FINAL PROJECT PROPOSAL

1. Group Members

Our group will consist of three members:

- Harry Nguyen
- Khoi Van
- Duc Bui

2. Problems and Background

Betting on sports, like soccer, is very popular around the world. Many people place bets hoping to win money. When they do this, they use something called betting odds. These odds tell people how likely something is to happen and how much money they could win. Bookmakers are the companies or people who set the odds. They use past data, expert guesses, and what other people are betting on to choose the numbers. In real life, many people lose money on sports betting even when they make smart guesses. For example, during the 2022 FIFA World Cup, lots of fans bet on big teams like Brazil or France to win. The odds looked good, but sometimes these teams lost to underdogs like Morocco. This shows that odds don't always match what really happens.

Also, many websites give slightly different odds for the same match. This shows that the odds are not always based only on true chances, but also on what will make the most money for the betting companies. Bookmakers add a small hidden cost to the odds (called the overround), which helps them profit no matter who wins.

This is a problem because most regular people don't know about this. They believe the odds are fair and based only on data. But if the odds are always slightly unfair, it means bettors are more likely to lose money in the long run, while bookmakers keep winning.

Because of this, we want to ask: Are the betting odds fair? Or are they made in a way that helps the bookmaker and hurts the person betting?

3. Data Sources and Structures

The data for this project comes from Football-Data.co.uk, a free website that shares match results and betting odds for soccer leagues around the world. It includes data from popular leagues like the Premier League, La Liga, Serie A, and more.

The files are available in CSV format. Each file includes one season of matches, with each row showing details of one game. The columns include the names of the teams, the

Time span?

score, the match result (home win, draw, or away win), and betting odds from different bookmakers. Odds are written in decimal format, and there are also stats like shots, fouls, and yellow/red cards. This format makes it easy to check if the odds match the real results and to look for patterns in how bookmakers set their odds.

4. **Questions**

In this project, we want to find out if soccer betting odds are fair or not. Betting odds are supposed to show how likely a team is to win, but sometimes they are changed so that bookmakers can make a profit.

We will examine the betting odds and compare them to the actual match results to see how closely they match. For example, if the odds suggest a team has a high chance of winning but that team loses more often than expected, the odds may not be accurate. If this pattern happens regularly, it could mean that the odds are set in a way that benefits the bookmaker. This would give the bookmaker an advantage, even when bettors make reasonable choices. By comparing implied probabilities with real outcomes, we aim to find out whether the odds are truly fair or designed to favor the house.

5. **Statistical tools**

Implied Probability Calculation

We will convert betting odds into implied probabilities using the formula:

1 / odds.

This shows the chance that the bookmaker believes each team has to win.

Chi-Square Test

This test checks whether the actual match results match the expected results based on the odds. If the difference is too large, the odds may not be fair.

Linear Regression

We use linear regression to see if implied probabilities predict real outcomes. A strong match means the odds reflect reality; a weak match suggests unfairness.

Hypothesis Testing

We will create null and alternative hypotheses to test if the differences between expected and actual outcomes are statistically significant. This helps us decide if the unfairness we see is just by chance or a real pattern.

MATH-DA 220 Final Project Proposal Rubric

- Background - Does the proposal explain a substantive domain area to be studied and motivate the project?

  17 _____ out of 20

- Research questions - Does the proposal clearly describe the key research questions of interest?

  22 _____ out of 25

- Data – Is it clear where data will come from, that the data are available and accessible to use, and that it will be structured conducive to the desired analyses?

  23 _____ out of 25

- Planned statistical analyses – Are the analyses appropriate for the questions to be answered? Are a wide range of methods from class employed?

  22 _____ out of 30