Over the course of this quarter, I developed a strong understanding of the fundamentals of statistical learning, exploring a wide range of techniques used for modeling and predictions. We started off with first learning about the different types of model, the difference between regression and classification at a very basic level. Through the reading, I explored the differences between supervised and unsupervised learning, and the tradeoffs between bias and variance when creating models.

The next chapters of the reading built upon this foundation and explored more advanced classification and clustering techniques such as K-means clustering and the Bayes classifier. I learned how both methods worked and also learned how to implement the programs using Python. Next, I explored the use of matrices in machine learning, and different regularization techniques such as Lasso and Ridge regression.

Once I had a basic understanding of different models, how to evaluate the models, and how to implement the models in Python, I began working my way through cleaning the data, while also learning the Python syntax that I needed from the textbook. The reading covered key python functions such as .iloc, lambda functions, as well as the zip function. The readings also touched on different ways to evaluate the predictions of models using statistical methods such as T statistics, P values, Residual standard errors, and R^2 statistics to name a few.

Using these concepts proved vital in the final project, as I had to show how my classification model was performing against the actual results. Later in the readings, I learned about more complex model architecture like Generative models for classification and linear discriminant analysis, which I did not use in my final project, however, I found them interesting to learn about. The most important part in the reading to my final project was learning more about how boosting models worked behind the scenes. Up until this point, I had used boosted models before but had a vague understanding of what was actually going on. Through reading and hands-on application of the XGBoost model to my dataset, I gained valuable experience in how to effectively use the model, understand its strengths and limitations, and fine-tune its parameters to maximize performance.

For my final project, I developed a tackle probability model to calculate the odds of a tackle occurring based on the location and movement of a defensive player. The data was sourced from the NFL Big Data Bowl, containing over 5 million frames of player tracking data from real games.

To make the problem more focused, I filtered the data to include only frames where a defender was within one yard of the ball carrier—situations in which a tackle was likely to occur. I trained an XGBoost classifier on these instances to predict whether a tackle would happen in a given frame.

This project allowed me to apply everything I had learned: data cleaning, feature engineering, model implementation, evaluation metrics, and model tuning. I compared the model's predictions with actual outcomes to evaluate performance and adjusted hyperparameters to improve accuracy. Through this process, I learned how to handle real-world sports data and build a predictive model that could eventually help teams evaluate defender efficiency and positioning.