FDS Kaggle Project A.Y. 2019/2020

Final private AUC	**Dario Ruggeri**	**Romeo Lanzino**	**Simone Ercolino**
0.78567			

For the final solution, we achieved the score of **0.78567** with LGBM classifier and running time of about 30m.

Our working final model is based on **LightGBM**, a gradient boosting framework that uses tree based learning algorithms, proved very useful in solving this particular problem. The strategies used in our model are:

- Anomaly detection algorithm (treating anomalies as NA)
- Removing columns with high ratio of missing values (more than 98%)
- One-hot encoding for categorical features
- Creating new features (polynomial features and ratios of features) starting from quantitative features with relative higher correlation with the target
- Merging all files and aggregate rows with multiple criterions (mean, median, mode) in this way creating new variables from the merged CSVs
- K-fold cross validation with 10 folds

In order to deal with overfitting and improve the AUC, we tuned some parameters of the LGBM. In particular, we used lambda L1 and L2 for regularization (0.03 and 0.08), tuned the number of estimators (10K with early stop at 200 rounds), used GOSS as boosting method because it improved our score and the speed of the algorithm and again, to deal with overfitting: number of leaves and max depth of trees set at 33 and 8, subsample rate of 0.85.

Previous models we implemented included many other pre-processing, feature engineering and classification methods. Most of these can still be found in our scripts unused. Previously we tried to use Logistic Regression, Naive Bayes, Random Forests, Neural Networks, and ADABoost as classifiers, with some more refined pre-processing and Feature Engineering methods, such as:

- Frequency encoding (for the FAMD)
- Imputing missing values with predictive algorithms (KNN, Iterative and Simple Imputer)
- Transformation of variables (normalization and log transformation)
- Discretization of continuous variables
- PCA, trying to select components based on explained variance or correlation with the target
- Delete also rows with high ratio of missing values

Moreover, this procedure led us to a maximum score under 0.77.

Therefore, we decided to adopt LightGBM, for which we noted that most of the pre-processing techniques applied in the previous models were not improving AUC score, and in some cases even worsened it. That brought us to our final solution.

The project is strusctured as follows:

- The main directory is named as FinalProjectFDS and is where the program expects the “data” folder, where the CSVs should be.
- Inside the main folder there is the script folder where all the .py modules are, main.py included.
    - Moreover, inside scripts there is the “feature_engineering” package where other .py modulesare placed.
- The submission csv will be created from the program in a folder that will be in the main directory in the “logs” folder, that will be created automatically by running the script.

Roma, 11/02/2020