**Predictive Model for Sports Betting on Prize Picks**

[Github Link](https://github.com/Isaiahmt0904/PrizePicks-Predictive-Betting-Model)

1. **Introduction**:

This project is a very simple passion project that aims at manifesting a tangible representation of the data analysis skills I’ve been personally developing for the past couple months. More so, the project's connection to Sports betting illustrates a real world practical application of data analysis skills that would not only look good on a resume, but also connects to my direct interests as well.

The Project can be broken up into 3 parts

1. Collecting data

Using **basketball-reference.com**, I obtained data from each webpage ready to collect based on inputted players first and last name and their past data based on bets. Using the request library in Pandas I can extract the necessary data from the website and parse it using beautiful soup, so that it’s formatted as a dataframe in Python. Afterwards, I cleaned the data and input it into one of my predictive models in order to obtain a desired result. To keep the model simple, I limited the predictions to output over and under on certain stats, based off of a multitude of factors listed in the predictors section.

1. Designing the Predictors

Utilizing Linear Regression and other analytic techniques with 4 different predictors, I create an automated function that takes in the player name and the base stat (ex. 4.5 rebounds) and have the model output whether or not to output over or under based on those parameters

1. Assessing the Modeling Algorithm (Still a Work in Progress)

Based on these results, I’ll use a similar, yet more primitive diagnostic process that evaluates how effective the algorithm is behaving, and is later used to train the model. It’ll simply be based on whether the algorithm was right or wrong (over vs. under) based on a computed value, and if so, by how much. If the algorithm was right, make sure the value specified is close to the base stat, and if it was wrong, penalize it based on how far it’s output was from that base stat.

1. **Data Preparation**:

* Collect relevant NBA betting data and WNBA player statistics with predictions and actual outcomes.
* Calculate the deviation between predictions and actual performance to serve as a basis for evaluating the consistency of a player
* Assess Correctness:
* Map the correctness of each prediction based on whether it was right or wrong compared to the deviation.
* Removed Extraneous columns, null values etc.
* Removed Injury Years (might deal with this differently later)
* Changed Data Types where necessary

1. **Modeling**:

Decision based on a multitude of factors a.k.a. predictors

Predictor Design:   
Influential Variables:

* Last 5 games average
* Consistency (Standard Deviation)
* Trend
* External Factors (Defense, Home vs. Away etc.)
* Minutes played
* Team Position

Each of these predictors held a different weight (that could vary in some predictors), which would increment or decrement a final decision score based on their recommended bet (over vs. under). That final score would also output Over or Under based on whether the value is positive or negative.

**Threshold Calculation**: Threshold is the cut-off value between how far I’d left the target variable deviate from the mean before changing the decision (over or under; is this player reliable or not?).

1. **Calculate Optimal Threshold:**

* Use precision-recall analysis with the deviation as scores and correctness as true labels.
* Find the threshold that maximizes the F1 score, balancing precision and recall.

1. **Compare to Base Stat:**

* Compare this threshold to the player’s base stats to gauge consistency.

1. **Betting Strategy:**

* Use the threshold to decide whether to bet "over" or "under" on future performances, focusing on players whose deviations are consistently above or below the threshold.

Essentially evaluates how wrong or right (how off) a series of basketball bets were, and compares this to the reliability of a player (standard deviation), and choses the most likely outcome for correctness being based on upper or lower.

1. **Evaluation**:

For my linear regression models in my Trend predictor and MP Predictors, I used the r\_squared metric from the SCIKIT LEARN Library to evaluate the efficiency of the predictor and determine its weight on the final decision score.

To assess the trade-off between precision and recall when predicting binary outcomes (Over vs. Under) for each bet, I used the F1 score to determine the optimal threshold for making decisions based on maximizing both precision and recall equally.

Lastly, to evaluate a player's variability for a given stat, and thus overall consistency, I used standard deviation as a quantitative representation of this which influenced the predict\_From\_Consistency function’s recommended output.

1. **Conclusion**:

Still looking for a way to retrieve data of PrizePicks bets directly to test the model performance overall. I plan on inputting synthesized data based on real bets from other sources to evaluate the model as an alternative (am open to suggestions of course).