About the Data:

The data set provided by Home Credit for the competition is in the form of csv. There are 8 files in total with main files being ‘application\_train.csv’ and ‘application\_test.csv’. These files contain the features from the loan applications filled by customers along with some numerical features already calculated and used by the company for credit scoring. Rest of the files are related to applicant’s history from previous loans. For this project, I will primarily use ‘application\_train.csv’ as it contains labelled data and the ‘application\_test.csv’, which contains unlabelled data, for imputing missing values and outlier detection and handling. The rest of the files contain information that would be helpful in achieving a better absolute performance but that is not relevant to this project as we are more concerned with their relative performance.

The reason for selecting this data set is that it contains the highest number of data points than the ones in previous studies. These numbers are shown in the table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| File name | Rows (Data point) | Columns (Features) | Numerical Features | Categorical Features | Boolean  Features | Miscellaneous  Columns |
| Application\_train | 307511 | 122 | 78 | 16 | 26 | 2 (ID, Label) |
| Application\_test | 48744 | 121 | 78 | 16 | 26 | 1 (ID) |

Table 1: Input files description

The only difference in the features of application\_train.csv and application\_test.csv is the ‘TARGET’ column that represents the label of data points in training file. The test file does not have labels, but it is still useful for the purposes mentioned above.

As can be seen from the table above, the features can be divided into three types, categorical ones depict nominal attributes with two or more types. The second type are the numerical features that have quantitative values. The third is nominal attribute with strictly two types of values. These will be treated separately then categorical types because the data set has them defined by integer that can take 0 or 1 (boolean) values.

A close up of a logo

Description generated with high confidenceAs it is real world data and the fact that Home Credit is still in business, we expect the data to be imbalanced i.e instances of good loans would be more than the bad loans. Otherwise, the company would be out of business for good. The data imbalance is as below:

|  |  |  |
| --- | --- | --- |
| Value | Instance Count | % |
| 0 | 282686 | 91.927118 |
| 1 | 24825 | 8.072882 |

Table 2: Data imbalance

Figure 1: Data imbalance in training file

Imbalance in class instances is a challenge for classification algorithms. As most of the real-world data is not balanced, it would be a meaningful characteristic to identify for the algorithms under study.

All null values in our data are considered as missing values. Total number of columns with missing values is 67/121 features. For the sake of brevity only top and bottom 10 features are shown in the table below.

|  |  |  |
| --- | --- | --- |
| Feature | Missing Values | % of Total Values |
| COMMONAREA\_MEDI | 214865 | 69.9 |
| COMMONAREA\_AVG | 214865 | 69.9 |
| COMMONAREA\_MODE | 214865 | 69.9 |
| NONLIVINGAPARTMENTS\_MEDI | 213514 | 69.4 |
| NONLIVINGAPARTMENTS\_MODE | 213514 | 69.4 |
| NONLIVINGAPARTMENTS\_AVG | 213514 | 69.4 |
| FONDKAPREMONT\_MODE | 210295 | 68.4 |
| LIVINGAPARTMENTS\_MODE | 210199 | 68.4 |
| LIVINGAPARTMENTS\_MEDI | 210199 | 68.4 |
| LIVINGAPARTMENTS\_AVG | 210199 | 68.4 |

Table 3: Top 10 Missing features

|  |  |  |
| --- | --- | --- |
| Feature | Missing Values | % of Total Values |
| DAYS\_LAST\_PHONE\_CHANGE | 1 | 0.0 |
| CNT\_FAM\_MEMBERS | 2 | 0.0 |
| AMT\_ANNUITY | 12 | 0.0 |
| AMT\_GOODS\_PRICE | 278 | 0.1 |
| EXT\_SOURCE\_2 | 660 | 0.2 |
| DEF\_60\_CNT\_SOCIAL\_CIRCLE | 1021 | 0.3 |
| OBS\_60\_CNT\_SOCIAL\_CIRCLE | 1021 | 0.3 |
| DEF\_30\_CNT\_SOCIAL\_CIRCLE | 1021 | 0.3 |
| OBS\_30\_CNT\_SOCIAL\_CIRCLE | 1021 | 0.3 |
| NAME\_TYPE\_SUITE | 1292 | 0.4 |

Table 4: Bottom 10 Missing Features

From table 3, some features that represent medians, modes and averages of the same variable have equal amounts of entries missing. This implies the missing value is not due to data entry but the variable itself is not applicable for this data instances. We can safely ignore such features to simplify our data set. It would impact the performance of our classifiers equally, so we don’t need to worry about them.

From table 4, some features have missing values count small enough that a simple strategy of imputing them would be a better than discarding those features for this project.

I have chosen to discard all features that have more than 20% of values missing. And the rest of the features would be imputed with mean values of those features. For imputing values, I will incorporate the application test data points as well.

After excluding the features with missing values greater than 20% we are left with the following:

|  |  |  |  |
| --- | --- | --- | --- |
| Categorical Features | Boolean Features | Numerical Features | Total Count |
| 11 | 26 | 35 | 72 |

Before we impute any features, it is important that I identify and handle outliers. Otherwise, the mean values that I choose to impute the data with, would be skewed by those outlier values. As an example of outlier values, figure below shows the distribution of ‘DAYS\_EMPLOYED’ column.

A screenshot of a cell phone

Description generated with very high confidence

Figure 2: Days employed feature distribution

It is evident from figure 2 that there is a value above 350000 that is present in both the labelled class instances of the training data as well as the test data. Also, according to the description of this column given by the company shows that the days are counted in negative from the date of application. Thus, a positive value would suggest that we treat this as a missing value. For now, I replace the value with NaN. The distribution after removal of this outlier is shown below.

A screenshot of a social media post

Description generated with very high confidence

Figure 3: Days employed distribution after correction

Now the data distribution seems to be correct. Outliers in the rest of the data set have been replaced with NaN as well.

Data Imputation Strategy: Nominal with modes, Numerical with mean.

Categorical Values Handling Strategies:

1. With Null values
   1. Without categorical features
   2. OHE
   3. WOE
2. Without Null Values
   1. Without categorical features
   2. OHE
   3. WOE
3. With Null Values Imputed
   1. Without categorical features
   2. OHE
   3. WOE
4. Native Strategy of LGBM i.e with null values and without any encoding
   1. Without categorical features
   2. With categorical features

Take the best overall data set from above and try feature selection (6 voters strategy)

Observe improvements.

Counter data imbalance with stratified K-fold training.