**Home Loan Credibility Assessment**

**Data Acquisition:**

The loan application data, which is provided by [Home Credit](http://www.homecredit.net/about-us.aspx), is used to predict whether or not an applicant will be able to repay a loan. The data is available on <https://www.kaggle.com/c/home-credit-default-risk/data>.

The data is acquired from seven different sources. The first dataset, application\_train/ application\_test, is the main training and testing data with information about each loan application. Each row is identified by the feature SK\_ID\_CURR. The TARGET feature in the training data represents load repaid by 0 and not repaid by 1. The second data source bureau provides the client's previous credits from other financial institutions. The third one is bureau\_balance, which provides monthly balances of previous credits in Credit Bureau. POS\_CASH\_balance provides monthly balance snapshots of the previous point of sales and cash loans that the applicant had with Home Credit. The fifth data source credit\_card\_balance presents the monthly balance snapshots of previous credit cards that the applicant has with Home Credit. All previous applications for Home Credit loans of clients who have loans are mentioned in the previous\_application. Repayment history for the previous loans is provided in the seventh data source installments\_payment.

This project will clean the data according to the following strategy:

**1. Missing value treatment:**

Missing data can lead to a wrong prediction or classification. There are 122 columns in the training data and 67 of them contain missing values. Most of the missing value columns contain more than 50% of missing values. Deleting the missing values is not a good option for this case. We can use imputation to fill in the missing values. XGBoost, LightGBM or other algorithms can be used to deal with the missing values.

**2. Outlier Detection and Treatment**

We will encode categorical variables before we detect outlier. We will use Label Encoding for those categorical variables which have only two categories. In our training data set we have only 3 such categorical variables. The rest of the categorical variables are encoded using One-Hot Encoding.

We will use various visualization methods, such as box-plot, histogram and scatter plot, to detect outliers. Each outlier will be treated separately. For example, the DAYS\_EMPLOYED column contains outliers and all the outliers have the same value. We will set those outliers with missing values, np.nan, and then create a new boolean column indicating whether or not the value was anomalous. Later, we will fill the nan values with the median of the column.

**3.** **Feature Engineering:**

Feature engineering is the process of extracting more information from existing data. There are two main steps in feature engineering Variable transformation and Feature creation. There are several feature engineering techniques are available but in this project, we will use Polynomial features construction method. It creates interactions between features, which can improve the predictive power of our model. For instance, we can create polynomial features using the EXT\_SOURCE variables and the DAYS\_BIRTH variable. We will also add the following features in our training data set.

1. DAYS\_EMPLOYED\_PERC: It is calculated by dividing DAYS\_EMPLOYED by DAYS\_BIRTH, which mention the percentage of days the client is employed.
2. INCOME\_CREDIT\_PERC: Calculated by dividing AMT\_INCOME\_TOTAL by AMT\_CREDIT. This feature represents the percentage of the credit for the client.
3. INCOME\_PER\_PERSON: The per-person income in a family is calculated by dividing AMT\_INCOME\_TOTAL by CNT\_FAM\_MEMBERS.
4. ANNUITY\_INCOME\_PERC: The percentage of the annuity of the client is calculated by dividing AMT\_ANNUITY by AMT\_INCOME\_TOTAL.

**4. Customer Analytical Record:**

All the records or rows in the training data set are unique for each client. SK\_ID\_CURR column is unique for each client.

The Exploratory Data Analysis (EDA) will be presented in the Data Story project.