**Step-by-Step Process:**

Our team determined that we would like to attempt to predict the final outcomes of National Football League (NFL) games using a machine learning program to determine if we could develop a system that predicts games better than the traditional gambling lines. This was done, as gambling lines are generally made to get even money (this does not mean 50%) on each side of the line to ensure profitability for the casino and not to predict final outcomes, although one would deduce that these ideas are highly correlated.

Once we chose this project, we wanted to set a time frame for data usage. Given the vast changes in playing style between the modern NFL and its historical style, we decided 2015 would be a good end point for our data collection, as previous data may not accurately reflect the offensive and defensive systems currently in place in the NFL. After making this determination, we began searching for available information on historical NFL lines and over/under totals for games. We located this at Australian Sports Betting (<http://www.aussportsbetting.com/data/historical-nfl-results-and-odds-data/>), where they have data dating back to the 2006 season. After retrieving this data, we clean the files to include only the date of the game, home team, away team, home score, away score, home line open, and total score open. We used the opening lines because most professional gamblers believe the open lines are where they can make their money and that if the money moves away from what they believe will happen (a team they will bet at -3, moves to -2) they will bet again later in the week too. After this was done, we added a year and week column to the data and removed the date. This change allowed us to merge our data easier with the Defense-adjusted Value Over Average (DVOA) data.

The DVOA data was developed by Football Outsiders and is based on the evaluation of every single NFL play. This analysis compares a team's performance to a league baseline based on unique situations occurring throughout the game in order to determine a value over the league average. It is best summarized as a measure of a team’s efficiency by comparing success on every single play to a league average based on situation and opponent. The DVOA data contains how a specific team performs as a rushing offense, passing offense, rushing defense, passing defense, and special teams. Teams also have a total DVOA and a total offense and total defense DVOA. We removed these columns from consideration, as they are incorporated in the individual passing and rushing numbers and we didn’t want them to unduly influence the model. While we attempted to scrape this data set for the web, we were presented with some challenges given the paywall and cookies from the source domain (We did pay for a subscription to access the data). Consequently, the data elements were retrieved manually and reviewed for accuracy.  Once the dataset was extracted from the source it was passed through some modest data cleaning when input into a CSV file.

Once the two data sources were compiled and cleaned, they were loaded into a data warehouse utilizing AWS/Postgres. The two tables were then joined based on the specific game matchups that had been passed through a query and were inserted into a table for extraction via pandas. This merged database was used to make our predictions. Our predictions are based on five (5) models made using the DVOA archive from FootballOutsiders and the historical odds data.

* The first model is a linear regression model that predicts the point differential of this week’s matchups. Predictions for this model are made from the perspective of the home team. For example, a prediction of +7 means the home team will win by a touchdown, while a -7 represents a loss by a touchdown. Model Score Differential: r2: .76, RSS: 43.51
* The second model is a linear regression model that predicts the total number of points scored by both teams in the game. This helps determine whether to bet the over or the under when compared to what the odds makers believe the total point score will be. Model Score Over/Under: r2: .33, RSS: 128.75
* The third and fourth models are both support-vector machines.
  + The first of these predicts whether the home team will win. Model Score Winner: r2: .85
  + The other predicts whether to bet the home or away team based on the current spread. Model Score Spread: f1: .81SVC
* Our final model is a support-vector clustering model used to predict what we believe the point differential based on how the two teams have performed offensively and defensively this year. This model groups the differential to a common football differential, like a field goal or a touchdown. Model Score Differential: r2: .57, RSS: 64.88 3.

Once the modeling process was completed and values extracted utilizing the most recent data sets, a predictions table was constructed and warehoused within the database, as well.

To create the tables and visuals, Tableau reads from the database server connection and joins the data based on the Year, Week, and specific team match up.

To keep our information up-to-date, we add the previous week’s game data to our database along with the current odds information. Once we do this, we plug next week’s games into the model using the average DVOA for the teams that are competing that week along with the other data. We then retrain the models, with all of the previous data, using the most current data sets and make the predictions for next week’s games. This is an iterative process that requires constant updating.