Hackathon Report

The final model used was a stack of LightGBM, catboost and random forest. This stack gave the best accuracy on the kaggle public leaderboard.

The first step was to do preprocessing on the data. Since most of the attributes were categorical, these were converted to ordinal data by assigning codes.

The classifiers used were support vector machine, random forest, gradient boosting classifier, xgboost, xgboost+dart, lightgbm, lightgbm+dart, lightgbm+goss, catboost. Random forest was observed to perform well with much fine tuning. Lightgbm+dart gave the best performance individually. Catboost was running the fastest and giving accuracy comparable to lightgbm. Thus a stack of these three models was used.

We also tried different encoding methods for categorical data which include using label encoders, one hot encoders, sum encoders and polynomial encoders. Polynomial encoder performed best and was thus selected.

Fine tuning was taking too long. Thus we also tried using the Flaml automl library. Flaml automl model was giving very good accuracy but it could not perform as well as manually tuned models.

The final parameters used for lightgbm+dart model were:

- n_estimators=1700
- drop_rate=0.37
- max_drop=65
- learning_rate=0.05
- reg_alpha=0.1

For the final random forest, 100 estimators were used. For the final catboost, 1000 estimators were used.