**OBJECTIVE**

To predict the 7-year survival rate of prostate cancer patients. The aim is to identify important indicators from the training data, select an algorithm and use it on the test data to derive the predictions.

**EXPLORATORY DATA ANALYSIS**

The training consists of 15,385 instances and 33 indicators including the target variable. Excluding ID, Symptoms and Diagnosis Date, the dataset had 5 ordinal, 2 nominal, 11 boolean and 12 numeric columns. The percentage of missing values in each column was calculated.

**UNIVARIATE ANALYSIS AND CORRELATION MATRIX**

Data Visualization for all the columns in the dataset which helps us to understand the distribution and outliers for numerical data and the variation of values in other types of data. It also gives us a glimpse into the variables that will have/not have a significant impact on prediction.

From the correlation matrix, multicollinearity between some of the variables was observed but it did not warrant any special treatment or transformation. Multicollinearity can alter the signs of certain coefficients, alter p-values and impact the significance of variables but it generally does not affect how well the predictive model fits.

**SYMPTOMS COLUMN TREATMENT**

The individual symptom values were stripped from each field and a list with just the unique symptoms was created. This list was used to traverse every row in the dataset to produce a column for each unique system indicating its presence.

**DATA TRANSFORMATION AND IMPUTING MISSING VALUES**

Dummy variables were created for Race and Side. The ordinal variables gleason\_score, t\_score, n\_score, m\_score and stage were factorized. MICE (Multivariate Imputation via Chained Equations) is one of the commonly used packages, especially when dealing with healthcare and pollution data. It predicts the missing values in one column by regressing against other columns and carries on iteratively until all columns are filled. By default, linear regression is used to predict continuous missing values. Logistic regression is used for categorical missing values.

**LOGISTIC REGRESSION MODEL**

The model was built to identify the significant variables in the dataset. This is very important from a business sense as some of the variables assumed to have significant predictive power from the univariate/multivariate analysis might not be statistically significant to the dependent variable. However, it does not mean that these variables are unnecessary to build the predictive model. Symptoms U05, P02, P03, S10, O11, O09, O08 were found to be statistically significant among other variables in the dataset.

**PREDICTION ALGORITHMS – LOGISTIC REGRESSION AND RANDOM FOREST**

The advantage with algorithms such as Logistic and Tree based ones are that they are simple, effective and easy to understand. Algorithms such as Neural networks and SVM are black box models which may give you better predicted values but offer no interpretability. The tree-based algorithms are better than these 2 algorithms in terms of handling missing values, robustness to outliers, computational scalability, ability to weed out irrelevant inputs and interpretability. Although some might argue that random forest is a black box model itself it provides good overall interpretation as opposed to local interpretation.

The other disadvantage of random forest is that it overfits easily. However, there are many ways to prune the trees and reduce its impact. I have chosen to go with Random Forest as my final model although other tree based boosting algorithms such as Gradient Boost Classifier and XGBoost can further improve model performance.

**VALIDATION, ROC/AUC AND PRECISION-RECALL CURVES**

Learning/Validation Curve gives us an idea as to how the accuracy or any other metric reacts to increase in the number of estimators or instances. The accuracy in our model does not decrease with the increase in estimators which indicates good model performance.

The ROC/AUC and Precision-Recall curves indicate how well our positive instances and negative instances perform as opposed to random guessing. Even though, we have a balanced dataset, the scores indicate that our model does well in identifying true positives and true negatives.