Submitted By: Alekhya Manem & Rachita Jain

Best Algorithm : **XGBoost**

Train AUC: **0.8329970253488329** Test AUC: **0.8126190938578574**

Code- Final Model

We consider the problem at hand as classification problem with binary outcomes - i.e, low or high length of stay (LOS). For this we tried the following three classes of machine learning algorithms:

Linear Classifiers - Logistic regression ; Decision Trees - Random forests; Gradient Boost - XGBoost & LGBM

Feature Engineering Strategies for XGBoost and LightGBM:

Method 1:

1. Drop all the columns with null values.
2. Find the hour based time difference between when the procedure actually completed (*PROC_DATE*) and when it was scheduled to be complete. *SCHED_START_DT_TM* - Delay
3. Remove all date based columns.

Method 2:

1. Columns with object data type to categorical.
2. Delay field same as Method 1.
3. Remove date time based columns.
4. Impute missing values with most frequent values.
5. PCA based dimensionality reduction.

Method 3:

1. Remove columns where null values > 60%
2. Columns with object data type to categorical.
3. Delay columns created same as in method 1.
4. Remove date time based columns.

Method 4:

1. Remove all columns with null values but *AGE_ON_CONTACT_DATE', 'BP_SYSTOLIC', 'PULSE', 'BP_DIASTOLIC.*
2. Impute the missing values with mean.
3. Create the Delay column as in method 1.
4. Remove date time based columns.

The model parameters were optimized by Hyperopt and the best parameters kept.

Feature Engineering Strategies for Logistic Regression and Random Forest

(Reference - 'Strategies for Handling Missing Data in Electronic Health Record Derived Data')

1. Filtered out rows with negative '*LOS*' values
2. Added a new categorical feature '*LOS_CAT*' with values for rows with '*LOS*'>5 as '1' and the rest as '0'
3. Filtered out rows where '*FEMALE*' column had null values
4. Dropped all 'closest lab value' and 'last hospitalization lab value' features as 'comorbidities' features signify the same
5. Converted '*BP_SYSTOLIC*' and '*BP_DIASTOLIC*' features to categorical feature with values 'low', 'high' and 'normal'
6. Null values in '*BMI*' and '*WEIGHT*' features are replaced with gender wise average values
7. Null values in '*BP_SYSTOLIC*', '*BP_DIASTOLIC*' and '*PULSE*' were set to avg human values
8. Null values in '*TOTALPREVIOUSHOSPVISITS*', '*TOTALPREVIOUSEDVISITS*', '*TOTALPREVIOUSPCPVISITS*', '*PREVIOUSSPECIALTYVISIT*', '*PREVIOUSURGENTCAREVISIT*' were set to 0
9. Finally one-hot encoded the categorical features

Results Summary

Algorithm	Parameter Configuration	AUC		Underfit/ Overfit/None
		Train	Test	
XGBoost	'max_depth':4, 'eta': 0.35, 'silent':1, 'objective':'binary:logistic', 'eval_metric': 'auc', 'learning_rate': 0.2, 'gamma': 3, 'min_child_weight': 10, 'subsample': 0.5, 'scale_pos_weight' : 3, 'maximize' : 'TRUE', 'n_jobs' : -1, 'n_estimator': 300	0.83	0.81	None
LGBM	'boosting_type': 'gbdt', 'objective': 'binary', 'metric': 'auc', 'max_depth' : 3, 'learning_rate': 0.1, 'num_threads' : -1, 'early_stopping_round' : 10, 'top_rate' : 0.3, 'other_rate' : 0.7, 'lambda_l1' : 30, 'lambda_l2' : 250	0.820	0.789	Overfit
Logistic Regression	‘solver’ = 'saga', ‘penalty’ = 'elasticnet', ‘l1_ratios’ = [0.1, 0.2, 0.3], ‘Cs’ = 20, ‘n_jobs’ = -1, ‘random_state’ = 0	0.503	0.467	Overfit
Random Forest	‘random_state’ = 0, ‘n_jobs’ = -1, ‘criterion’ = 'gini', ‘max_features’ = 5, ‘max_depth’ = 11	0.623	0.612	None

Deployment:

Since this is a medical use case and thus it can be expensive to ignore the false negative cases. They may lead to ignorance towards a potential medical requirement for a patient and thus AUCPR is the best evaluation technique which can be used to deploy the model. Various thresholds can be used to identify the high risk patients from the low risk patients on the basis of AUCPR scores and keeping the threshold to a maximum of 0.5 can reduce the chances of ignoring false negatives.