Sports betting is a multi-billion-dollar industry, with betting lines constantly set for a variety of sports and teams and updated every minute of every day. In this project, our group explored the possibility of not only using structured tabular data to make sports lines, but also collecting unstructured sentiment data to fine-tune money lines and point spreads in line with the public. The typical goal for a sportsbook office is to tease lines down to achieve a close to 50/50 public bet on opposite sides of the same bet, assuming an identical or similar price. Sportsbooks thrive when betting lines are well-balanced, as overly predictable or inaccurate lines could lead to consistent wins for bettors, ultimately threatening the sportsbooks' profitability. However, this scenario is far from reality. NBC estimated that gamblers would wager over $35 billion on the NFL this year alone. This underscores the critical importance of creating balanced betting lines, as well educated betting. An effective line ensures a guaranteed profit for sportsbooks, while an ineffective line allows a profitable opportunity for the public to capitalize on. Our goal is to improve the accuracy of sports betting lines, particularly during the NFL season, helping sportsbooks maximize their weekly profits. Additionally, bettors using our model may also gain an edge by identifying and capitalizing on inaccuracies in lines set by other sportsbooks. The sports betting industry continues to grow rapidly, and factoring in public sentiment about games presents a unique opportunity for more accurate predictions. By leveraging sentiment analysis alongside traditional statistical data, our approach could benefit both bettors seeking to make informed decisions and sportsbooks looking to refine their strategies.

Our primary audience includes individuals engaging in sports gambling as well as sportsbooks themselves. By using this model, bettors could potentially make smarter wagers by relying on comprehensive data analysis rather than gut feelings or surface-level statistics. For sportsbooks, the model offers a tool to integrate public perception into their line-setting process. Understanding whether sentiment skews positive or negative for a team could help them adjust their initial lines to balance bets more effectively, minimizing risk when one side of the line attracts the majority of bets. Sportsbooks aim to create balanced lines so that roughly half of bets land on either side. This approach mitigates risk if one side wins and ensures consistent profits, referred to as “hold”, typically between 2-6% on average for a balanced wager. The ability to forecast public betting behavior from the opening line would be incredibly valuable in achieving this balance.

Currently, there is limited publicly available documentation on how major sportsbooks set their lines. This is largely due to the industry's competitive nature and its relative newness. Sharing intellectual property could lead to significant losses if competitors gain an advantage. Similarly, if the public gains access to the data and models used by sportsbooks, it could empower bettors to make more educated decisions, potentially reducing the sportsbooks' edge. When sportsbooks set an initial line, it often shifts as bets accumulate on one side. The goal is to balance the wagers, but this adjustment process is largely influenced by sentiment and external factors. We believe sentiment can be integrated as a quantitative measure within a predictive model, making the process more analytical and precise.

To create our model, we collected data from two primary sources: **Reddit API** for sentiment analysis and **Pro Football Reference** for quantitative NFL performance metrics. For sentiment analysis, the Reddit API provided access to a wealth of unstructured data in posts from relevant subreddits like *r/NFL*. In our analysis, we focused on the NFL subreddit rather than diving into team specific subreddits due to the potential bias that would result from drawing sentiment analysis from a large group of fans that may have a more pessimistic or rosy outlook on their team. Those posts on the NFL subreddit are likely to be more objective than those that are put out by specific teams. Posts were filtered by date to ensure that only data from the week of the game was used for analysis, making the sentiment more relevant. This data was processed and standardized using natural language processing (NLP) techniques, specifically the Vader Sentiment Analyzer to extract sentiment scores and after tying observations to a specific team by searching observations for keywords that typically included the city name, star players, coaches, and other relevant terms. An average sentiment score by team and week was then acquired which was stitched together with our structured data from Pro Football Reference.

For performance data, **Pro Football Reference** served as our source of structured NFL statistics. This included essential metrics such as passing yards, rushing yards, turnovers, defensive and offensive efficiency on third downs, and more. The platform's comprehensive dataset provided historical performance trends and real-time updates, making it ideal for predicting game outcomes and team performance. We aggregated and filtered the data by team and game week to align with our predictive model's requirements. Both datasets were integrated into a unified framework for analysis. Unstructured sentiment data from Reddit was stored in a NoSQL database, MongoDB, to handle its variability. After cleaning and transforming our data as well as ensuring it was properly formatted for analysis, we were able to begin modeling.

Results

Our overall best model was an ensemble composed of a Logistic Regression, XGBoost, SVM, and Random Forest model. We used propensity-based classification in classifying a win or a loss. Our overall record in the test data was 63-23, meaning we successfully predicted 63 games and failed to predict the outcome of 23 games. This amounted to an overall accuracy of 0.73, a recall of 0.77, precision of 0.72, and f1-score of 0.74. We then transformed each win probability into a spread and moneyline utilizing a 2.5% sportsbook hold. We were able to conclude that in week 13, our spreads were more accurate than closing spreads (covers.com) in 81.25% (13/16) of games. However, we also noticed that our model was not perfect. In particular, we were often off on games that included some sort of factor not in our model, such as a key player injury, weather, or other external factors. Overall though, we were very pleased with the performance of our model.

Conclusion

Our model can be used by both sportsbooks and bettors. Sportsbooks could implement our model as a way of setting better initial lines that may result in a more evenly distributed public bet. As mentioned earlier, the closer a public bet is to 50/50 on a particular line, the more guaranteed profit a sportsbook makes after the conclusion of that game, regardless of outcome. It is important to note that our previous claim also assumes the price on opposite sides of the same line is the same. However, for spreads, it is quite typical for both sides of the line to be -110, which results in a 4.55% hold regardless of outcome in a 50/50 split public bet.

Bettors can use our model to identify potentially profitable markets before the lines adjust due to the public. For example, for Buccaneers @ Panthers in Week 13, the sportsbook line closed at -6 Buccaneers. However, our model closed at -2 Buccaneers. This indicates that our model believes this will be a closer game than the sportsbook lines offer. A bettor could take note of this, and as such bet on the Panthers to cover +6. In this case, the result of the game was a 30-27 Buccaneers victory, meaning a bet on the Panthers to cover +6 would have won, whereas a bet on the Buccaneers at -6 would have lost. Additionally, our line in this case was only 1 point off the actual outcome, whereas the sportsbooks missed by 3 points. Although bettors can not directly bet on our lines, they can utilize our insights to help place more educated bets and beat the books.