With the common prostate cancer diagnostic among men, the goal of this analysis is to find the variables that could predict the chances of a tumor has penetrated of prostatic capsule. A dataset was provided by the Ohio State University Comprehensive Cancer Center to analyze how demographic and screening results could cause the likelihood of penetration. A tool was built using multiple logistic regressions to predict the chances of a patient getting diagnosed with prostate cancer using three variables results of digital rectal exam, Prostatic Specific Antigen value, and total Gleason score. This model was proved to be the best model with an AUC of \Sexpr{round(test\_auc, 2)[1][[1]]}, an accuracy of \Sexpr{round(accuracy/100, 2)}, Sensitivity of 0.75, and Specificity of 0.79. Some interesting findings includes a 269\% increase in chances of having prostate cancer when a patient has a unilobar nodule from the left on their results of digital rectal exam compared to a patient with no nodule. Furthermore, patients who have bilobar nodule results after their digital rectal exam will have higher chance of having prostate cancer compared to patients with no nodule, 260\% more to be exact. In addition, patients could gain 133\% more likelihood of being diagnosed with prostate cancer when they have a higher Gleason score. Knowing these three variables (results of digital rectal exam, Prostatic Specific Antigen value, and total Gleason score) are significant factors of the likelihood of penetration, health care professionals would be more cautious in treating each patient. In addition, patients could detect their prostate cancer stages earlier and seek help accordingly.

Prostate cancer is one of the most common cancers among men. The survival rate is vary depending on the stages of the cancer. Prostate cancer has a high probability of being treated if detected at early stages. There are many important characteristics that can predict whether a tumor has penetrated the prostatic capsule. With the results of penetration rate, doctors and nurses would prioritize the patients with the high-risk level. On the other hand, if the patients have low rate in penetration, health care workers could find the best treatments for them to slow down prostate cancer growth rate or even treat the patients. In other words, hospitals can plan their budgets and supplies to focus more on high risk patients and minimize supplies to low risk patients. Knowing the penetration rates does not only benefits health professionals, but it would also assist male patients to detect if their tumors have penetrated of prostatic capsule. This way, patients would seek professional help at an earlier stage to increase their survival rate. Many questions have proposed in favor of these issues including 1) What variables cause the penetration rate to increase or decrease? 2) What is the likelihood of penetration rate? 3) How accurate is the model proposed in this analysis? 4) How can we implement the model? and 5) How can we improve the current model? This report will answer all of the questions mentioned above by going through all crucial part of the data analysis processes such as data exploratory analysis process to explore the relationship of each variable, model development using multiple logistic regression to build the models, and diagnostics process to examine the quality of the model. This model will serve the purpose of predicting penetration rate for prostate cancer in men to minimize human errors and provide an effective tool for health care workers.

A subset of data was provided from the Ohio State University Comprehensive Cancer Center to find variables to predict if a tumor has penetrated the prostatic capsule. This dataset contains a total of 380 subjects, out of which, 153 subjects had a tumor that penetrated the capsule. There are three observations contain missing values will be removed. After eliminating missing values to build more quality models, the dataset now has a total of 377 observations. These data contains one identification column with numbers of each subject, two demographic columns including age and race of the subjects, and four health measurements such as the results of digital rectal exam, detection of capsular involvement, Prostatic Specific Antigen value, and total Gleason score. Health measurement variables will be explained in detail in the exploratory data analysis section. The goal of this analysis is to find the average tumor penetration of prostatic capsule rate. To accomplish this goal, we will first analyze the relationship between each predictor and target variable, select models using bootstrapping method, consider interaction terms to be in the final model, check for the quality of the model, and conclude with an interpretation of the model outputs. All analysis will be done in R Studio with R version 3.6.2.

Before selecting variables for our model development process, it is important to access the relationship of each response variable with the target variable. Table~A\ref{descrips} shows descriptive statistics of each independent variable. As seen in this table, the maximum value of Prostatic Specific Antigen is relatively high compared to the median and mean values suggesting there might be outliers. In addition, there are disproportion weights of each factor in race of the subjects and detection of capsular involvement. This indicates that there might be biased in predicting the outcome if we use these two variables in the models. The rest of the variables seems proportionally balanced and no skewness were detected among them. In terms of the tumor penetration of prostatic capsule, the response variable has a balance proportion of 40\% of tumor has penetrated the prostatic capsule out of 377 subjects. Figure~\ref{explore1}.1 show the histogram of the chances of the tumor penetrated of prostatic capsule (1 = penetration, 0 = no penetration).

Histogram of 1 (Left): Tumor penetration of prostatic capsule (1 = penetration) and 2 (Right): Results of Digital Rectal Exam (1 = no nodule, 2 = unilobar nodule left, 3 = unilobar nodule right, 4 = bilobar nodule) with average penetration rate in black.

The digital rectal exam is a simple exam to examine the prostate to determine the size of the prostate and any abnormalities. Clearly, different abnormalities would have different chances of penetration of prostatic capsule. Figure~\ref{explore1}.2 shows four categories of the results of digital rectal exam including no nodule, unilobar nodule (left), unilobar nodule (right), and bilobar nodule noted using number from 1 to 4, respectively. Here, this variable is identified as categorical variables with four levels instead of a continuous variable. As seen in plot, these four categories obviously have different average penetration rates. For example, a subject with no nodule would be less likely to penetrate compare to a person with bilobar nodule. It makes sense because if there is no nodule detected in the prostate area, then there is less chances of the tumor will penetrate the prostatic capsule. In addition, Figure~\ref{explore1}.2 also depicts the number of subjects in each category. Unilobar nodule (left) is the most common and bilobar nodule is the least common among prostate cancer. However, the proportion of each factor for digital rectal exam are not too different. Thus, this variable would provide significant impact for the model.

Detection of capsular involvement can describe if the tumor involving or extending beyond the prostate capsule. The tumor is involved when there is capsular involvement present. Noticeably, the patients who have capsular involvement present can be riskier than patients who do not have capsular involvement. Figure~\ref{explore2}.1 shows the count and proportion of penetration for the detection of capsular involvement. This variable has two levels (1 = no, 2 = yes) to suggest if a subject has capsular involvement or not. As seen in the figure, patients with capsule involvement would be much more likely to have tumor penetrated of prostatic capsule compared to patients that do not have capsule involvement. However, the count for each level is not proportional. Majority of the subjects have level 1 with no capsular involvement. This variable could be a good variable for our model since it shows a strong signal. However, caution about the disproportional needs to be considered before finalizing the model.

Plots of each variable with average penetration rate in black of 1 (Left): Detection of capsular involvement along with the frequency of each factor, 2 (Center): Prostatic Specific Antigen value with mg/ml unit by decile, and 3 (Right): Total Gleason score along with its histogram.

Prostatic Specific Antigen (PSA) can be found in the prostate gland cells. This screening test measures the amount of prostate-specific antigen in patients' blood using milligram per milliliter (mg/ml) of blood. Higher PSA level could lead to a higher possibility of diagnosing with prostate cancer. Figure~\ref{explore2}.2 shows a plot of the relationship between PSA and the average penetration of prostatic capsule rate. Here, the data point is being sorted from lowest to highest, divided into 10 equal buckets, and assigned a decide for visualization purposes. As one can see, as the decile number increases by one unit, the chances of penetration rate increases, on average. Since PSA in higher buckets are generally the largest, they have higher risk of getting prostate cancer than PSA in lower buckets. Therefore, this proves that PSA could provide a significant relationship for the likelihood of tumor penetrated of prostatic capsule.

Last but not least, total Gleason score can measure the abnormality of the cells to see how the cells are arranged using a scale from 1 to 10. Higher scores lead to a higher possibility of penetration in prostatic capsule. On the other hand, cancer cells that looks similar to normal cells can be considered as low risk and have lower scores. Figure~\ref{epxplore2}.3 illustrates a histogram of total Gleason score overlayed by the average risk of having cancer. Total Gleason score seems to have a normal distribution with less subjects having low or high scores and most subjects having median score of 6. In addition, the average penetration rate is increasing as the total Gleason score is higher. This mean that if a patient's cancer cells look similar to normal cells, they have less chances of being diagnosed with prostate cancer. Nevertheless, if the Gleason scores are high, the patients would have high risk of being diagnosed with prostate cancer. Hence, this variable can also provide excellent signal for the response variable. Other variables did not mention above have little to no relationship with tumor penetration including age and race of the subjects.

After accessing the relationship of predictors and response variable, it is important to verify the correlation between predictors. Fortunately, all correlation between predictors are less than 40\%. Total Gleason score and PSA values have the highest correlation score of 38.60\% which is considered as low. This suggests that all variables can be in the same models without raising the multicollinearity issues. However, to make sure, VIF scores will confirm if there are multicollinearity issues after the modeling process. Therefore, we will go ahead and build our model using four variables including results of digital rectal exam ("dpros"), detection of capsule involvement ("dcaps"), PSA value ("psa"), and total Gleason score ("gleason").

After four variables were discovered to be significant, logistic regressions will be used in model development process. Bootstrapping method is a common method for variables selection using random partitions. Here, the models are generated using a combination of each variable using different random partition of training and testing data. An average AUC and AIC of each model was computed after to compare. Table~A\ref{reg\_bootstrap} demonstrates the top 10 models with the highest AUC scores. Models with three variables including results of digital rectal exam, PSA values, and total Gleason scores seems to be the best model with overall AUC above 80\% and low AIC scores compared to other models. Therefore, the model development process will continue using the three variables mentioned previously with a train set of 70\% and hold-out test set of 30\%.

Before finalizing the model, all combinations of interaction terms were added to the model with three variables including results of digital rectal exam, PSA value, and total Gleason score. The results of this model are shown in Table~A\ref{reg\_summary\_int}. The coefficients of all three variables seem to increase in the model with interaction terms. This proves that the model with interaction terms’ coefficient has steeper slopes than the model without interaction terms (see Table~\ref{reg\_summary\_final}). However, p-values for all interaction terms are quite high suggesting that interaction terms are not needed in the model. To further prove this argument, a stepwise model selection was run using AIC as a validation metric. The best model appears to be the model with the original three variables: results of digital rectal exam, PSA value, and total Gleason score. As seen in Table~\ref{reg\_summary\_final}, the final model's coefficients are positive consistent with the exploratory analysis. In addition, all variables in this model have low p-values of less than 0.05 and odd ratios’ confident intervals do not contain 1 suggesting all variables are significant. Lastly, this model has a low AIC of \Sexpr{round(summary(fit\_final)$aic, 2)}, high AUC of \Sexpr{round(test\_auc, 2)[1][[1]]}, and high accuracy of \Sexpr{round(accuracy/100, 2)} consistent with high probability of producing precise predictions.

Summary regression of the final model with three variables: results of digital rectal exam (dpros), PSA value (psa), and total Gleason score (gleason). The columns represent the feature names, coefficient estimated values, standard errors, test statistics, significant p-values, odd ratios, and 95 percent lower and upper confident interval of odd ratios, respectively from left to right.

To access how good the model fit, predictions were generated using the hold-out test set. Figure~\ref{model\_plot\_1}.1 depicts a lift chart of empirical and indicated likelihood of penetration of prostate capsule. Here, the predicted proportion were sorted from smallest to largest, divided into 10 equal buckets, and assigned a decile number from 1 to 10. The lift chart illustrates that the average predicted proportion of each bucket are similar to the average empirical values. This implies that the model is predicting precisely.

1 (Left): Lift chart of empirical (black) and indicated (blue) penetration rate sorted from smallest to largest and assigned by decile, and 2 (Right): Cook's Distance plot with labeled outlier observations in blue.

Examining the quality of the model is one of the most important steps in a data analysis. First, multicollinearity would not be an issue in this analysis since VIF scores are smaller than 2 for all the variables in the model. Secondly, Figure~\ref{model\_plot\_1}.2 shows the Cook's distance of each observation point. A few observations did not meet the cutoff point of \Sexpr{round(cutoff, 2)} including rows 4, 7, 87, 121, 157, 196, 239, 254, 267, 272, 289, 307, and 365. These points influence all fitted values. In other words, these subjects with row number mentioned previously have huge impacts on the chances of getting penetrated of prostate capsule.

The studentized residuals are also very essential to validate while accessing the quality of the model. Figure~\ref{model\_plot\_2}.1 portrays a residuals plot of the model. The plot is color coded by two section, one for underestimate points and one for overestimate subjects. The number of overestimated values and overestimated values are approximately the same which constructed a mean value for residuals of approximately 0, or \Sexpr{round(mean(diag$.std.resid), 2)} to be exact. Furthermore, there are no patterns to be detected in the residuals consistent with constant variance. Most residuals are between -2 and 2 which is ideal for all residual values. However, there are a few points that evidently have large residuals including rows 87, 239, and 289. This indicates that these three subjects are outliers.

1 (Left): Studentized residuals plot with blue color are positive residuals (overestimate values) and black are negative residuals (underestimate values), and 2 (Right): Diagonal hat matrix plot with influential points colored in blue.

Last of all, large hat values show potential outlying observations with respect to each predictor. As seen in Figure~\ref{model\_plot\_2}, rows 115, 215, 185, and 335 have higher hat values than others. These subjects potentially have high/low values consisted in one or more of their predictors. To further access the outliers and influential points, Table~A\ref{outlier\_obs} was constructed to show the details of each observation that were detected as outliers or influential points. Most subjects noticeably have really high PSA scores compared to the mean PSA scores of \Sexpr{round(mean(prostate$psa), 2)}. However, all subjects were decided to be kept in the model since there are no valid explanations to eliminate them from the model.

After completing model development and validation, the model is ready for interpretation. All interpretation values used in this section are from Table~\ref{reg\_summary\_final}. The results of digital exam have great impact on penetration. For instant, a patient with a unilobar nodule (left) would be 2.28 times or 128\% riskier than a patient without a nodule. A person with a unilobar nodule (right) would be 3.69 times or 269\% more likely to have their tumor penetrated with prostate capsule. Lastly, a person with bilobar nodule would be more likely to be diagnosed with prostate cancer 3.60 times higher compared to patients with no nodule. In addition, if a patient’s PSA score increases by 1 mg/ml of blood, on average, the patient will have 4\% increase in the risk of having a prostate cancer. Last but not least, having a low Gleason score would decrease the chances of having cancer since an increase of 1-unit in Gleason score would cause the patient to be 2.33 times riskier.

Generally, there are protocols for doctors and nurses to follow when it comes to important diagnosis like prostate cancer. Manually, it is quite difficult to avoid human errors when providing a diagnosis to a patient. With this model, health professionals can be able to minimize predicting errors and generate the likelihood of a male patients getting prostate cancer more efficiently. The three main factors that change the chance of a tumor has penetrated of prostate capsule are the results of digital rectal exam, PSA score, and total Gleason score. Penetration rates are likely to increase as the results of digital rectal exam are in a critical nodule, PSA values are higher, and total Gleason scores are larger. On the other hand, having no nodule in the results of digital rectal exam, low PSA scores, and low Gleason scores would cause the likelihood of penetration to be very low. This leads to lower risk of having prostate cancer. With an AUC of \Sexpr{round(test\_auc, 2)[1][[1]]} and accuracy of \Sexpr{round(accuracy/100, 2)}, we can consider this model to be the best model that can provide the likelihood of penetration rate in male patients. There are many limitations for this model. First, the sample size of this dataset is relatively small. The subset of this small sample size could cause the results to be not as accurate as the full dataset. If we have access to the full dataset from the Ohio State University Comprehensive Cancer Center, then our results might be different. In addition, there were only six predators provided to use in the model. If there are more variables, the model could provide better results. For example, we can consider genetic as a potential a factor that cause the risk of having prostate cancer in the family of the patient to improve the current model.

Iteration log of model bootstrapping process. Showing the top 10 models that have the highest AUC scores. The columns represent random partition for train set, number of rows in the training data, number of rows in the testing data, predictors, AIC scores, and AUC scores, respectively from left to right.

Summary regression of the final model with all combination of interaction terms. The columns represent the variables, coefficient estimated values, standard errors, test statistics, and significant p-values, respectively from left to right.

Table of all potential outlier and influential values. The columns represent the tumor penetration of prostatic capsule, results of digital rectal exam, PSA scores, total Gleason scores, and row number of each subject, respectively from left to right.