Data Wrangling Techniques Used in Dataset Preparation for Capstone (HOME CREDIT DEFAULT PREDICTION)

1. Introduction

- Credit default risk is one of the critical factors in determining the profitability of banks and lending institutions, which in turn, affects the stability of the economy. Being able to develop a mathematical model to quantify the risk associated with a borrower, is crucial for lenders.

2. Source

- The dataset has been obtained from a past Kaggle competition organised by a HOME CREDIT, non-banking financial company that lends primarily to people with very less or non-existent credit history. Hence, judging the ‘credit-worthiness’ of the borrower and quantifying credit risk is very important to minimize losses.

3. Dataset Description

- The main data is present in the file “application_train.csv”. There are many auxiliary files that contain details about the customers credit card balances, previous inquiries for credit, credit rating bureau data, etc. .

- There are 123 columns in the dataset, comprising categorical and numeric features. The pandas types allocated to the features are “int64”, “float64” and “object”. But we need to apply appropriate conversions, depending on the feature’s meaning.

- Features like client’s income, credit amount, etc are numeric, and no preprocessing is required for them as they have been identified correctly by pandas. Null values in these columns are replaced with either a central statistic(mean/median/mode), or 0, depending on the significance of the feature on the TARGET. The detailed process for handling null values has been described below.

- Some columns are stored as type ‘object’, but in reality, they are categorical. The values in these fields are in the correct format, and no string preprocessing is required as these will be converted to categorical type for modelling. Null values will be handled based on the importance of the feature on the TARGET, as described below.

4. Removal of NULL values

- Columns having a small number of null values - some features like customer income have a very few number of null values. The rows having null values contain valuable information about other features. Hence, the null values were replaced by the median values of those columns. The data contains some outliers, hence the median was used instead of the median.

- Some columns have a huge number of null values. For example, the feature ‘OWN_CAR_AGE’ has more than two-thirds of the values as null. This is expected, since many customers of Home-Credit do not own cars. To handle such features, we evaluated the statistical significance of the feature on the target variable. If the feature was statistically significant, we will proceed with replacing the null values, with either 0, or a central number representative of the feature (mean or median). To test the statistical significance of the feature, we will use either F-test in regression or ANOVA analysis. Here, since the output is categorical variable and inputs are numeric, ANOVA would be a better choice. If the feature was not significant, the it was dropped.

5. Handling of data in incorrect format

- The numerical columns contain numbers in the correct format. There are no unnecessary commas, or other special characters. Hence, no preprocessing is required initially. For modelling, these features may be normalized depending on type of model used and feature variability.

- The categorical columns contain data in correct format. The values are informative and meaningful. As these columns will be converted to ‘category’ type from ‘object’ type, no preprocessing is required on the values. As feature selection progresses, the correct method for handling the categorical columns (one-hot encoding, label encoding, etc.) will be decided.

6. Handling of outliers

- Since, the dataset has 123 columns, hence, complex relationships may exist between the input columns and the TARGET. As such, handling of outliers needs to be done with caution.
- As the dataset has been obtained from a credible source, very few outliers were observed. Visual and descriptive EDA was carried out to statistically analyse the features and remove outliers. For example, boxplot of the feature ‘AMT_INCOME_TOTAL’ was plotted and data points beyond a limit of 2 times the IQR were removed.

7. Removal of insignificant features (feature selection)

- Since the dataset contains a large number of features, there may be multi-collinearity in the features. This is readily evident in some cases, for example, a higher amount of loan will have a higher annuity (EMI). In other cases, it may not be visible. Such multicollinearity will be handled using statistical analysis. Appropriate methods like univariate regression analysis and ANOVA will be used.