risk assessment model that can identify potential defaulters more accurately.
2. **Feature Engineering and Selection:** The project involves in-depth exploration and analysis of various features and factors that contribute to

3. credit default risk. Through feature engineering and selection, the aim is to identify the most relevant variables that significantly impact the likelihood of default, thereby refining the model's predictive capabilities.
4. **Model Interpretability:** The project places emphasis on creating models that are not only accurate but also interpretable. By employing

5. interpretable machine learning techniques, stakeholders within Home Credit can gain insights into the factors influencing credit decisions, facilitating more informed and transparent decision-making processes.
6. **Scalability and Efficiency:** The developed models will be designed to be scalable, accommodating the diverse and large-scale data sets typical of Home Credit's operations. This ensures that the models can be integrated seamlessly into the existing credit assessment workflow, improving efficiency without compromising accuracy.

Methodology: The methodology encompasses several key steps, including data collection and preprocessing, exploratory data analysis, feature engineering, model development, validation, and interpretation. Various machine learning algorithms such as logistic regression, decision trees, random forests, and gradient boosting will be explored and compared to identify the most suitable model for predicting credit default risk.

Data Sources: The project will utilize Home Credit's historical loan application data, encompassing information on customer demographics, financial history, previous credit behaviour and other relevant features. External data sources, such as macroeconomic indicators and credit bureau information, may also be integrated to enhance the predictive power of the models.

Expected Outcomes: Upon completion of the Home Credit Default Risk Capstone Project, the anticipated outcomes include:

1. **Improved Credit Risk Models:** Enhanced models for predicting credit default risk, contributing to more accurate decision-making in the loan approval process.

2. **Interpretable Insights:** Clear and interpretable insights into the factors influencing credit decisions, aiding in better understanding and management of credit risk.
3. **Scalable Solutions:** Developed models that can be seamlessly integrated into Home Credit's existing infrastructure, ensuring scalability and efficiency in handling a large volume of loan applications.

4. Data Description:

- It includes all the previous records of credits provided by other financial institutionsreported to credit bureau in the **Bureau table** consisting of **Columns :**
  SK_ID_CURR, SK_ID_BUREAU, CREDIT_ACTIVE , CREDIT CURRENCY , DAYS_CREDIT,CREDIT_DAY_OVERDUE [17 Columns

- Monthly Balances of previous credits in credit bureau in the **Bureau Balance table**
  consisting of **Columns :**
  SK_ID_BUREAU, MONTHS_BALANCE, STATUS [3 Columns]

- **Point of Sale Balances (POS Cash Balance)** table which provides insights about MonthlyBalance snapshots of POS (Point Of Sales) and cash Loans that the applicant had with Home credit. It consists of **Columns:**
  SK_ID_PREV, SK_ID_CURR, MONTHS_BALANCE, CNT_INSTALLMENT, CNT_INSTALLMENT_FUTURE, NAME_CONTRACT_STATUS, SK_DPD, SK_DPD_DEF
  [8 Columns]

- **Credit Card Balance** table which provides insights about the Monthly Balance of the previous credit cards that the applicant has with home credit. It consists of **Columns:**
  AMT_CREDIT_LIMIT_ACTUAL,
  AMT_DRAWINGS_ATM_CURRENT,

AMT_DRAWINGS_CURRENT,
AMT_INST_MIN_REGULARITY,
AMT_PAYMENT_CURRENT,
AMT_PAYMENT_TOTAL_CURRENT,
AMT_RECIVABLE_PRINCIPAL, AMT_RECIVABLE

[23 Columns]

- **Previous Application** table which provides insights about all previous applications for home credit Loans of clients who had applied for loan in the past. It consists of **Columns:**

  SK_ID_PREV, SK_ID_CURR, NAME_CONTRACT_TYPE,
  AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT,
  AMT_DOWN_PAYMENT, AMT_GOODS_PRICE,

  WEEKDAY_APPR_PROCESS_START,
  HOURS_APPR_PROCESS_START,
  RATE_DOWN_PAYMENT [37 Columns]

- **Instalments Payments** table which provides insights about Repayment History for the previously observed distributed Loans from the organization with details of paymentsmade and missed. It consists of **Columns:**

  SK_ID_PREV, SK_ID_CURR, NUM_INSTALMENT_VERSION,
  NUM_INSTALMENT_NUMBER, DAYS_INSTALMENT,
  DAYS_ENTRY_PAYMENT,

  AMT_INSTALMENT, AMT_PAYMENT [8 Columns]

Data Shape:

Number of tables	9
Number of combined columns	120
Number of Rows	307,512

5. Data Dictionary of tables:

- **Bureau :**

All client's previous credits provided by other financial institutions that were reported toCredit Bureau.

- **Bureau Balance :**

Monthly balances of previous credits in Credit Bureau.
This table has one row for each month of history of every previous credit reported toCredit Bureau.

- **POS Cash Balance :**

This table has one row for each month of history of cash balance for each client.

- **Credit Card Balance :**

This table has one row for credit limit for each loan applicant.

- **Previous Application :**

All previous applications for Home Credit loans of clients who had applied for loan in thepast with application amount, down payment and credit details.

- **Instalments Payments :**

Repayment history for the previously disbursed loans from the organisation with detailsof payments made and missed.

From these tables we have selected the relevant features using SQL and applied the below approach on it.

6. Approach:

1. **Data Analysis, Cleaning/ Preprocessing:** The pre-processing of the dataset before performing ML functions involves the following:

   a. **Descriptive Analysis:** Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. Measures of variability help communicate the spread of distribution by describing the shape and spread of the data set.
   b. **Inferential Analysis:** Validating the inferences which are found with the help of descriptive analysis (Graphs) with the help of respective statistical tests if needed.
   c. **Treating Outliers:** Checking and analyzing for presence of Outliers in the numerical columns and treating those outliers using the IQR method or any relevant method.
   d. **Treating Missing Values:** Null values in the variables must be treated with suitable methods.
   e. **Encoding Categorical Variables:** Since, machine learning models are based on Mathematical equations and we can intuitively understand that it would cause some problem if we can either keep the Categorical data by encoding the categorical variable or we can drop by checking whether we need the variable for further modelling process because we would only want numbers in the equations.
   f. **Dropping Unnecessary Columns:** We are removing the columns which do not contribute to the model building or the columns which are of less, or of no importance.

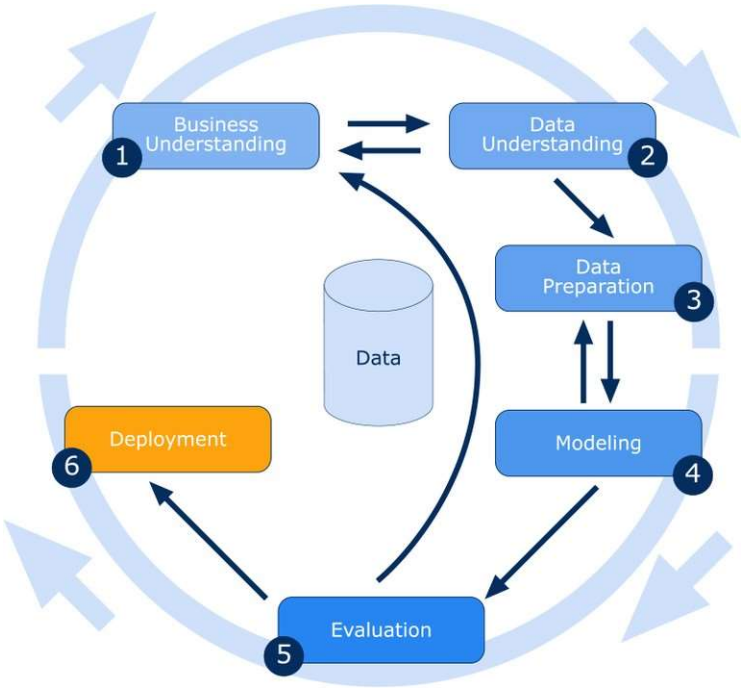
2. Exploratory Data Analysis:

Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modeling the data in order to apply Machine learning. Plotting in EDA consists of Bar-plot, Box plot, Scatter plot and many more using Univariate, Bivariate and Multivariate Analysis.

- **METHODS TO BE FOLLOWED:**

In this step, we will save the best model(pickle) and come up with a method or functionwhich takes patient data as input and re-admission status as output.

We can try to productionize the deployment using flask.



3. Data Preparation:

Data preparation is a crucial step to ensure the quality and relevance of the data for model training. Here are a few key steps that we have introduced in our project i.e. Home Credit Default Risk :

- Data Cleaning
- Garbage value/ missing value treatment
- Outlier treatment

Data Cleaning:

In the systematic process of data cleaning, we have undertaken a comprehensive approach to refine our dataset. Our efforts began by scrutinizing the various features, employing rigorous steps to ensure the highest data quality. By implementing a series of cleansing techniques and using advanced data manipulation methods, we have successfully enhanced the reliability and integrity of our dataset.

Garbage Value / Missing value treatment:
For numerical columns we have used the median value to fill the numerical columns, whereas for categorical columns we have used the mode value to fill up the categorical columns.

Missing value treatment

```
# for numerical columns we will use the median value to fill in the missing values.
```

```

```

```
for i in num_col.columns:

    # Impute missing values with the median
    median_value = num_col[i].median()
    num_col[i].fillna(median_value, inplace=True)
```

```
num_col.isnull().sum().sum()
```

```
0
```

```
# for categorical columns we will use the mode to fill in the missing values.
```

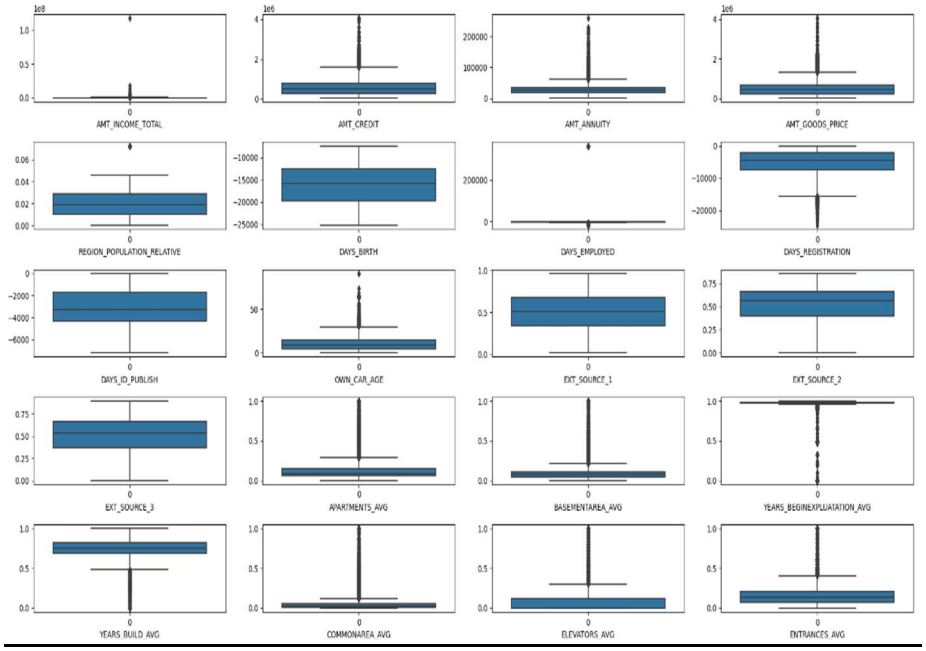
```
for i in cat_col.columns:

    mode_value = cat_col[i].mode().iloc[0]
    cat_col[i].fillna(mode_value, inplace=True)
```

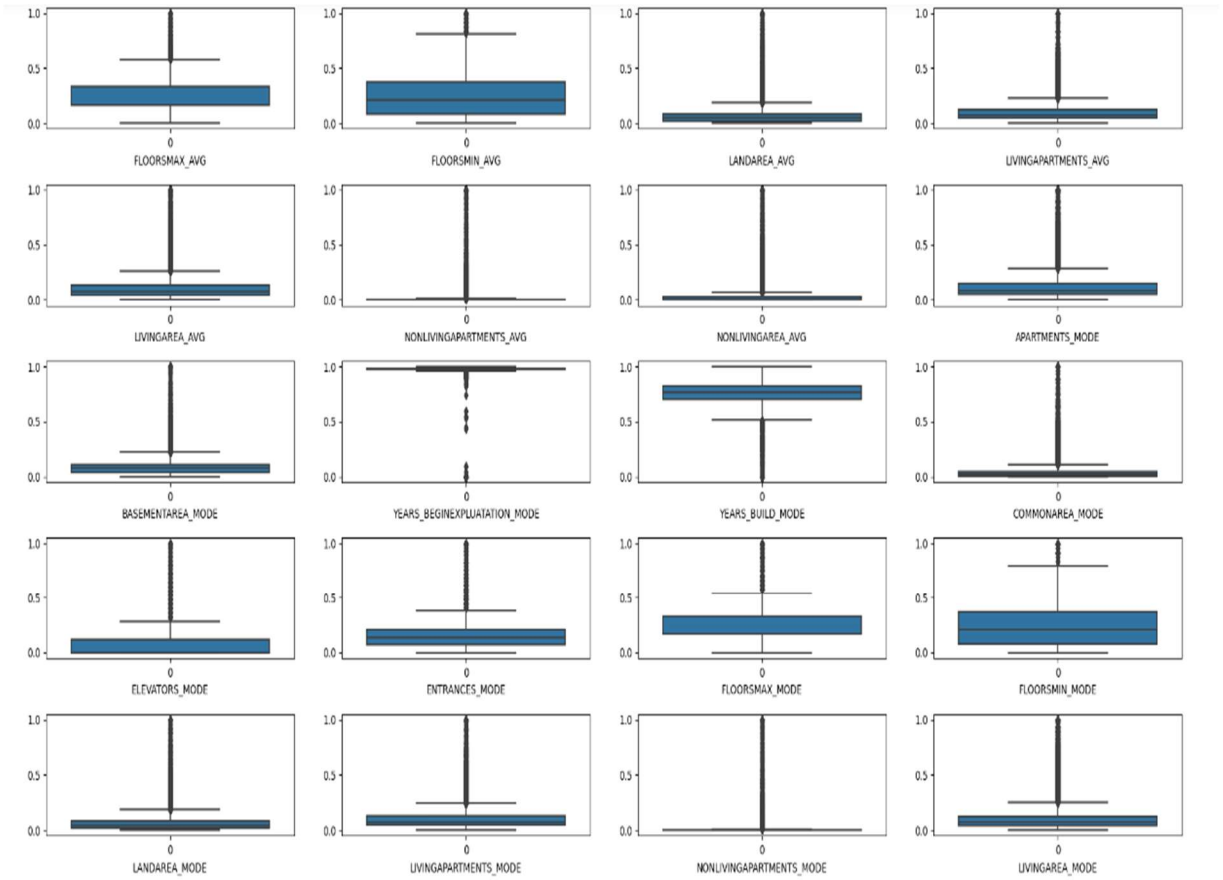
```
cat_col.isnull().sum().sum()
```

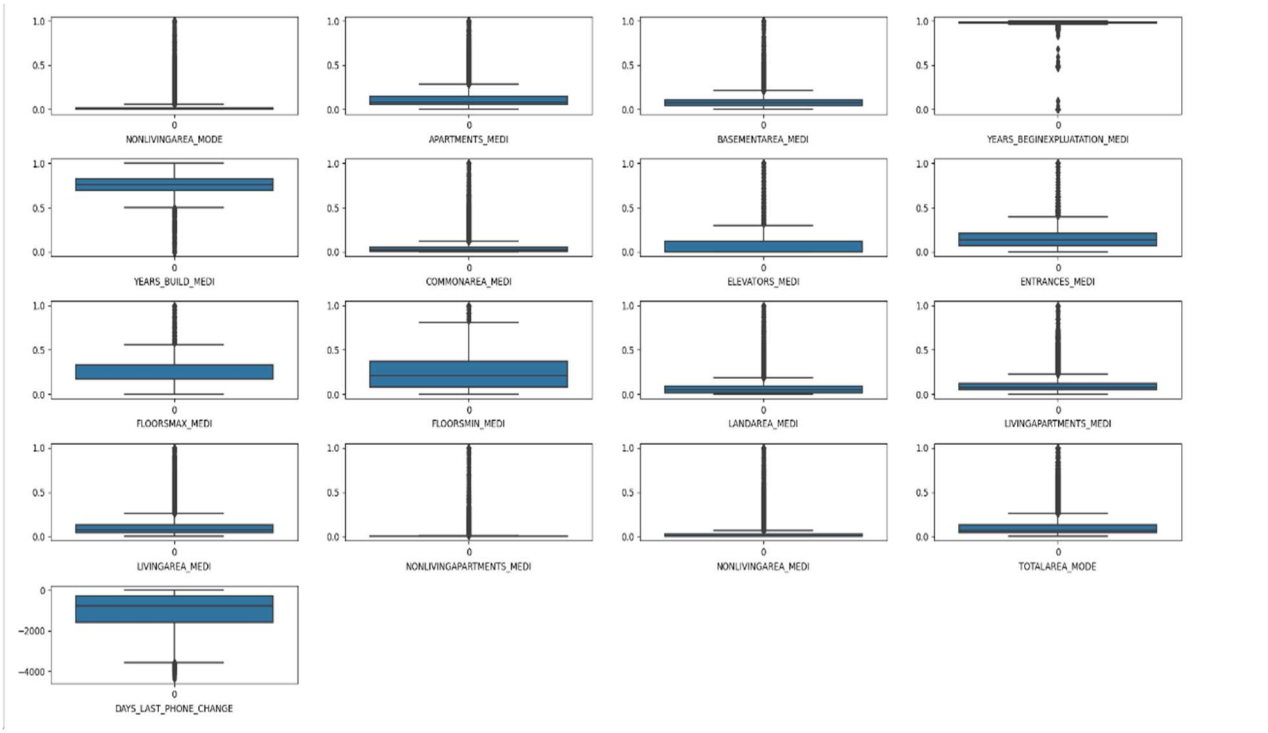
```
0
```

4. **Data Exploration**:

Univariate Analysis:


AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
OWN_CAR_AGE
EXT_SOURCE_1
EXT_SOURCE_2
EXT_SOURCE_3
APARTMENTS_AVG
BASEMENTAREA_AVG
YEARS_BEGINEXPLUATATION_AVG
YEARS_BUILD_AVG
COMMONAREA_AVG
ELEVATORS_AVG
ENTRANCES_AVG

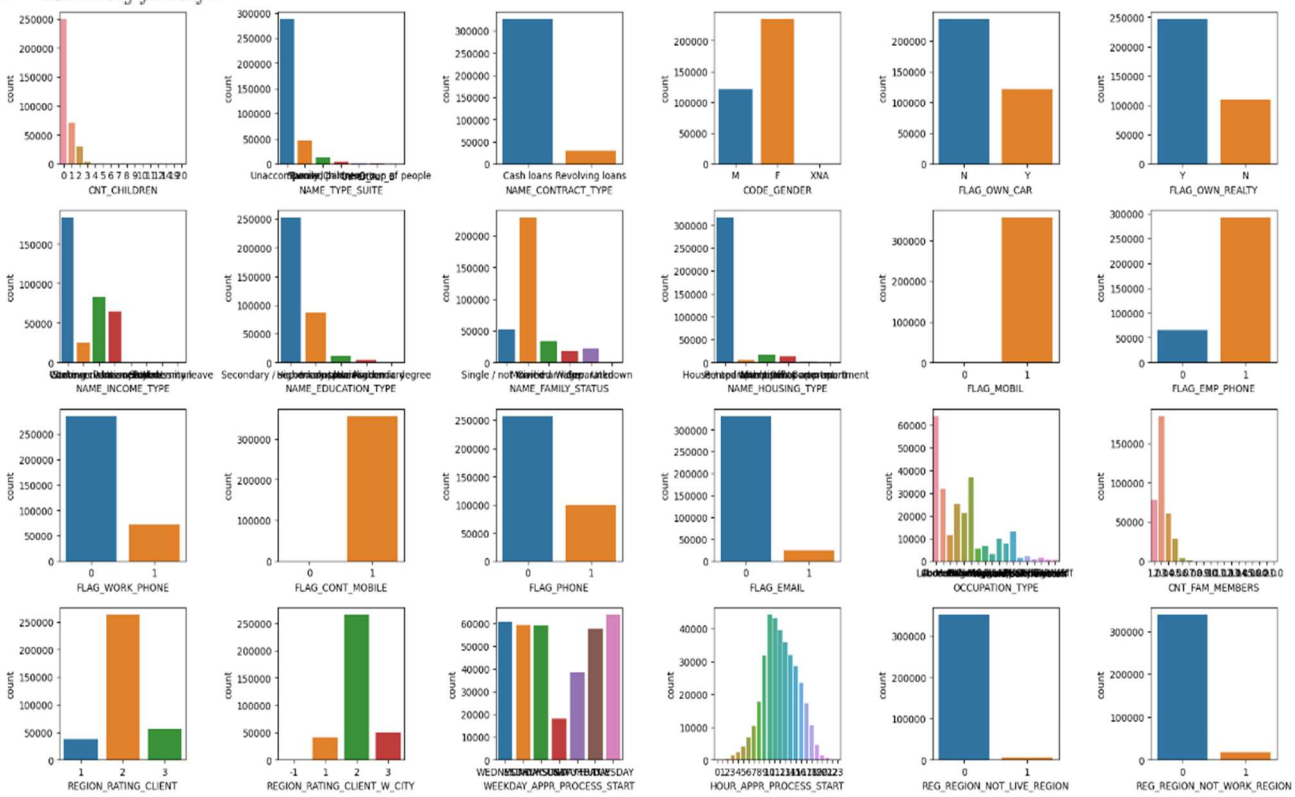

FLOORSMAX_AVG
FLOORSMIN_AVG
LANDAREA_AVG
LIVINGAPARTMENTS_AVG
LIVINGAREA_AVG
NONLIVINGAPARTMENTS_AVG
NONLIVINGAREA_AVG
APARTMENTS_MODE
BASEMENTAREA_MODE
YEARS_BEGINEXPLUATATION_MODE
YEARS_BUILD_MODE
COMMONAREA_MODE
ELEVATORS_MODE
ENTRANCES_MODE
FLOORSMAX_MODE
FLOORSMIN_MODE
LANDAREA_MODE
LIVINGAPARTMENTS_MODE
NONLIVINGAPARTMENTS_MODE
LIVINGAREA_MODE

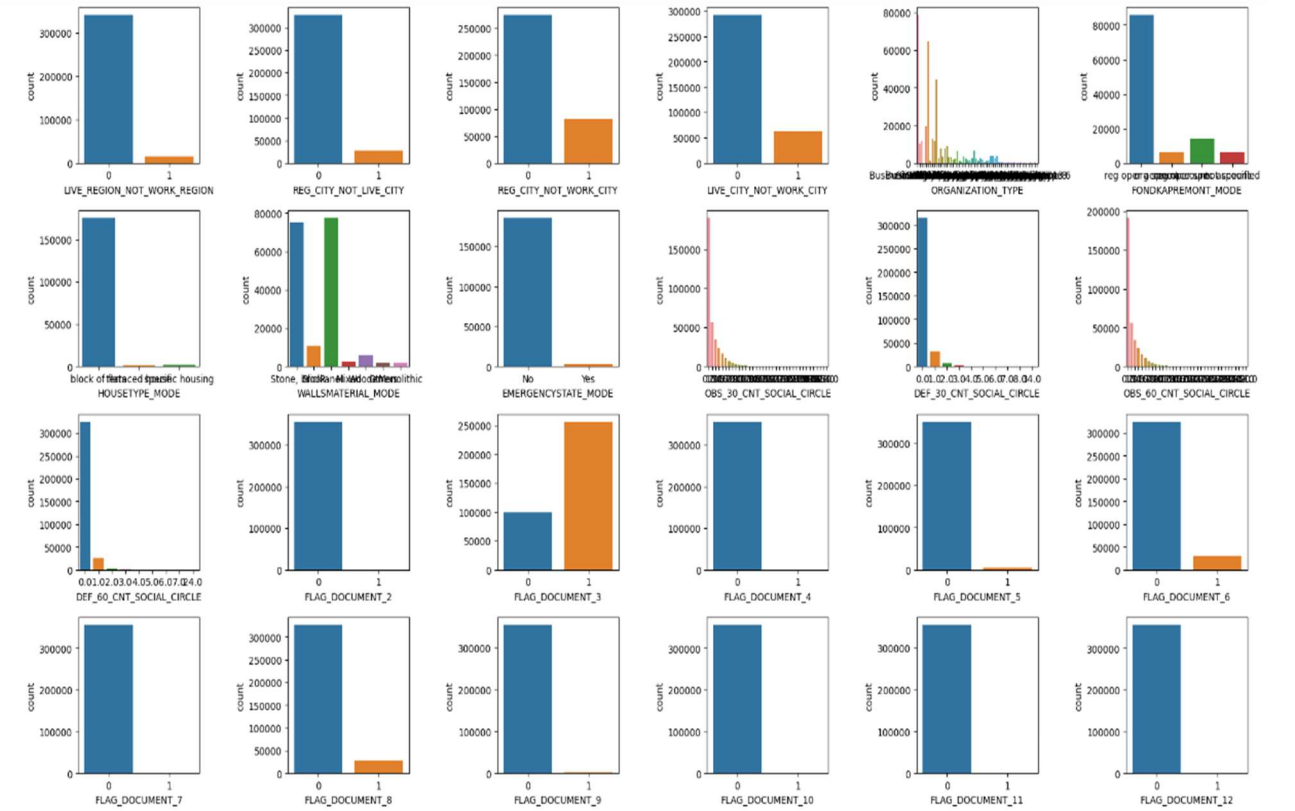


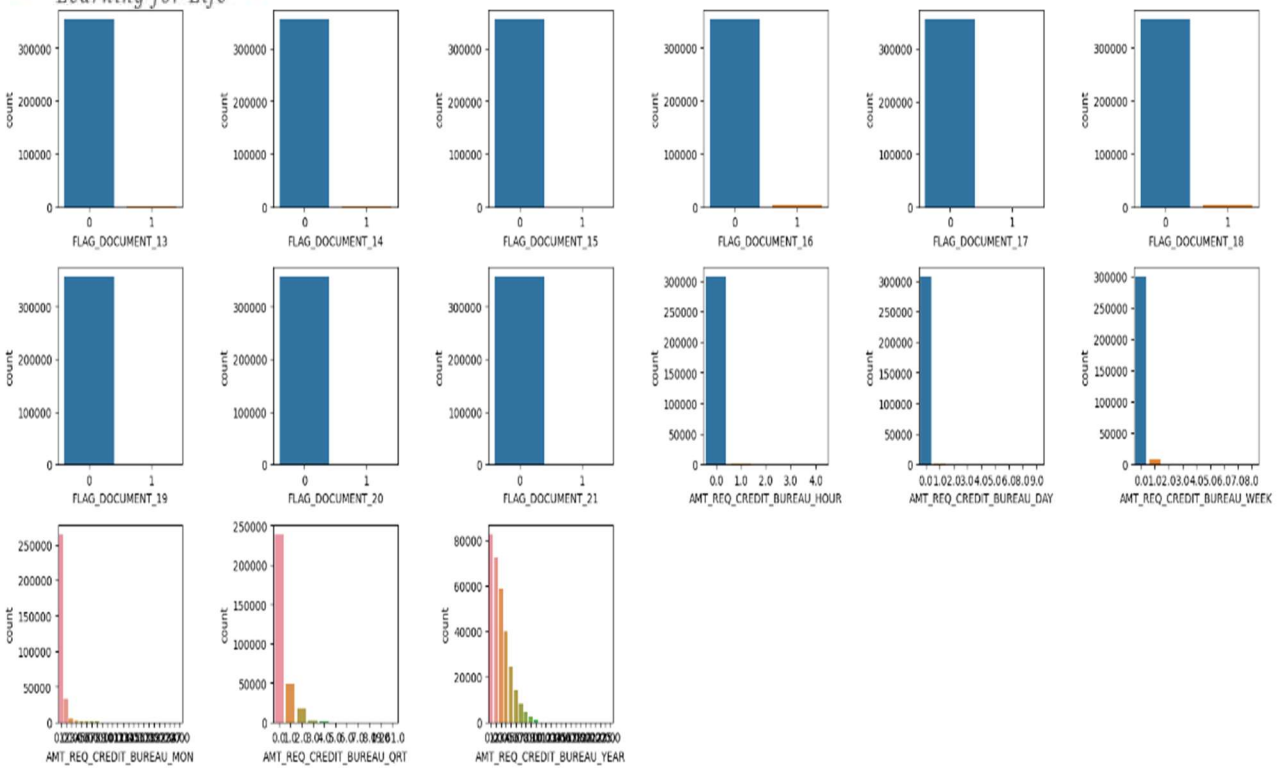
Numerical Columns:

'AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRICE','REGION_POPULATION_RELATIVE','DAYS_BIRTH','DAYS_EMPLOYED','DAYS_REGISTRATION','DAYS_ID_PUBLISH','OWN_CAR_AGE','EXT_SOUCCE_1','EXT_SOURCE_2','EXT_SOURCE_3','APARTMENTS_AVG','BASEMENTAREA_AVG','YEARS_BEGINEXPLUATATION_AVG','YEARS_BUILD_AVG','COMMONAREA_AVG','ELEVATORS_AVG', 'ENTRANCES_AVG', 'FLOORSMAX_AVG','FLOORSMIN_AVG','LANDAREA_AVG','LIVINGAPARTMENTS_AVG','LIVINGAREA_AVG','NONLIVINGAPARTMENTS_AVG','NONLIVINGAREA_AVG', 'APARTMENTS_MODE','BASEMENTAREA_MODE','YEARS_BEGINEXPLUATATION_MODE','YEARS_BUILD_MODE','COMMONAREA_MODE','ELEVATORS_MODE','ENTRANCES_MODE', 'FLOORSMAX_MODE','FLOORSMIN_MODE','LANDAREA_MODE','LIVINGAPARTMENTS_MODE','NONLIVINGAPARTMENTS_MODE','LIVINGAREA_MODE','NONLIVINGAREA_MODE','APARTMENTS_MEDI','BASEMENTAREA_MEDI','YEARS_BEGINEXPLUATATION_MEDI','YEARS_BUILD_MEDI','COMMONAREA_MEDI','ELEVATORS_MEDI','ENTRANCES_MEDI','FLOORSMAX_MEDI','FLOORSMIN_MEDI','LANDAREA_MEDI', 'LIVINGAPARTMENTS_MEDI','LIVINGAREA_MEDI','NONLIVINGAPARTMENTS_MEDI','NONLIVINGAREA_MEDI','TOTALAREA_MODE','DAYS_LAST_PHONE_CHANGE']]

greatlearning
Learning for Life







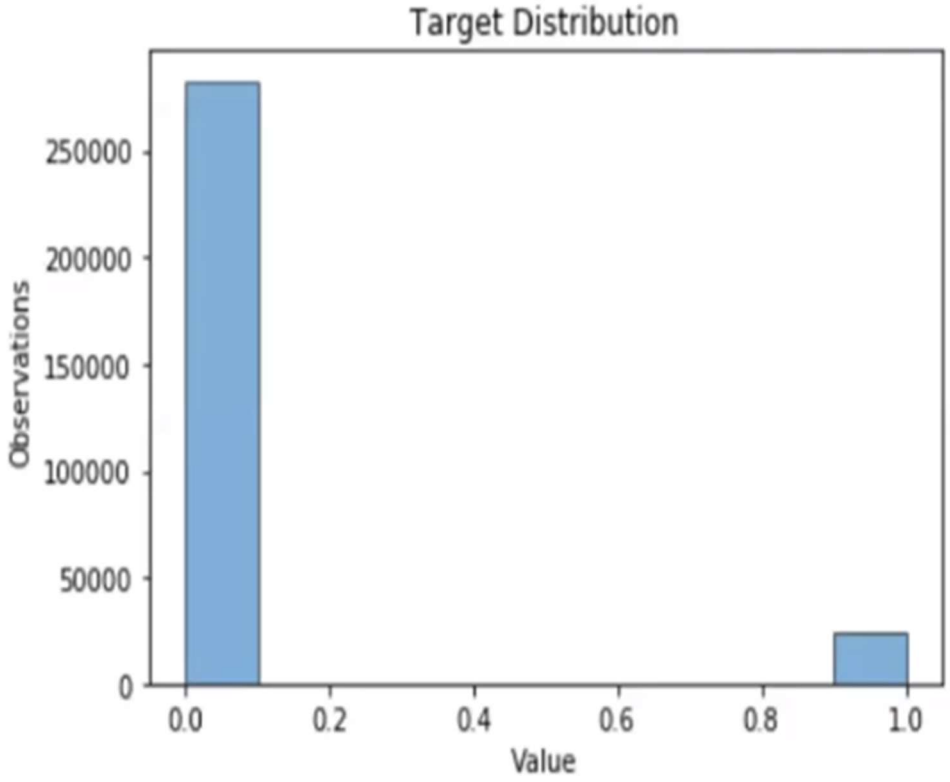
Categorical columns:

'CNT_CHILDREN','NAME_TYPE_SUITE','NAME_CONTRACT_TYPE','CODE_GENDER','FLAG_OWN_CAR', 'FLAG_OWN_REALTY','NAME_INCOME_TYPE','NAME_EDUCATION_TYPE','NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE','FLAG_MOBIL','FLAG_EMP_PHONE','FLAG_WORK_PHONE','FLAG_CONT_MOBILE','FLAG_PHONE','FLAG_EMAIL','OCCUPATION_TYPE','CNT_FAM_MEMBERS','REGION_RATING_CLIENT','REGION_RATING_CLIENT_W_CITY','WEEKDWWEEKDAY_APPR_PROCESS_START','HOUR_APPR_PROCESS_START','REG_REGION_NOT_LIVE_REGION','REG_REGION_NOT_WORK_REGION','LIVE_REGION_NOT_WORK_REGION','REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY','LIVE_CITY_NOT_WORK_CITY','ORGANIZATION_TYPE','FONDKAPREMONT_MODE','HOUSETYPE_MODE','WALLSMATERIAL_MODE','EMERGENCYSTATE_MODE', 'OBS_30_CNT_SOCIAL_CIRCLE','DEF_30_CNT_SOCIAL_CIRCLE','OBS_60_CNT_SOCIAL_CIRCLE','DEF_60_CNT_SOCIAL_CIRCLE','FLAG_DOCUMENT_2','FLAG_DOCUMENT_3','FLAG_DOCUMENT_4','FLAG_DOCUMENT_5','FLAG_DOCUMENT_6','FLAG_DOCUMENT_7','FLAG_DOCUMENT_8','FLAG_DOCUMENT_9','FLAG_DOCUMENT_10', 'FLAG_DOCUMENT_11','FLAG_DOCUMENT_12','FLAG_DOCUMENT_13','FLAG_DOCUMENT_14','FLAG_DOCUMENT_15','FLAG_DOCUMENT_16','FLAG_DOCUMENT_17','FLAG_DOCUMENT_18','FLAG_DOCUMENT_19','FLAG_DOCUMENT_20','FLAG_DOCUMENT_21','AMT_REQ_CREDIT_BUREAU_HOUR','AMT_REQ_CREDIT_BUREAU_DAY','AMT_REQ_CREDIT_BUREAU_WEEK','AMT_REQ_CREDIT_BUREAU_MON','AMT_REQ_CREDIT_BUREAU_QRT','AMT_REQ_CREDIT_BUREAU_YEAR'.

greatlearning
Learning for Life

TARGET DISTRIBUTION

Target Distribution



- Total obs: 307,511, Imbalance data: 8%

- Here we observe an imbalanced data, which we will further deal through under-sampling method of balancing the data.

TREATMENT OF IMBALANCE DATA

Random Under Sampler

- the "Random Under Sampler" is a technique employed to address the issue of class imbalance in the dataset, specifically when dealing with binary classification problems like predicting loan repayment. Class imbalance occurs when one class (e.g., loans repaid on time) significantly outnumbers the other class (e.g., loans with payment difficulties), which can lead to a biased model that performs poorly on the minority class.

- The Random Under Sampler tackles this problem by randomly removing instances from the majority class (over-represented class) until a more balanced distribution between the two classes is achieved. This is done to prevent the model from being overly influenced by the majority class and to ensure that it gives due consideration to the minority class, which is often of more interest in scenarios such as predicting loan defaults.

]:
```
from imblearn.under_sampling import RandomUnderSampler
```

]:
```
# Split the original dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Combine the training features and target variable into a single DataFrame
train_data = pd.concat([X_train, y_train], axis=1)

# Resample the majority class using random under-sampling
undersampler = RandomUnderSampler(sampling_strategy=0.32, random_state=42)
X_resampled, y_resampled = undersampler.fit_resample(X, y)
```

]:
```
y_resampled.value_counts(normalize=True)
```

]:
```
0.0    0.757575
1.0    0.242425
Name: TARGET, dtype: float64
```

]:
```
X_resampled.shape, y_resampled.shape
```

]:
```
((102403, 120), (102403,))
```

]:
```
# Let's work on the resampled data now
```

Statistical Analysis:

Statistical test

```
# lets check the chi-sqaure test on categorical columns

# null hypothesis = there is no relation between columns
# alternate hypothesis = there is a relation

# Assuming alpha as 0.05
```

```
insignificant_features = []

for i in cat_col.columns:
    table = pd.crosstab(df['TARGET'],df[i])
    ch1,p_value,dof,e = stats.chi2_contingency(table)
    print(i)
    print('there is a relation', p_value<0.05 ,p_value)

    if p_value>.05 :
        insignificant_features.append(i)
```

```
CNT_CHILDREN
there is a relation True 2.135381009734265e-29
NAME_TYPE_SUITE
there is a relation True 3.446825818703331e-06
NAME_CONTRACT_TYPE
there is a relation True 1.0235150721172847e-65
CODE_GENDER
there is a relation True 1.1290217848908289e-200
FLAG_OWN_CAR
there is a relation True 9.330994431109667e-34
FLAG_OWN_REALTY
there is a relation True 0.0006681470317545887
```

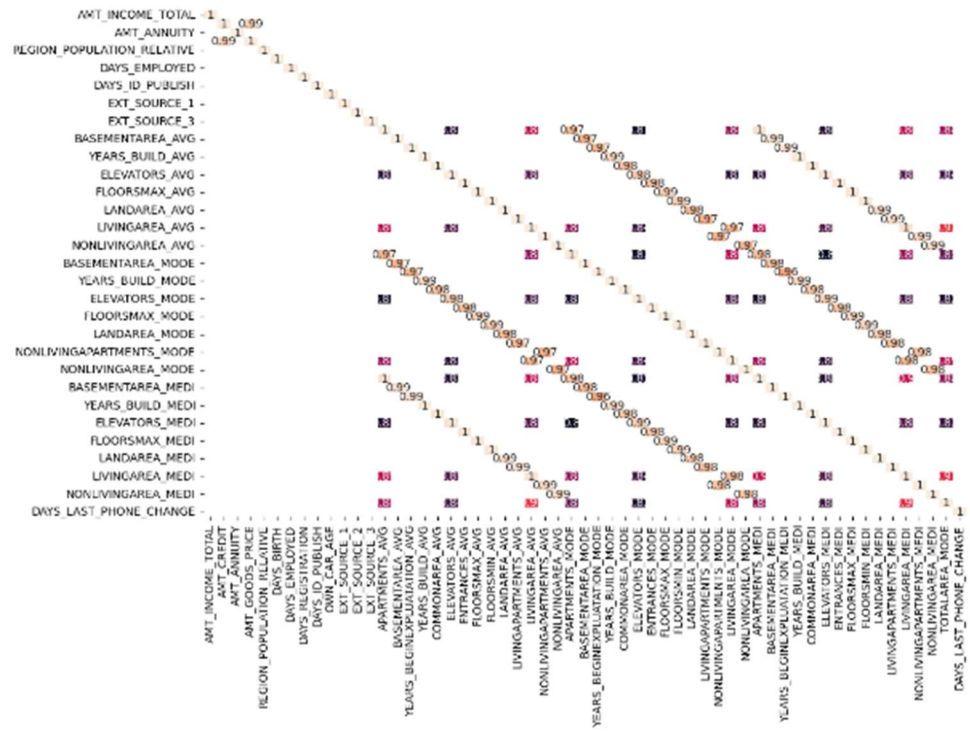
```
insignificant_features
```

```
['FLAG_MOBIL',
 'FLAG_CONT_MOBILE',
 'FLAG_EMAIL',
 'LIVE_REGION_NOT_WORK_REGION',
 'FLAG_DOCUMENT_4',
 'FLAG_DOCUMENT_5',
 'FLAG_DOCUMENT_7',
 'FLAG_DOCUMENT_10',
 'FLAG_DOCUMENT_12',
 'FLAG_DOCUMENT_17',
 'FLAG_DOCUMENT_19',
 'FLAG_DOCUMENT_20',
 'FLAG_DOCUMENT_21',
 'AMT_REQ_CREDIT_BUREAU_HOUR',
 'AMT_REQ_CREDIT_BUREAU_DAY',
 'AMT_REQ_CREDIT_BUREAU_WEEK']
```

CORRELATION MATRIX:

greatlearning
Learning for Life

```
plt.figure(figsize = (12,8))
sns.heatmap(num_col.corr()[(num_col.corr()>.80) | (num_col.corr()< -.80)    ] , annot=True , cbar=False)
plt.show()
```



MODELLING:

In the context of our project, the modelling phase plays a pivotal role in extracting meaningful patterns and making predictions based on the available data. Employing machine learning techniques, our modelling approach aims to uncover relationships and dependencies within the dataset. We utilize a systematic process of feature engineering, where relevant variables are identified and transformed to enhance their predictive power. The selection of an appropriate model is guided by the nature of the problem at hand, considering factors such as data distribution, complexity, and interpretability. Following model selection, rigorous training and validation procedures are implemented to ensure robust performance and mitigate overfitting. The effectiveness of the model is evaluated using various metrics, and fine-tuning may be applied to optimize its predictive accuracy. Throughout this modelling phase, our goal is to develop a reliable and interpretable model that contributes valuable insights to the objectives of our project, ultimately facilitating informed decision-making based on the data at hand.

After building and training various machine learning models on the Home Credit Default Risk dataset, it was observed that the AdaBoost classifier consistently achieved the highest accuracy compared to other models. AdaBoost, an ensemble learning technique, demonstrated superior performance in predicting the risk of default for home credit applicants.
The ensemble model combines multiple weak learners to create a robust and accurate predictive model. AdaBoost's ability to sequentially adapt and give more weight to misclassified instances during training contributed to its effectiveness in capturing complex patterns within the dataset.

The high accuracy of AdaBoost suggests its suitability for the Home Credit Default Risk prediction task, showcasing its potential as a reliable model for assessing credit risk and making informed lending decisions. Further evaluation metrics and detailed analysis may provide additional insights into the model's strengths and areas for improvement.

Name of the Model	Accuracy	Precision	Recall	F1
1. Logistic Regression	0.777794	0.227329	0.585314	0.327472
2. Logistic Regression (New Base)	0.777355	0.218301	0.5865314	0.318182
3. Decision Tree	0.675260	0.368691	0.334575	0.350805
4. Bagging	0.762756	0.211325	0.503667	0.297731
5. Adaboost	0.781065	0.236767	0.601668	0.339812
6. Random Forest	0.777452	0.162290	0.624803	0.257655