ISyE 6740 – Fall 2022
Using Daily Fantasy Salaries to Beat Vegas

Author: Robert Bingaman
December 6, 2022

Table of Contents

1. Background

In the last decade, fan interest in professional sports has increased significantly through gambling and gaming opportunities. In 2018, the United States Supreme Court decided in Murphy v. National Collegiate Athletic Association to overturn a clause in the law, Professional and Amateur Sports Protection Act, which previously had been interpreted to restrict states from regulating sports gambling. This decision changed to allow states to regulate sports gambling however they see fit if the federal government does not regulate it directly. After this decision, there was a very quick emergence of the sports gambling industry in the United States. In 2021, the industry generated $4.29 billion in revenue and through June of 2022, they had generated $3.04 billion in revenue. (Thorsburg, 2022) In addition to the rise in sports gambling, fantasy sports have been a significant part of the sports entertainment industry for many years and continue to grow. The Supreme Court decision and fantasy sports have changed the sports entertainment industry in the United States in how fans follow sports, how sports are covered in media, and a revolution in the data analytics industry related to sports.

Within fantasy sports, there is a more specific segment called Daily Fantasy Sports in which users pay a fixed entry fee (ranging from free to hundreds of dollars), then pick a lineup from a selection of players that have an assigned salary (in the range of $500 - $10,000) but must stay within a predetermined budget. Each entry for selecting players is in a closed pool of other players and the best performing teams in the pool receive prize money for winning. This segment [of fantasy sports] has immense focus on how lineups can be optimized, and value found with players having lower salaries than others. Globally, this industry generated $20.36 billion in 2020 and is expected to grow to $38.5 billion by 2025. (G & M News, 2021)

2. Problem Statement/Objective

It is interesting that casinos are making offers on the sports gambling markets in addition to publishing data about specific players competing in those games. I believe there could be an opportunity to find value in the sports gambling markets given the published "salaries" of players for the daily fantasy competitions. First, the value of the players might correspond to the offers provided for sports gambling on each game. If this is true, we can use machine learning methods to identify trends in which daily fantasy salaries might present meaningful value in relation to the sports gambling offers. Second, the daily fantasy salaries do not fully represent the same value offered in the sports gambling markets and using machine learning methods, we can exploit this disparity to find value in the sports gambling markets. The objective is to develop a predictive model that gives accurate results on the outcomes of:

1. Who wins or loses an NFL game
2. Who wins a handicapped game, called a point spread.
3. Which side of an over/under point total will win.

After getting accurate results for these outcomes, the further objective is to use the results to gain profitable results against the casino in gambling on these offers.

3. Data Sources

To explore these objectives, I chose to get all relevant details for games from 2016 through 2021 to develop the models. After identifying the optimal models from those years, I used 2022 data to evaluate the potential usefulness of these results for gambling if I were to have followed the models' recommendations during the current season. To do this, I leveraged three data sources:

1. Daily fantasy salaries from fantasydata.com for 2016 – 2022 games. (fantasydata, 2022)
2. Gambling offers on Moneyline bets (a bet that is on a team to win or lose the game), spread bets, and over/under point totals from bettingdata.com. for games from 2016 – 2022. (bettingdata, 2022)
3. NFL game results for 2016 – 2022 from profootballreference.com. (Pro Football Reference, 2022)

3.1 The daily fantasy salaries were for every player on the active roster of the NFL team for every game starting in 2016 through 2022. Due to the high number of players this resulted in, I chose to limit the salaries used in the predictive models to the highest number of players at following positions:

a. Top salaried quarterback – teams typically only use one quarterback in a game.
b. Top two salaried running backs – teams typically do use multiple running backs in every game.
c. Top four salaried wide receivers – teams typically do throw to many wide receivers.
d. Top two salaried tight ends – this position varies by team, some use one tight end primarily, but some use two tight ends regularly.
e. Salary of each team's defense/special teams' position – every team only has one defense/special teams' position since it is combined as a single unit.

3.2 The gambling data from bettingdata.com provided several valuable pieces of information:

a. Point spread which is an offer made on a game intended to make the matchup more equal. A favorite will have a negative point spread, while an underdog will have a positive point spread. For the favored team to win in this situation, they would need to beat the under dog by more than the point spread. For example, if the favorite's point spread is -3.5 and they win the game 25 to 21, they would also "win" with the point spread because (25 – 3.5 = 21.5) which is still greater than 21. The odds for both sides of point spread bets are assumed to be -110. This means that the bettor would need to bet $110 to win $100.
b. Point totals which is an offer made on a game of how many total points will be scored between the two teams in the game. Bettors can bet over or under this point total offer. For example, if the point total for a game is set at 40.5 total points, and the hypothetical result from part (a) happens, the two teams would combine for 25 + 21 = 46 points, the "over" bet would win. The odds of both sides of this bet are also assumed to be -110, as in the example above.
c. Moneyline odds on both the favorite and underdog. Different from the point spread, this bet is just for picking who will win the game, team A or team B. The difference is that the casinos adjust for the risk, not with a point spread, but a different offer of odds for each of the teams. A big favorite might have odds of -300, requiring a bet of $300 to win $100, while a big underdog might have odds of +250, meaning a win of $250 for betting $100.
d. Which team in each matchup was favored or the underdog.

All the odds described above can be converted to an implied probability. The most important to note is -110 implies a 52.4% implied probability. A calculator and common odds conversions can be found at AceOdds.com (AceOdds, 2022).

3.3 The NFL game results from profootballreference.com were acquired from scraping the webpages of each respective year's results table in Python. After a bit of cleaning, the final data acquired from this website contained the following attributes:

a. Season (2016 – 2022)
b. Week of season (1 – 17 in 2016 through 2020, 1-18 starting in 2021)
c. Home team name
d. Away team name
e. Home points scored
f. Away points scored

After acquiring all three separate data sources, significant wrangling was required to sufficiently structure them to be joined and trained with machine learning models. Due to the complexity of joining these primary data sources together, time did not allow to pursue other features that were mentioned in the project proposal such as aggregated performance statistics or advanced metrics published online.

4. Methodology

To identify if the daily fantasy salaries were useful in identifying opportunities to win bet offers from the casinos, I built a series of classification models to predict the outcome if the team won or lost the game, if the point total in the game exceeded the offered point total, and if the favored team covered the spread in the game. For the win and loss predictions, the target classification was if the home team won the game, indicated by a 1. For the point total predictions, the target classification was if the game resulted in going over the offered point total from the casino, indicated by a 1. Finally, for the spread bets, the target classification was if the favored team covered (won the game by more than their handicapped points), indicated by a 1.

In each of these three sets of training data, I included the following training data:

a. The home and away positional daily fantasy salaries for the positions indicated in section 3.1.
b. The home and away team's moneyline odds.
c. The offer of the point total from the casino.
d. The point spread from the casino.
e. Aggregations of the home and away offensive positions daily fantasy salaries, then a total of both of those aggregations.

In the point spread training data I also included a binary indicator if the home team was the favorite or not.

After identifying the target variable and the training points for the three objectives, the first step was to build a set of candidate models that performed the best on the classification problems. To do this, Scikit-learn's Grid Search Cross Validation was used with varying values of parameters for the following classification machine learning algorithms:

a. K-Nearest Neighbors
b. Support Vector Machine
c. Naïve Bayes
d. Random Forest
e. Light Gradient-Boosting Machine (LightGBM)
f. eXtreme Gradient Boosting (XGBoost)

In the project proposal, I suggested using a one-class support vector machine as well, but with many attempts at tuning, the data proved to not vary enough to detect anomalies and therefore was abandoned before testing on all the different objectives.
After each of the six candidate models were optimized and selected, hypothetical bets using each of the six candidate models on the historical data in 2016 – 2021, and games from 2022 through week twelve of the NFL season were run to see how much return on investment (ROI) would have been generated. In the evaluation of bets, each candidate model was evaluated at increasing margins compared to the implied probability of the offer from the casino. In theory, if the models are accurately predicting the probabilities, as the margin of being greater than the implied probability increases, the return on investment should also increase.

5. Evaluation and Final Results

In this situation, since the correct prediction of each result is most important, and the data is reasonably balanced on the historical outcomes, the accuracy measure is the most critical. However, precision, recall, and AUC are still of interest and important in this classification problem. In this section, each target result will be evaluated independently.

In addition to the results of the prediction being evaluated, the ROI will be evaluated for the hypothetical bets. Some key considerations are that on moneyline bets (wins and losses), the odds vary significantly in each matchup. For moneyline bets, the team to bet on will be selected by which one has the higher advantage over the implied probability. This means that underdogs will be selected to be bet on, if they are the team that has a higher predicted probability than their implied probability from the odds.

Meanwhile, for spread and over/under bets, all games are assumed to be at -110 odds, therefore whichever side of the bet is predicted higher will be selected if it is above 52.4% (the implied probability) by at least as much as the indicated margin.

5.1 Wins and Losses

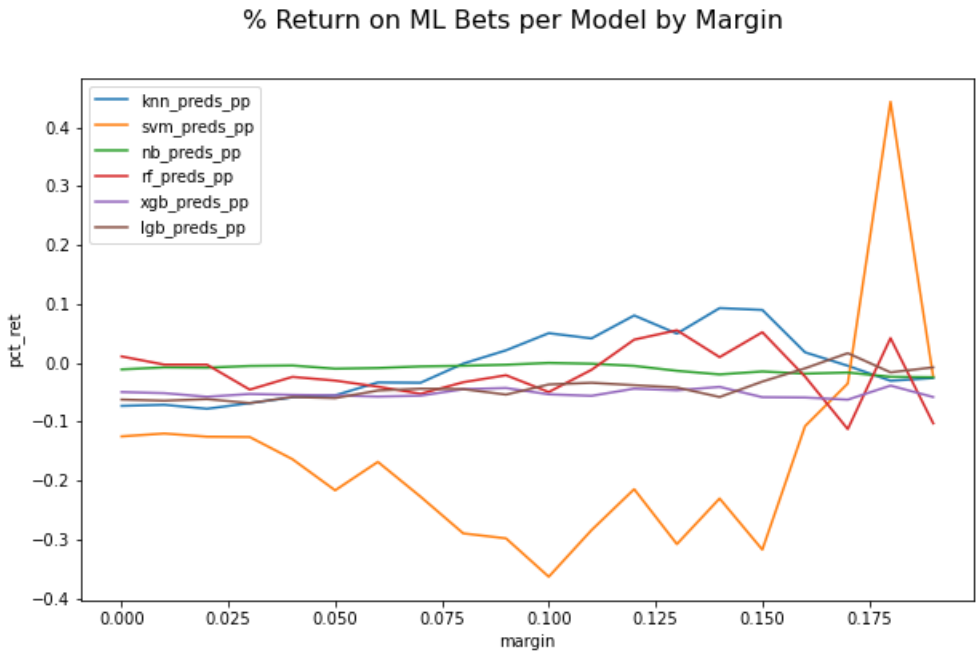
First to evaluate how the models perform on wins and losses, the following metrics were observed:

	Accuracy	Recall	Precision	AUC
XGB	0.634167	0.815113	0.621324	0.619847
LGBM	0.642796	0.660772	0.669381	0.641373
RF	0.672994	0.681672	0.700826	0.672307
KNN	0.647972	0.646302	0.681356	0.648105
SVM	0.673857	0.742765	0.679412	0.668403
NB	0.663503	0.680064	0.688925	0.662192

From this, Random Forest and Support Vector Machine really stand out as strong performers, having over 67% accuracy each. Random Forest also displays the highest AUC and precision. The XGBoost results are the most imbalanced with a remarkably high recall, but a correspondingly low precision, meaning that true positives are identified at a high rate, but there is also a higher rate of false positives than the other models. Since the moneyline bets have widely varying odds, it is important to predict

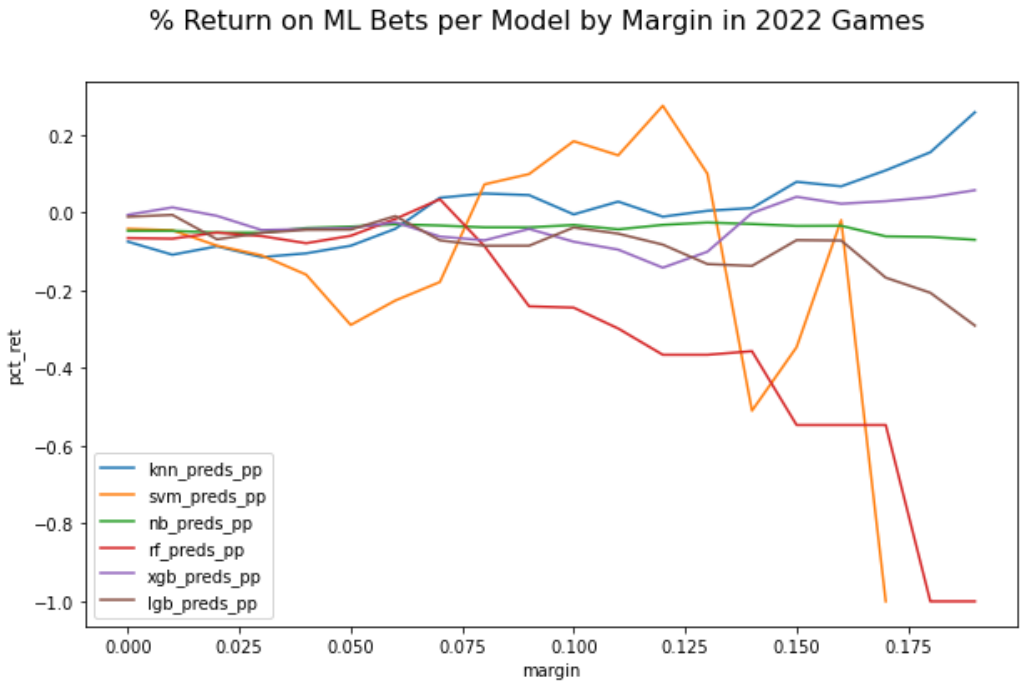
the accuracy of these games well so that the right team can be selected. This is a promising start.

Now to explore the potential returns on this model. First to examine the returns by



model by the expanding margin as described in the methodology section:
This chart highlights that most of the models are generating negative returns when using the predicted probabilities to compare to the implied probabilities of the moneyline offers. The Random Forest and K Nearest Neighbors models do provide positive returns at higher margins, but it is not as consistent enough, or high enough to trust for generating consistent returns. The Support Vector Machine shows a strong return at the highest margins, but this is due to a small number of games selected that meet the criteria, and this induces the corresponding high variability seen as the return drops off significantly.

After examining the potential returns for 2016 – 2021 hypothetical bets, the same logic was applied to 2022 games through week 12 (Sunday November 27,2022) to see if it would have generated positive returns this season.



Again, there are large swings of performance in the model. Surprisingly, Support Vector Machine produces strong returns in 2022, even though 2016 – 2021 it mostly produced negative returns. In 2022, KNN and XGBoost produce results that are following the pattern that makes sense, increasing returns as the margin increases.

What is most surprising is that the Random Forest had some of the best performance metrics when evaluating the results of the model, but when applying the predictions to gambling, the predictions are not resulting in a profitable outcome. Overall, the results on these bets vary significantly and lead to a conclusion of not trusting them as is to make accurate picks on games for NFL moneyline bets.

5.2 Spread Bets

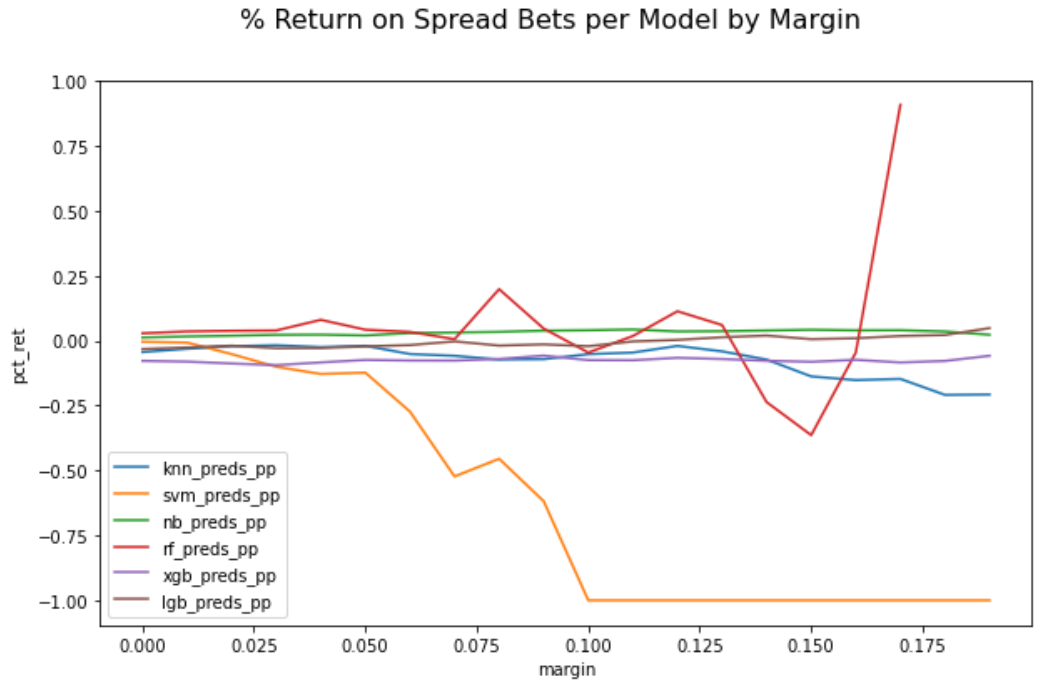
To evaluate how the models perform on spread results, the following metrics were observed:

	Accuracy	Recall	Precision	AUC
XGB	0.482312	0.673469	0.461245	0.494799
LGBM	0.506471	0.445269	0.467836	0.502473
RF	0.530630	0.437848	0.494759	0.524569
KNN	0.503020	0.397032	0.460215	0.496096
SVM	0.537532	0.031540	0.548387	0.504480
NB	0.532355	0.311688	0.495575	0.517941

Compared to selecting who wins and loses games, these results are relatively poor, but when considering the implied probability on a spread bet is 52.4% - in theory, if the model can correctly identify the correct result at a higher rate than that, it could be

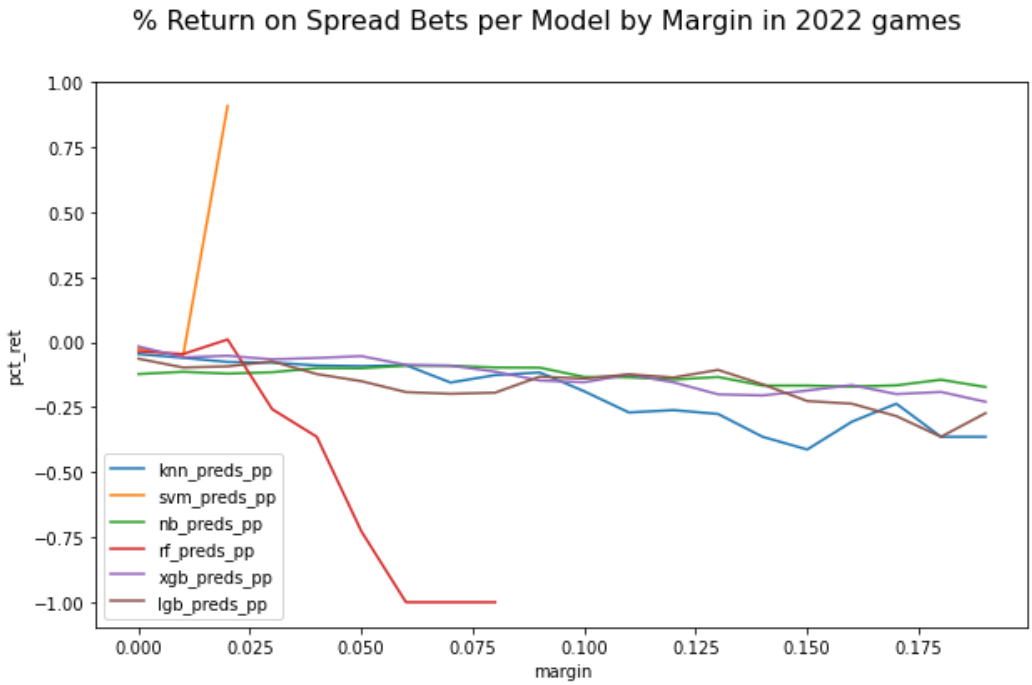
profitable. Again, Random Forest and Support Vector Machine produce some of the highest accuracy, but Support Vector Machine can barely predict a true positive, meaning the value of the model is not incredibly significant if it is only selecting one side of the bet every time. XGBoost is under 50% accuracy, but is again showing high recall, but with lower precision. Random Forest is the most balanced and correct model, but it does not show a lot of promise in being profitable to gamble with.

Below are the returns by model on games from 2016 – 2021 in hypothetical bets:



This chart shows a lot more consistency than in the wins and loss profitability by model. Naïve Bayes shows consistent, positive ROI, but it is quite low. Random Forest does show positive ROI through most of the spread of margins, with large swings at the higher margins, but that is due to a small number of games being selected. Support Vector Machine is generating extremely negative returns, showing the value of needing strong recall, not just accuracy.

Now to compare these results with 2022 games:



On 2022 games, Support Vector Machine and Random Forest had very few games selected to bet on and produced erratic results. Naïve Bayes again proved to be one of the strongest performing models, but that is not valuable because it still generated negative returns. All the other models also produced negative and consistently declining ROI results on spread bets.

It is easy to conclude that these models are not suitable for betting on spread bets with the criteria to bet in this scenario.

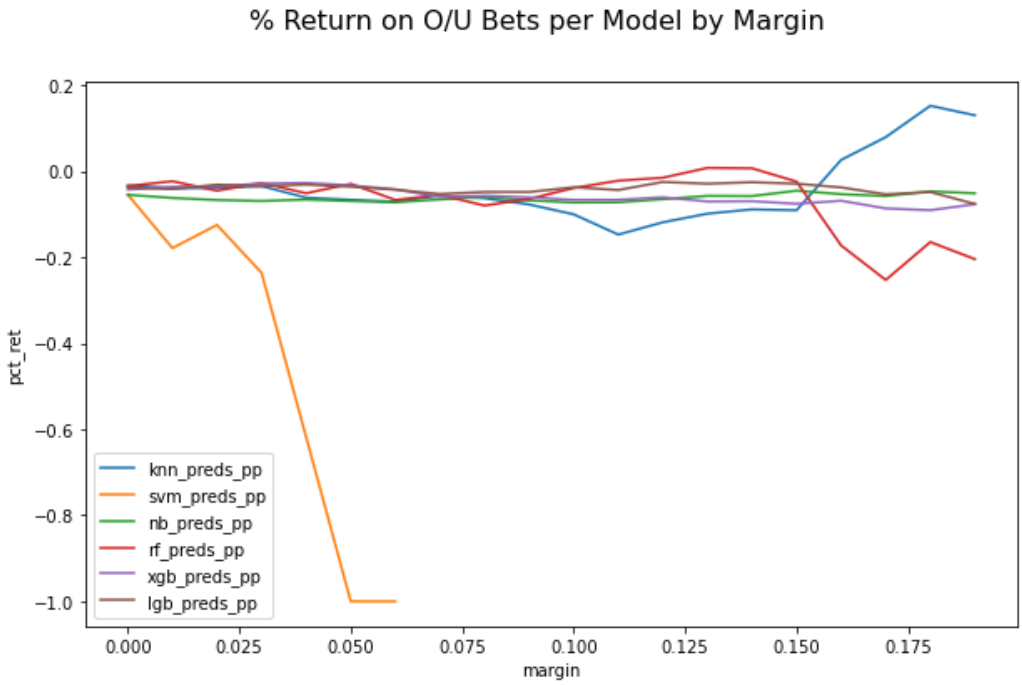
5.3 Over/Under Bets

To evaluate how the models perform on over/under results, the following metrics were observed:

	Accuracy	Recall	Precision	AUC
XGB	0.500431	0.761404	0.494869	0.504641
LGBM	0.511648	0.512281	0.503448	0.511658
RF	0.525453	0.524561	0.517301	0.525439
KNN	0.505608	0.459649	0.497154	0.504867
SVM	0.500431	0.138596	0.473054	0.494595
NB	0.493529	0.664912	0.489032	0.496293

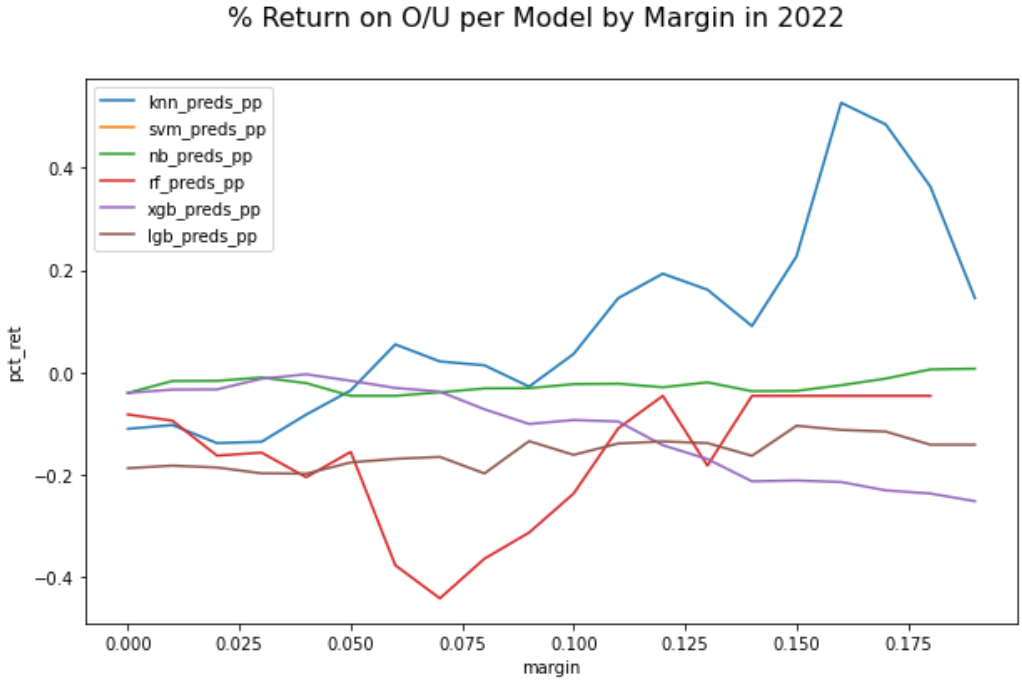
Again, Random Forest stands out as the top performer and XGBoost shows the strongest recall of all the candidate models. Overall, none of the candidate models show much hope, aside from Random Forest, in producing results strong enough to consistently identify over/under bets to find a profitable margin for gambling.

Below are the results for over/under hypothetical bets in the 2016 – 2021 seasons:



These results are consistent with the metrics observed from the results. Incredibly low, or negative, ROI and some variability at high margins, with Support Vector Machine once again producing completely losing results. KNN might show promise at higher margins as it increases sharply at those high values.

And, finally, the results of over/under hypothetical bets on 2022 games:



The ROI for over/under bets in 2022 shows a potential candidate for gambling selection. KNN shows a similar pattern in 2022 as it did in 2016 – 2021, a strong surge

of ROI at high margins. This is an encouraging development compared to the other model evaluations. All the other models generate negative returns consistently through the range of margins. The predictions from the Support Vector Machine model did not have any games qualifying for selection in 2022.

6. Conclusions and Next Steps

Overall, I am concluding that daily fantasy salaries are not enough, in conjunction with odds offers, for finding value in sports gambling markets using the identified methods and outlined gambling strategy. One interesting finding was that the LightGBM model consistently did find the salary data as the most important in its models. For all three objectives, wins/losses, spread results, and over/under results, four out of the top five features in importance were salaries of players from either team. This contrasts with XGBoost which always had the home and away team moneyline odds as some of the most important features. From this, I do believe that the daily fantasy salaries prove they provide some amount of value in assisting with identifying results, but the casinos are not "giving away" immense insight to untapped value by publishing these salaries.

There are a few ways I believe this could be expanded upon to potentially find value in these gambling offers. First, I would incorporate more data as potential training features. Advanced player metrics that are published could provide value to help identify the results more successfully. Second, I believe that ensembling some of these predictions could help. Perhaps connecting the results of the win/loss result predictions with the daily fantasy salaries could help the spread predictions. Third, a regression approach could be used with the spread and over/under results. For example, the models could predict a point total for each team given their players' salaries, or a total for the game, then compare that to the offer from the casinos and make the gambling decision on that. Finally, there could be other decision criteria for the decision to bet or not bet on an offer. With the classification models providing predicted probabilities, the intuition is that if the predictions are higher than the implied probability, it is good value. This intuition does not hold up in general, and other strategies, with new selection criteria, need to be identified. For example, on moneyline bets, only betting on results with a predicted probability over 0.7, regardless of the odds – or on spread bets, if a game has a spread of more than 7 points, do not bet on it.

References

AceOdds. (2022, December 5). *Odds Converter*. Retrieved December 5, 2022, from AceOdds: https://www.aceodds.com/bet-calculator/odds-converter.html

bettingdata. (2022, December 5). *NFL Oddds*. Retrieved November 28, 2022, from bettingdata: https://bettingdata.com/nfl/odds

fantasydata. (2022, December 5). *NFL Daily Fantasy Salaries*. Retrieved November 28, 2022, from fantasydata: https://fantasydata.com/nfl/daily-fantasy-football-salary-and-projection-tool

G & M News. (2021, September 1). *DFS market to grow 9.5% to USD 22.3 billion in 2021*. Retrieved November 7, 2022, from Gaming & Media News: https://g-mnews.com/en/dfs-market-to-grow-9-5-to-usd-22-3-billion-in-2021/

Pro Football Reference. (2022, December 5). *NFL, AFL and Pro Football History*. Retrieved November 28, 2022, from Pro Football Reference: https://www.pro-football-reference.com/years/

Thorsburg, C. (2022, September 8). *Sports betting — a hugely popular, multibillion-dollar industry — is poised to become an American epidemic*. Retrieved November 7, 2022, from GRID News: https://www.grid.news/story/economy/2022/09/08/sports-betting-a-hugely-popular-multibillion-dollar-industry-is-poised-to-become-an-american-epidemic/