README

# Practical Application 3

In this application assignment, our task was to compare the performance of the classifiers (k-nearest neighbors, logistic regression, decision trees, and support vector machines) using a dataset from the UC Irvine Machine Learning Repository related to the marketing of bank products over the telephone.

* Business Objective: Determine which classification method (k-nearest neighbors, logistic regression, decision trees, and support vector machines) has the best performance in predicting if a client will subscribe to a term deposit (target variable y).

#Data Understanding and Preparation

I began by reading in the dataset and performing initial exploratory data analysis to inspect and take a quick look at the data to understand the features and state of the data (quality, completeness, format etc.)

Next, I did data preparation such as replacing unknown values with NaN, one hot encoding categorical features, establish ordinal encoding for education level, converting feature data types and dropping rows with null values.

From there, I developed visualizations for both the categorical and continuous features against the target variable why to see which features had the strongest predictive indicators and which features were good candidates for further feature engineering.

To begin the modeling phase, I split my dataset into train and test sets, and created baseline f1 score for to understand where what the F1-score baseline performance is that we are aiming to beat.

#Modeling

Before doing any feature engineering from observations gained in the visualizations, I set up the models for Logistic Regression, K Nearest Neighbors, Decision Tree and Support Vector Machine. Then, I made adjustments to certain features such as dropping unneeded or irrelevant feature columns, handling outliers, and creating quadratic and interaction features, retraining the 4 models and monitoring for model improvements along the way.

Finally, I cross validated my results and tuned hyperparameters with grid search as well as tested two other modeling techniques we’ve covered in the course so far, Random Forest and Gradient Boost.

#Evalutation

All the models tested performed better than the baseline F1-scrore. However, across all modeling techniques, Decision Trees performed the best, regardless of the feature engineering and other model improvement efforts. Of those efforts, the Decision Tree model with the best performance had a F1\_score of .3762 which is an improvement from our baseline f1-score of .1216.

#Next Steps and Recommendations

If a business were to move forward with one technique for deployment or further optimization, I would recommend focusing on Decision Trees. They performed a little better than the rest but also are more easily interpreted and explainable to stakeholders.