Enova Data Challenge

Objective: To predict if a person survived for 7 years after being diagnosed of prostate cancer on the basis data given in training_data.csv

Data exploration: The dataset had 15385 observations of 35 variables.

- Variables *psa_6_months* and *tumor_6_months* had more than 60% of NA values. Given the context of our problem and research about the subject, the values could not be missing due to random. The two columns were dropped.
- Symptom is an important factor to consider in the problem so the null values in symptom were removed.
- **MICE** package was used to impute the remaining missing values.
- The dataset had a mix of categorical and integer variables and few were multimodal distributions. So, the variables were converted to the appropriate datatypes.

Data Preparation:

Variable	Transformation
height and weight	BMI
gleason_score	Gleason score <=6 (Critical-Low) Gleason score = 7 (Critical-Medium) Gleason Score >8 (Critical-High)
tumor_diagnosis and tumor_1_year	Tumor_1_year - Tumor_diagnosis (Difference)
psa_diagnosis and psa_1_year	Psa_1_year - psa_Diagnosis (Difference)
age	Age less than 70 Age between 70 and 80 Age between 80 and 109(based on distribution)
symptoms	One hot encoding was performed for all the symptom codes creating new columns

Variable selection:

Chi square tests: The variables *smoker, family_history* and *side* were removed as they were not significant and highly skewed.

ANOVA tests: The variables *BMI, tumor_change* and *psa_change* were highly significant in ANOVA tests and considered in the model.

Predictive model: Logistic Regression

Three models were built based on the important variables in the tests and by using forward step wise regression. False negative rate (1 - sensitivity) FN/(TP + FN) is an important factor in

this dataset as incorrectly classifying a non-surviving patient is more costly than for a surviving patient. The final model was chosen based on accuracy and False negative rate. The dataset was divided into Training and Validation Datasets with a random selecting of ratio 70%-30%. Below are some of the most significant variables and their odds ratio.

	Coefficient	Odds Ratio
multi_thrpy1	-0.276035	0.758
BMI	-0.022577	0.977
Gleason_Critical-Low	0.473813	1.606
S10 1	-0.437445	0.6456

Interpretations:

1) The chances of survival of a patient undergoing multi therapies is 25% less compared to a patient who is does not undergo multi therapies.

2) A unit increase in the BMI will reduce the likelihood of survival after 7 years by 3%

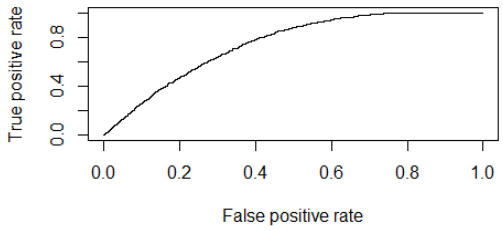
3) The odds of survival of a patient with low gleason score is 1.6 times more than a patient with high gleason_score

4) The chances of survival of a patient having S10 symptom is 35% less than a patient without S10.

Model performance: testing data

Model obtained an overall accuracy of **67.23%** on validation data set and a sensitivity of 0.6251 with an AUC (Area under curve) value of 0.746,

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1797  703
         1  744 1172

               Accuracy : 0.6723
                 95% CI : (0.6583, 0.6862)
    No Information Rate : 0.5754
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.3313
 Mcnemar's Test P-Value : 0.293

            Sensitivity : 0.6251
            Specificity : 0.7072
         Pos Pred Value : 0.6117
         Neg Pred Value : 0.7188
             Prevalence : 0.4246
         Detection Rate : 0.2654
   Detection Prevalence : 0.4339
      Balanced Accuracy : 0.6661

       'Positive' Class : 1
```



Model predictions on score data:

The model has predicted the score data : 4730 as patients who survived (flag =1) and 6801 as patients who did not survive (flag=0).