**Classification Models (Milestone 2)**

**Team ID: SC-5**

|  |  |
| --- | --- |
| Section | Name |
| 4 | نوران السيد عبد الحميد فرحات |
| 4 | ياسمين خالد محمد عبد الحميد |
| 4 | مريم هشام عيد حسن |
| 1 | إسراء عزت عبد المنعم علي |
| 2 | حسناء إيهاب عبد الشافي حسن |

**Preprocessing:**

there are many approaches in preprocessing some of them are applied to all models and some are not.

|  |
| --- |
| Applied to all models |
| * Drop column “CreditGrade” & “TotalProsperPaymentsBilled” as more than 70% of their values are null. |
| * We have dropped all rows that have null values in “ProsperRating (Alpha)” column. |
| * We have dropped “ListingNumber” as it is just like ID of the row. |

|  |
| --- |
| Not applied to all models |
| 1. Removing outliers from all numerical features. ["BorrowerAPR", "EmploymentStatusDuration","CreditScoreRangeLower",   "CreditScoreRangeUpper", "RevolvingCreditBalance", "BankcardUtilization", "AvailableBankcardCredit", "TotalTrades”, "DebtToIncomeRatio", "BorrowerRate", "StatedMonthlyIncome", "LoanOriginalAmount","LoanNumber"] |
| 1. Replace null values by (mean or median or mode). |
| 1. Apply “Normalization” on numerical data. |
| 1. Apply “Feature Encoding” on string data. |
|  |

Functions:

* Remove outliers: In this function we take “Dataframe” & “Feature columns” as a parameter and return the dataframe without outliers.

Graphical user interface, text, application, email

Description automatically generated

We plot the “Box plot” to see the values before removing outliers and after it the following table appear:

|  |  |
| --- | --- |
| Before removing Outliers | After removing Outliers |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

* Mean, Median, Mode functions:

Graphical user interface, text, application, email

Description automatically generated

In these functions we take “Dataframe “ as an input and fill the null values in this dataframe with “Mean or Median or Mode” and then return the modified dataframe and “Mean or Median or Mode” values to be used later in testing phase.

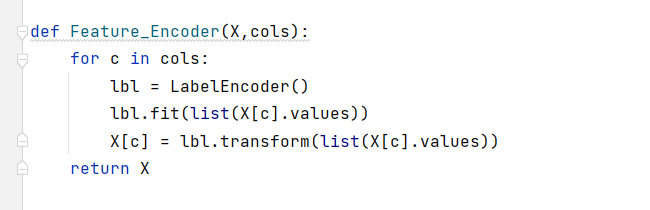
* Feature Scaling function:

Graphical user interface, text

Description automatically generated

In this function it takes dataframe and the range of normalization as input and return Normalized dataframe.

* Feature Encoding function:



This function takes dataframe and string columns as input and return modified dataframe as an output.

* To\_null function:

Graphical user interface, text, application

Description automatically generated

This function is for converting a specific value in a certain feature into null if it’s meaning is similar to null, so it takes dataframe and the value which we want to replace it with null as inputs and return modified dataframe.

**Feature Selection:**

We applied “mutual information” to visualize the correlation between the features and the target column and selected the first 13 feature that have stronger correlation with the target column.

#picture

**Flow of code:**

* We split the data into X\_train, X\_test, Y\_train, Y\_test as 80% of the data are for training and 20% are for testing.
* We dropped the rows that has null values in the output column.
* We dropped the columns that we have mentioned above in preprocessing.
* We converted “Not displayed'' in “IncomeRange” and “Not available” in “EmploymentStatus” with null.
* We removed outliers from numeric columns.
* We replaced null values of the numerical data with the median of the column.
* We replaced null values of the categorical data with the mode of the column.
* We normalized the numerical data.
* We applied “Feature Encoder” on string data.

**Models:**