Week 10 submission
Project: Bank Marketing Campaign

Group Name: Evolve Data
Name: Dmitry Sharukhin
Email: sharuhinda@gmail.com
Country: Russia
College/Company: Finval GC
Specialization: Data Science

Problem description
ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).
Bank wants to use ML model to shortlist customer whose chances of buying the product are more so that their marketing channel (tele marketing, SMS/email marketing etc.) can focus only to those customers.
This will save resource and their time (which is directly involved in the cost (resource billing)).

The task is to create binary classifier to forecast the probability of customer's agreement to open term deposit

Github repo link: https://github.com/sharuhinda/bank_marketing_campaign/tree/review

EDA notebook:
https://github.com/sharuhinda/bank_marketing_campaign/blob/review/2_eda.ipynb

Reviews thread: https://github.com/sharuhinda/bank_marketing_campaign/pull/2

Based on the Weight of Evidence approach the predictive powers of features were evaluated. All features were left except 'duration'.

Dataset is highly imbalanced so missing-like data was retained to encode it later.

3 highly correlated features found. They are 'euribor3m', 'nr.employed' and 'emp.var.rate' (Pearson's $r > 0.9$). This is a potential problem for some ML model types.

Final recommendations:
- Perform features encoding using 3 main strategies (and their mixed options with dropping any of high correlated and/or low variance features) and compare their impact on model's metrics:
    - Use WoE approach to encode all category and category-like features without imputation
    - Use WoE partially only on truly categorical features without imputation, numeric features convert to reasonable values and cutoff outlying values
    - Don't use WoE on any feature, impute values and encode categorical features by ordinary way
- As the problem is of binary classification class and has to be a relatively interpretable we can try Logistic Regression, Bayes Naïve Classifier, Decision Tree, Random Forest and LightGBM / CatBoost (with SHAP explanation) classes