MS4610: Introduction to Data Analytics

Project Report by Group 001

Abstract:

In this project we were asked to predict whether the credit card applicants will go default or not in 12 months from the time of application submission, by training various machine learning models over the 83k sample point training dataset, provided by Amex company. The basic approach carried out by our group is data pre-processing, exploratory data analysis (EDA), feature generation, model training and evaluation.

We have chosen **LightGBM** model, a gradient boosting framework that uses tree-based learning algorithm, and trained it using the train split as it can handle large size of data well and takes lower memory to run. Another important reason is that it focuses on the accuracy of results. Although, the model results were compared to those of two other types of classification models namely, Logistic Regression (LR) and Random Forest algorithm. On comparing the model performance using the F1 score evaluation metric, the LightGBM model had superior predictive performance among the models trained with a **F1 score** of **0.536**. We have used the popular python programming language for our data processing, visualization and modelling purposes.

Understanding the dataset:

We are provided with the training and test datasets along with the data dictionary sheet, that describes each of the variables. The training dataset contains the customer application and bureau data with the default tagging i.e., if a customer has missed cumulative of 3 payments across all open trades, his default indicator is 1 else 0. The first column “application_key” in the datasets will serve as the unique reference value for each individual record, as a resident of the city can submit only a single application form.

Approach Methodology:

Data pre-processing:

- The NumPy and Pandas data science libraries are imported and used for data manipulation and analysis. Also, the Seaborn library is imported to plot graphs for visualization as part of EDA.
- By importing ‘preprocessing’ package from ‘sklearn’ library of python the categorical variable “mvar47” from the dataset is encoded using the LabelEncoder( ) function. It tells the type of product the applicant has applied for (C = Charge, L = Lending).
- As there’s inconsistency in the way missing values are entered in the dataset, which are given using both ‘na’ and ‘missing’, they are all replaced with ‘NaN’ values by using the replace( ) function of the NumPy library.

Exploratory Data Analysis:

- By plotting the distribution curves for all the variables we notice that some of the variables are highly skewed with large skewness value, so we applied box-cox transformation technique for making normality by using the boxcox1p( ) function, under scipy.special library, with the lambda (exponent value) as 0.13.
- **Feature generation:** With our level of domain knowledge in credit card terms we have added four additional variables as columns to the datasets.
    - By adding the variables ‘mvar3’, ‘mvar4’, ‘mvar5’ we created a variable called ‘Severity_sum’
    - Similarly, we have, ‘num_active_full’ = ‘mvar17’ + ‘mvar18’
    - ‘max_credit_active’ = ‘mvar7’ + ‘mvar8’
    - ‘num_active_75%’ = ‘mvar19’ + ‘mvar20’

Model training and evaluation:

- The training dataset has been split into train and validation data by assigning 20% into test_size parameter of the train_test_split( ) function.
- We train the LightGBM model over both training and validation data splits with the required parameter values and a maximum number of iterations as 50
- The AUC score in the training split is obtained as 0.798.
- The threshold value used is 0.5, which implies the prediction is 1 (default) when the probability value is > 0.5 and 0 (non-default) otherwise.

Evaluation metric values of all the applied models:

	Accuracy	F1 score	Balanced Accuracy	Gini index
LightGBM	0.782	0.536	0.679	0.411
Random Forest	0.686	0.468	0.619	0.460

- **Gini** can assess the accuracy of a prediction around whether a loan applicant will repay or default. **Gini** is measured in values between 0 and 1, where a score of 1 means that the **model** is 100% accurate in predicting the outcome.

Conclusions:

- From the metric score table, we can determine that LightGBM model has the best accuracy and efficiency based on evaluation by F1 score and accuracy values.
- And we can say that lightgbm model has showed higher performance than other well-known models of Logistic Regression and Random forest.
- Although the Gini index value is lesser for lightgbm model but it is still comparable with that of random forest and both are less than 0.5.
- Also the lightgbm model, as the name ‘light’ suggests, runs much faster than the other classification models tested even with large sets of data, which shows its effectiveness in performance.