Introduction:

Home Credit Default Risk dataset offers a chance to explore credit risk assessment and financial inclusion for the unbanked population. We'll use EDA and feature engineering to find patterns and build strong credit risk models.

Understanding credit risk is paramount for financial institutions to mitigate defaults and facilitate right lending decisions. This dataset provides crucial insights into loan applicants, enabling us to gauge their creditworthiness.

Our goal is to analyze the dataset, identify key features, and create new ones to identify clients capable of repayment not being rejected. We'll use statistics and correlation analysis to gain insights into making better lending decisions.

Methodology:

Our methodology encompasses a systematic approach to data exploration, preprocessing, and feature engineering, handling outliers with transformation. The following outlines the key steps undertaken in this process:

1. Data Collection and Familiarization:
    - Data Source: We obtained the data from the Home Credit Default Risk competition on Kaggle.
    - Data Download: Using Jupyter Notebook or Google Colab, we downloaded all ten available CSV files from the Kaggle, uploaded them in the drive and read the files from there.
    - Data Model Review: We reviewed the data dictionary provided by Kaggle to understand the structure, variables, and relationships within the datasets. This helped us gain a preliminary understanding of the features relevant to creditworthiness.
2. Data Exploration:
   Conducted preliminary data exploration to understand the data types of variables, distribution, identify outliers as well as potential anomalies and gain insights into the dataset's characteristics.
    - Data Types: We employed pandas library functions like *dtypes* to identify the data types of each variable in the dataset. This helped us ensure compatibility during analysis and avoid potential errors.
    - Distribution: We analyzed the distribution of features using techniques like histograms, scatterplot, and boxplots. This visualization helped us identify any skewed distributions, particularly in numerical features like Days_Registration. Skewed distributions are transformed using quantiles after trying out multiple techniques to convert them into Gaussian distribution.

- Correlation: We calculated the correlation matrix and created heatmaps. This helped us understand which selected numerical variables are highly correlated with the target variable.
- Outlier Detection: We used Z-scores to identify outliers within numerical features. We leveraged z score greater than 2 and 3 to understand the outliers. After performing transformation, we again analyzed the outlier distribution.

3. Target Variable Analysis:

    - Target variable: 1 implies client with payment difficulties: The person made late payment more than X days on at least one of the first Y installments of the loan in our sample and 0 is for all other cases.

    We calculated the percentage of target variable. This provided insights into the class imbalance within the dataset. We noticed that there are 91.92% values for 0 and 8.07% values for 1 implying an imbalance in class distribution.

4. Handling Missing Values:
    - Missingness Assessment: We found missing values by .isnull().sum() so that we understand the missing values in the new merged data frame. In the new data frame, we have only included the selected numerical variables, categorical variables, and the new variables created via feature engineering.
    - Imputation Techniques: For categorical values, we replaced nan with unknown and for numerical data, we replaced nan values with zero along with creation of a new column indicating missingness. This new column creation is done to enhance model in machine learning process.

5. Handle categorical Features:
    - Cardinality Analysis: We assessed the cardinality within the selected five categorical variables. We noticed out of 5, 4 features have the number of categories less than 10 implying low cardinality. But one variable, namely organization type has high cardinality with 58 different values.
    - Rare Value Identification: For rare value identification, we used 5% threshold. We used this threshold on all the unique categories of every variable to undertsand which ones belong to the rare values / categories.
    - Encoding Techniques: Below is the list of the selected categorical variables and their appropriate encoding techniques:

        1. NAME_CONTRACT_TYPE: One-Hot Encoding

        2. CODE_GENDER: One-Hot Encoding

        3. ORGANIZATION_TYPE: Target Encoding

4. WEEKDAY_APPR_PROCESS_START: One-Hot Encoding

5. EMERGENCYSTATE_MODE: Binary Encoding

Findings:

1. Data Overview:
    - Primary Identifier: Confirmed as SK_ID_CURR (unique values match total rows).
    - Target Imbalance: Significant imbalance (8.1% defaults vs. 91.9% non-defaults) requiring special handling during model training.

2. Missing Values and Feature Selection:
    - For categorical values, we replaced nan with the word unknown.
    - For numerical data, we replaced nan values with zero along with the creation of a new column indicating missingness. This new column creation is done to enhance model in machine learning process.

    Feature selection:

    - Selected numerical variables:
        i. CNT_CHILDREN
        ii. AMT_INCOME_TOTAL
        iii. AMT_CREDIT
        iv. AMT_ANNUITY
        v. REGION_POPULATION_RELATIVE
        vi. DAYS_BIRTH
        vii. DAYS_EMPLOYED
        viii. DAYS_REGISTRATION
        ix. DAYS_ID_PUBLISH
        x. HOUR_APPR_PROCESS_START

    Variables with high correlation with the target variable:

	vars	TARGET
1	DAYS_BIRTH	0.078239
2	DAYS_ID_PUBLISH	0.051457
10	DAYS_EMPLOYED	0.044932
3	DAYS_REGISTRATION	0.041975
9	REGION_POPULATION_RELATIVE	0.037227

    - Selected categorical variables:

i. NAME_CONTRACT_TYPE
ii. CODE_GENDER
iii. ORGANIZATION_TYPE
iv. WEEKDAY_APPR_PROCESS_START
v. EMERGENCYSTATE_MODE

3. Feature Exploration:
    - Correlation analysis Identified strong correlations between numerical features and TARGET. Top 5: DAYS_BIRTH, DAYS_ID_PUBLISH, DAYS_EMPLOYED, DAYS_REGISTRATION, REGION_POPULATION_RELATIVE
    - When we looked at z-score over 3, we noticed outliers in DAYS_REGISTRATION, REGION_POPULATION_RELATIVE

4. Feature Engineering and Transformation:
    - After trying out multiple transformations, quantile transformation was applied to normalize skewed features.
    - We created 5 new features that might be useful for the model.
        i. Average amount balance
        ii. Max credit limit
        iii. Total active credits
        iv. Average days credit
        v. Avg Credit utilization
    - Categorical variables analyzed for cardinality and distribution (e.g., NAME_CONTRACT_TYPE with two categories, one dominant at 90%+) indicating potential class imbalance.

These findings offer valuable data characteristics for credit risk analysis modeling, including structure, missing data, correlations, and potential features.

Challenges and Solutions:

1. Outliers: We identified outliers in features DAYS_BIRTH, DAYS_ID_PUBLISH, DAYS_EMPLOYED, DAYS_REGISTRATION, REGION_POPULATION_RELATIVE. We addressed these by either removing them, after transforming the variables tried the following transformation:
    - Logarithmic transformation (log)
    - Square root transformation (sqrt)
    - Cube root transformation (cube_root)
    - Exponential transformation (exp)
    - Inverse transformation (inv)
    - Logit transformation (logit)

- Box-Cox transformation (boxcox)
- Quantile transformation (quantile)

From which we finalized Quantile transformation.

2. Feature Engineering Complexity:
   Merging the two datasets: previous applications and application training, and aggregating the following:
   - Averaging the AMT_BALANCE.
   - Taking the maximum of AMT_CREDIT_LIMIT_ACTUAL.
   - Summing the ACTIVE_CREDITS.
   - Calculating the average of CREDIT_UTILIZATION and DAYS_CREDIT.

Conclusion:

Key Findings from Credit Risk Analysis (Home Credit Default Risk dataset):

Data exploration provided insights into the data by examining dimensions, cardinality, missing values, correlations, and visualizations such as correlation heat maps, scatter plots, bar plots, violin plots, and box plots. These insights supported the selection of methods for future steps. Helped to identify the primary key i.e SK_ID_CURR, the correlation heatmap between the target variable and other variables gives an overview of the correlation. The box plot between target and other variables gives the distribution which suggests how the target is distributed over the variables.

Upon closer examination, it was observed that the new features contain a significant number of null values. Particularly, the 'days employed' feature exhibited an unusually high number of outliers. Future analysis will involve feature selection, including dimensionality reduction and detection of multicollinearity. Features will be selected using methods such as PCA, Lasso and Random Forest. Subsequently, based on the selected features.Additionally, we plan to run various models such as XGboost, stratified K fold and utilize different metrics to evaluate the performance of each model.

Colab Link: https://colab.research.google.com/drive/1KFB-LC51bzy1vFJfo9VpR4LWZGq6Bp_P?usp=sharing#scrollTo=981O-D_6nCcE