**Project Proposal**

MET CS 699 – Data Mining

Oren Shapira

Michael DeMers

**Purpose of Project**: Our project would like to assess how weather impacts the performance of players in NFL games. Based on the combination of temperature, humidity, and wind (mph), we would like to be able to predict how many passing yards a Quarterback would throw in a given game. This prediction would only be feasible for weather conditions that are deemed to have a statistically significant impact on Quarterback performance (i.e. we don’t expect a difference in performance if the weather is 60 vs. 70 degrees F). While weather is our main predictor of interest, we will also account for other covariates that can influence passing yards such as a Quarterback’s passing performance in prior games, and the strength of their opponents Defense.

If time allots, we plan to explore how weather conditions impact other performance metrics of a Quarterback, such as pass completion percentage, average yards per completion, interceptions, or touchdowns thrown in a game. We also may explore the prediction of other team/player metrics, such as total team points scored, rushing yards, or field goal attempts made.

**Data Set Descriptions:** We will be using four different datasets for this project. These datasets may be subject to change depending on any preprocessing obstacles we encounter.

1. **Weather\_2013\_1231.csv:**
   1. Summary: This dataset contains weather information for every NFL game between 1960-2013. It has 11 attributes, and 11,192 tuples.
   2. Source: http://www.nflsavant.com/about.php
   3. Attributes: The ones in bold are particularly relevant to our project
      1. Id
      2. **Home\_team:**
      3. Home\_score
      4. Away\_team
      5. Away\_score
      6. **Temperature**
      7. Wind\_chill: Likely won’t use this, since it’s only available for some games
      8. **Humidity**
      9. **Wind\_mph**
      10. Weather: A concatenation of the other weather columns
      11. **Date**
2. **Game\_Logs\_Quarterback.csv**
   1. Summary: This dataset contains passing statistics for every Quarterback, for NFL game between 2000 – 2016. This includes total passing yards, pass attempts, Touchdowns, Interceptions, etc. This will be joined with the weather dataset above on overlapping years. This dataset has 29 attributes and 40,247 rows.
   2. Source: <https://www.kaggle.com/kendallgillies/nflstatistics>
   3. Attributes: The ones in bold are particularly relevant to our project
      1. Player Id: Unique Identifier
      2. **Name**: Name of Quarterback
      3. Position: Player position
      4. **Year**: Year of game. Will be joined with weather dataset for overlapping years
      5. Season: Distinguishes preseason vs. regular vs. postseason
      6. Week: Season week of game
      7. **Game Date**: Will be joined with weather dataset
      8. **Home or Away:** This and Opponent will determine where the game was played
      9. **Opponent:**
      10. Outcome: Win or Loss
      11. Score
      12. Games Played
      13. Games Started
      14. **Passes Completed**
      15. **Passes Attempted**
      16. **Completion Percentage**
      17. **Passing Yards:** This is the attribute we would like to predict
      18. **Passing Yards** Per Attempt: We may look at this if time allots
      19. TD Passes
      20. Ints
      21. Sacks
      22. Sacked Yards Lost
      23. Passer Rating: Numeric rating of QB’s performance in the game
      24. Rushing Attempts
      25. Rushing Yards
      26. Yards Per Carry
      27. Rushing TDs
      28. Fumbles
      29. Fumbles Lost
3. **Career\_Stats\_Passing.csv**
   1. Summary: This dataset contains every NFL player’s overall passing stats for the entire season, and spans between 1924-2016. It will be used primarily to normalize across Quarterbacks that have different averages of passing yards per game. There are 23 attributes and 8,525 rows.
   2. Source: <https://www.kaggle.com/kendallgillies/nflstatistics>
   3. Attributes: The ones in bold are particularly relevant to our project
      1. Player Id
      2. **Name**
      3. **Position:** We would only filter on the position of Quarterbacks
      4. **Year**: NFL season year. Will be merged with other datasets on overlapping years
      5. **Team**: The team the Quarterback plays for
      6. Games Played
      7. **Passes Attempted**
      8. **Passes Completed**
      9. **Completion Percentage**
      10. **Pass Attempts Per Game**
      11. **Passing Yards**
      12. **Passing Yards Per Attempt**
      13. **Passing Yards Per Game:** This will be the key metric used to normalize for the impact of weather on passing ability
      14. TD Passes
      15. Percentage of TDs per Attempts
      16. Ints
      17. Int Rate
      18. Longest Pass
      19. **Passes Longer than 20 Yards**
      20. **Passes Longer than 40 Yards**
      21. Sacks
      22. Sacked Yards Lost
      23. Passer Rating
4. **Sportsref\_download\_2013.xls**
   1. Summary: This dataset contains team defensive stats for an entire NFL season. This will be used to normalize the passing yards that can be expected against a given opponent. A Quarterback that plays an opponent with a strong Defense against passing is likely expected to throw for fewer yards in that given game. The dataset we have submitted is for NFL season 2013. We will have to manually download datasets for each NFL season of interest (likely 2000 – 2013). There are 28 attributes and 32 tuples.
   2. Source: <https://www.pro-football-reference.com/years/2013/opp.htm>
   3. Attributes:
      1. Rk: Defensive rank
      2. **Tm: Team name**
      3. G: Games
      4. PF
      5. Yds
      6. Ply
      7. Y/P
      8. TO
      9. FL
      10. 1stD
      11. **Cmp: Completed passes against defense**
      12. **Att: Attempted passes against defense**
      13. **Yds: Passing yards against defense**
      14. TD: Passing Touchdowns against Defense
      15. Int: Interceptions
      16. NY/A
      17. 1stD
      18. Att
      19. Yds: Rushing yards against defense
      20. TD: Rushing Touchdowns against defense
      21. Y/A
      22. 1stD
      23. Pen
      24. Yds
      25. 1stPy
      26. Sc%
      27. TO%
      28. EXP