**Data Analytics Capstone Topic Approval Form**

**Student Name:** Jonathon Fryman

**Student ID:** 000974544

**Capstone Project Name:** NFL Betting Model – Comparing historical results with betting lines to develop prediction model

**Project Topic**: As increased legalization of sports betting has spread to states throughout the country, a new segment of the population is being introduced to the concepts of favorites, game spreads, and over/under for the first time. In this analysis, historical game data will be reviewed to compare the pre-game betting lines and the actual results to develop a prediction model to help determine the best games for bettors to wager their bets.

**This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** Can a model be developed that provides insight and predictions into which bets historically have been the safest and most reliable.

**Hypothesis**:   
**Null hypothesis**- The predictive model is not able to sufficiently predict the results of the historical games.   
  
**Alternate Hypothesis**- The predictive model is able to sufficiently predict the results of the historical games.

**Context:** *Explain why the situation or question would benefit from a data analysis in less than 500 words.*The historical games dataset contains data on 13,516 NFL games. This data contains games dating from 1966 that contains both the regular season and post season games.   
  
The majority of records within the existing data contains information related to the original spread favorite and over/under data for nearly 11,000 of the records.

**Data:** *Identify data you will need to collect that is relevant to the situation or question.*Click here to enter text.

The dataset to be used for this analysis is publicly available at <https://www.kaggle.com/datasets/tobycrabtree/nfl-scores-and-betting-data>  
  
The full dataset to be used consists of three unique .csv files. The files contained within the data set are ‘nfl\_stadiums.csv’, ‘nfl\_teams.csv’, and ‘spreadspoke\_scores.csv’  
  
The ’nfl\_stadiums.csv’ file contains information related to the stadiums the included games were played in. This includes information such as the stadium’s name, geographic location, years of operation, if the stadium is indoor or outdoor, weather type, seating capacity, and the playing surface.  
  
The ’nfl\_teams.csv’ file contains information related to the full teams name, mascot, team abbreviation for both the city and mascot, conference, and division.   
  
The spreadspoke\_score.csv’ file contains information related to the individually played games. This includes the home/away team, home/away score, date of game, the betting spread and favorite, the betting over/under, the stadium the game was played in, if the game was played on a neutral field, and weather conditions at time of the game.   
  
The data being used for this analysis is provided under a [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/) license that allows users to share and adapt the data with proper credit given to the original data provider.

**Data Gathering:** The data-gathering methodology to be used for this analysis is documents and records. This methodology consists of examining existing data. For this specific analysis, this includes examining existing records related to historical game data and the associated historical betting lines for the recorded game data.

**Data Analytics Tools and Techniques**:For this report, the primary data analytic technique to be used is factor analysis. Factor analysis is a technique used to reduce many variables into a smaller number of factors. This technique extracts the maximum common variance from all variables and puts them into a common score. This will include the principal component analysis method which starts by extracting maximum variance and puts them into the first factor. At this point, the variance explained by the first factor is removed and the maximum variance for the second factor is extracted. This process continues for all remaining factors.

**Justification of Tools/Techniques:**

The planned tools to be used for this analysis include the Python programming language, Jupyter Notebooks, and Tableau.

Jupyter Notebooks will be utilized for any required data cleaning, data analysis, data visualization, and model development for use with generating predictions and comparing against the historical game dataset. This will be achieved using the Python programming language and its many available packages and libraries that facilitate data analysis and model development.

Additionally, Tableau is planned to be utilized to create additional data visualizations for use within the written report associated with this project.

**Project Outcomes**: The anticipated outcomes for this project are to determine if a model can be developed that is better at determining the winner of historical games in comparison to the pregame betting indicators, such as favorite, point spread, and the over/under line.

**Projected Project End Date**: 3/30/2023

**Sources**:

1. Spreadspoke. (2023, February 14). *NFL scores and Betting Data*. Kaggle. Retrieved February 17, 2023, from <https://www.kaggle.com/datasets/tobycrabtree/nfl-scores-and-betting-data>
2. *Creative Commons License Deed*. Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International - CC BY-NC-SA 4.0. (n.d.). Retrieved February 17, 2023, from <https://creativecommons.org/licenses/by-nc-sa/4.0/>

**Course Instructor Signature/Date:**

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Instructor’s Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.