FDS PROJECT

Matricula: 1911545/ Name: MONA KASHANI

AUC score:

```
Fold  3 AUC : 0.846555
Full AUC score 0.847163
```

1. Data preprocessing:

Data preprocessing is one of the crucial steps in data mining. This process can include noise values, modifying values and normalization numerical values.

- Modifying invalid values:

Data collection methods are often uncontrolled and therefore unreliable values may be found in the data. As an example of invalid values include out of range values, such as negative execution time impossible data composition and unspecified values. These can cause the process of data mining and data extraction in some problems. Therefore, representing and controlling data quality is a priority before performing analysis. Parts of the processing process has been done manually and a number of unreliable data have been deleted which are included in 4 records with XNR values from the application_train table and the CODE_GENDER attribute. Since this database has a lot of NULL values, two methods have been used to remove these values.

'Categorical properties' has been used for removing NULL values with 'the most frequently used' method.

To eliminate NULL values 'numeric properties' has been used with the 'mean value' method.

This table shows some of the properties with NULL values.

- Convert categorical data to numeric data:

Since some features are categorical, they must convert to numerical. For this purpose, each of the modes in that feature is assigned a unique number.

- Normalization:

In order to normalize numerical data, minimal and maximum mapping normalization to the range [-1,1] has been used. In this method, each property is mapped separately so that the minimum value of the property is -1 and the maximum value is +1.

$$x_n = 2 \times \frac{x - min}{max - min} - 1$$

In this regard, minimum and maximum are the minimum and maximum values, respectively.

- Data balancing:

For balancing the data smooth and oversampling are the methods which have been used.

2. Extracting features and data integration:

In this stage, first these five features

(DAYS_EMPLOYED_PERC, INCOME_CREDIT_PERC, INCOME_PER_PERSON, ANNUITY_INCOME_PERC, PAYMENT_RATE)

from the 'application_train' table will be extracted and then combined with two tables 'bureau' and 'bureau_balance' using the 'SK_ID_BURE' feature. And a table with the number of 240 attributes has been created, which will be used in the next step pf this database.

3. Customers classification:

Finally, at this stage, the use of existing classification algorithms is a model of prepared data which in the previous steps has been extracted in order to classify customers. For this purpose, ‘Decision Tree’, ‘Logistic Regression’, ‘Random Forest’, ‘LightGBM’ and ‘XGBClassifier’ algorithms have been used. And the results of each of these methods can be seen in the table below.

score	Algorithms
0.66514	Decision Tree
0.73360	Logistic Regressor
0.70930	Random Forest
0.77018	Light GBM
0.72934	XGB Classifier
0.77499	LGB