Comp Sci 349 Final Project Proposal

Vatsal Bhargava, Coel Morcott, Will Pattie, Aidan Villasenor, Alex Romanenko

Fall '23

What task will you address, and why is it interesting?:

We are going to predict whether or not an NFL team will cover their spread in a given playoff game. This is an exciting problem for us because beating the bookmakers is a challenge that we would love to succeed at. We are also very big fans in the NFL, and making this model is interesting to us.

How will you acquire your data?:

ESPN has a website with all historical season average stats. It contains season average stats for each NFL team on both offense and defense. We can easily convert this data into a csv file which we can use to predict the scores of each game.

Team Offensive/Defense/Special Teams/Turnover Stats NflFastR Python Spread Data

Which features/attributes will you use for your task?

For our features and attributes, we will look at a plethora of team stats. The team stat attributes are: games played, total yards, passing yards, rushing yards, total points, total yards against, total passing yards against, total rushing yards against, total points against, interceptions, fumbles, interception takeaways, fumble takeaways, Vegas lines for team totals and spreads, kickoff and punt averages.

What will your initial approach be? What data pre-processing will you do, which machine learning techniques (decision trees, KNN, K-Means, Gaussian mixture models, etc.) will you use, and how will you evaluate your success (Note: you must use a quantitative metric)? Generally you will likely use mean-squared error for regression tasks and precision-recall for classification tasks. Think about how you will organize your model outputs to calculate these metrics:

We will do preprocessing to structure our data so that each row represents an NFL matchup and the columns represent the stat categories for each team, or our attributes. The class would be whether or not team1 covered during a specific game. We will also be using random forests to classify whether or not team1 covers the spread. For this, we will need to change our ID3 Random Forest model to allow for splits based on non-discrete values. For example a split may be based on if passing yards of team1 is >= 300 yards versus <300 yards.

In addition to our random forest, we will be using a Neural Net for the same task to compare our models and see which makes better predictions..

We will be trying to minimize the misclassification for both of our models, given that our output is a binary Cover or No Cover, this is the most suitable metric to minimize.