Data Management and Data Analytics Capstone Topic Approval Form

Capstone Topic Approval Form

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of the following areas so you will have a complete and realistic overview of your project. Your course instructor cannot approve your project topic without this information.

Student Name: Logan Donlan

Student ID: 002470457

Capstone Project Name: NBA Statistical Analysis

Project Topic: Analyzing NBA statistics from 2000 - 2023 to look for correlations to winning in the regular season.

Research Question: What statistics or group of statistics correlate to winning in the regular season, and do those change with the evolution of play throughout the years?

> **Hypothesis:** Null Hypothesis: There is no correlation to any one statistic or group of statistics over the 24 season period that equates to winning in the regular season. Alternative Hypothesis: There will be a correlated grouping of statistics that correlate to winning in the regular season and these will change with the evolution of play style through the years.

Context: An analysis of this dataset will benefit multiple organizations. First, each NBA team is always looking for an edge. Seeing what the winning teams are doing well compared to others will give them insight into possible scheme or personnel changes to better fit the mold of the current top NBA teams. Secondly, being able to predict winners and losers based on statistical analysis will help sports gambling organizations more accurately depict lines for all kinds of bets. Lastly, being able to see these trends will help the gambler make more informed decisions about which bets to take and when to look elsewhere. The statistical data analysis on this dataset will be an integral part of decision making for multiple facets of the sports world.

Data: I will collect data that contains various statistics from all NBA teams across the regular season spanning from 2000 to 2023.

> The dataset I will be using is called NBA Team Stats and can be found for download on Kaggle.com following the link https://www.kaggle.com/datasets/mharvnek/nba-team-stats-00-to-18?resource=download. The dataset consists of 29 columns and 716 rows with all of the data needed to answer my question.
> This dataset is provided by Kaggle and is accessible as unrestricted data to Kaggle and users. It is downloadable as a zip file directly from the site page.

Data Gathering: I will be using Jupyter notebooks to import the data as a CSV file with the pandas library. The cleaning and analysis of the data will be done in a Jupyter notebook, as well.

Data Analytics Tools and Techniques:

Tools: Python will be the primary language used to clean and analyze the data. I will be utilizing the pandas, matplotlib, numpy, seaborn, and sklearn libraries, along with Tableau to perform the statistical analysis and create visualizations to display findings. I will be creating a supervised learning model to determine which statistics have the most significant impact on the outcome of games.

Techniques: I will be using the pandas library in the Jupyter notebook to import and clean the dataset. The numpy library will be used to perform some statistical analysis. The matplotlib and seaborn libraries will allow me to create visualizations to help analyze the data. I will use Lasso Regression to create a model and perform regression analysis on the dataset. Tableau will be used to visualize final findings.

Justification of Tools/Techniques: Pandas and numpy will allow me to import, clean, and manipulate the data efficiently in the Jupyter notebook environment. Matplotlib and seaborn will provide me with flexibility in creating visualizations within the environment. I'm choosing a Lasso Regression model based on its ability to hand multicollinearity and reduce the impact of overfitting. It will also allow me to create a simpler, more interpretable model. Tableau is an industry-leading visualization platform that will allow me to create sleek and digestible visuals to depict the results to an audience.

Application Type, if applicable (select one):

☐ Mobile
☒ Web
☐ Stand-alone

Programming/Development Language(s), if applicable: Python

Operating System(s)/Platform(s), if applicable: Windows 11

Database Management System, if applicable: NA

Project Outcomes: I believe I will find a small group of statistics that correlate with winning on a consistent basis. I also think that these groups will shift slightly based on different year groupings (i.e. 2000-2009 vs. 2010-2020).

Projected Project End Date: 5/24/2024

Sources: Michael H., 2024, NBA Team Stats, https://www.kaggle.com/datasets/mharvnek/nba-team-stats-00-to-18?resource=download

---

Human Subjects or Proprietary Information

Does your project involve the potential use of human subjects? (Y/N): No
Does your project involve the potential use of proprietary company information? (Y/N):No