Report

March,2024

1 Customer Churn Prediction

For this dataset, we will build two classifiers, Adaboost and Random Forest to predict whether a customer will churn or not, and compare their performances. An appropriate metric to evaluate our model would be **recall**.

- We perform the train test split before proceeding, this helps us achieve realistic modeling, since we use only test data for evaluation and help and exclude it from all the pre processing steps. Without it, there is a chance of information leakage from the test data.
- We define utlity functions to avoid redundancy in our code.
- We define a **data preprocessing pipeline**, which is applied to our train and test data seperately.
- We compare our model's performances with baseline classifiers to check for the improvement that our models bring in.
- For tuning the parameters of our model, we perform **Bayesian Optimization**.
- We perform EDA on our dataset and also do a sanity check for fitting the models.
- For evaluation of both the models, we use evaluation metrics such as precision, recall, accuracy score, ROC-AUC curve and the confusion matrix.
- For our particular problem, we can conclude that Adaboost (with base learner as decision tree) is a better model for classification.

2 SuperMarket sales

The given task is to predict the gender of a customer of a supermarket, using two models, a decision tree and a random forest. Also, we predict the rating that a customer assigns, using two models- linear regression and decision tree regressor. The given dataset contains features such as branch of the store, product type purchased, quantity, etc.

- The data processing and EDA are performed in the same manner as in the previous task. Similar evaluation metrics are also used.

- For the task of predicting gender, we can see that the decision tree performs better than the random forest model, in terms of AUC score.

- We fit a linear regression model for predicting the rating. The results are not very satisfactory because we can see from the **QQ plot** that our data violates the assumptions of fitting a linear regression model. Also the **p-values** are not large enough, which indicates that none of the attributes contribute significantly to the final result.

- From the report values, the decision tree regressor performs better for predicting the ratings, as compared to the linear regression model.