Queen City Data science Hackathon

Motive:

The main motive behind this competition is to create a tool to help government officials and health professionals more effectively use their time and resources combating the negative effects of substance abuse in the United States. Goal is to accurately predict whether patients will complete their treatment as well as their length of stay. It is Regression and Classification problem. Submissions are judged based on accuracy and root mean squared error, respectively.

Data:

Substance Abuse Data set was provided which has 66 features and 1.6 million rows. Data includes Demographic and Geographic data. It also includes patients current and past substance abuse issues. Target labels are Length of stay and Reason for leaving treatment.

Exploratory Data Analysis:

Exploratory Data Analysis has been performed to understand all the 66 features and their correlation with other features. Data visualizations has been performed in tableau software. It played a crucial role in understanding the data well and how target are variables are varying based in independent features.

Data Preprocessing:

1. Removing nulls, NA and replacing with mean, median value.
2. Normalizing the data

Baseline model:

We first created our baseline model using knn for classification task and linear regression for regression task to see how well we can improve our models and to check how well our model performs in comparison to others. We got an accuracy of 55% and RMSE of 14.5 for our baseline models.

Feature engineering:

We then performed featured engineering to improve the models performance. We found importance of each feature with respect to target variables using xgboost feature importance function. Then we implemented each of the above models removing irrelevant features. The models miserably failed after implementing this.

Beating Baseline model:

We then started developing enhanced models that can accurately fit large data sets. We wanted to try Ensemble models which are very good with large data sets and they even train faster compared to Neural Networks. We have used LGBM classifier which is Boosting algorithm to improve accuracy and RMSE values. Our basic LGBM classifier produced an accuracy of 71% and RMSE of 9.1 without any Hyper parameter tuning which was great improvement from our Baseline model.

Why LGBM:

Light GBM is a gradient boosting framework that uses tree based learning algorithm.Light GBM grows tree vertically while other algorithm grows trees horizontally meaning that Light GBM grows tree leaf-wise while other algorithm grows level-wise. It will choose the leaf with max delta loss to grow. When growing the same leaf, Leaf-wise algorithm can reduce more loss than a level-wise algorithm. LGBM is one of the powerful algorithms when participating in Kaggle competitions.

Hyper parameter tuning:

Extensive Hyper parameter tuning has been performed to improve the results and make it to the top. Parameters we have tried changing are Number of leaves, Number of epochs, Depth, Number of estimators, Regularisation factor.

Our final model achieved a best accuracy of 74.2% and RMSE of 8.2 .