DATA 622 Assignment 1

Erick Hadi

The Bank Marketing dataset contains data from a marketing campaign conducted by a Portuguese bank, with the goal of predicting whether a client will subscribe to a term deposit. The dataset contained a binary target variable, yes or no. The dataset also contained both numerical and categorical features. In order to perform some EDA, the dataset was split into categorical and numerical features. The first thing that was checked was missing values. Initially there were no missing values after running a simple check for null values. However, upon checking through the data there were placeholders such as unknown and 999 indicating that value was not known. So, those values were changed to null. The overall distribution was checked for each variable. Bar plots were made for categorical values and histograms were mad numerical values. There were a lot of imbalances and skewed data, which impacted which model would be selected for this dataset. Outliers were also checked through visual box plots, interquartile range, and z-score. Outliers can be significant in altering the distribution and relationship of the features and it is difficult to know what to do with them. Correlation between numerical values were checked and there were high correlations between euribor3m, emp_var_rate and nr_employed. High correlation between features is important to note because they may indicate redundancy and multicollinearity. The central tendency and spread of each variable were also checked to validate the visual check earlier. Lastly one of the patterns in the data was that some months had significantly less subscriptions than others. For example, March, September, October, and December had much higher no rates than the others.

After completing some EDA, it was fairly clear as to which algorithm would work well on this dataset. Three algortihms seem to be well fit for the task: Logistic Regression, Random

Forest Classifier, and Gradient Boosting. Logistic Regression is one of the more fundamental algortithms and the baseline for binary classifcation tasks. It is simple and fast to train, has interpretable coefficients which is highly useful for business, it handles categorical and numerical data well if they are encoded, and provides probabilities as well. However, it assumes linearity between features and the log odds of the target and is sensitive to multicollinearity. It also underperforms when trained on complex or nonlinear patterns. Random Forest is an ensemble algorithm of decision trees. It handles nonlinear relationships much better then logistic regression. It performs well when there are outliers and multicollinearity. However, it is less interpretable than logistic regression and much more computationally expensive. Also, if not tuned properly may be prone to overfitting. Lastly Gradient Boosting, is one of the more accurate and advanced algorithms in performing classifications tasks. It has built in handling for missing data and can handle imbalances in the datasets with tuning. It also has very good control over bias variance through hyperparameters. However, it slower to train and harder to interpret. It also has a high risk of overfitting on smaller datasets. The size of the dataset affects which algorithm performs better. More advanced algorithms are prone to overfitting on smaller datasets. Ultimately, the algorithm that seems best out of the three for this task is Random Forest. Since the dataset has both categorical and numerical features which interact in nonlinear ways, Random Forest will perform better than logistic regression. Also, there is no need for heavy preprocessing and the dataset is not that large. Random Forest robustness to overfitting would put it over gradient boosting.

The next step before training a Random Forest was data preprocessing. In this step, placeholder values were replaced with null values. Correlated features (emp_var_rate and nr_employed) were removed to reduced multicollinearity. Some new features were engineered such as contacted_before and has_credit_risk, this simplifies the data by grouping categorical features. Lastly, categorical features were encoded using one-hot encoding for nominal categorical variables. Ordinal variables were label encoded. After this, the data was split into training and testing sets, with stratification on the target to preserved class distribution. A Random Forest Classifier was trained using class_weight=’balanced’ to try to counter the class imbalance without altering the dataset much. The result, Accuracy: 92%, Precision: 0.70, Recall: 0.44, F1-Score: 0.54, ROC-AUC: 0.95. These results that the model performed well at separating positive and negative cases, but still struggled to capture all positive cases (low recall). However, this is a very highly imbalanced dataset (0.89 no, 0.11 yes).

In conclusion, according to the data customer subscription is rare with only ~11% customers subscribed to a term deposit. Which suggests that mass marketing is inefficient, a high cost for low yield. So, a more targeted marketing approach is required to improve the ROI. The datasets also have seasonality with specific months had higher rejection rates, March, September, October, and December. Certain factors such as longer calls and whether they were contacted previously seemed to be more successful in getting subscriptions. Although a lot of potential positive clients may be missed with the initial model it can still be used to prioritize leads. Which would enable efficient market targeting rather than mass marketing.