PREPARED FOR:

2023 Spring | MSDS 498 | Team 55

PREPARED BY:
Timothy Steed
Timothy Chiu
Christopher Kradjian
Garrett Lynch

GAME CHANGER

Sports Wager Analytics

- Capitalizing on online sports betting market inefficiencies

Final Report

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Project Overview

We presume a market of this size but with relatively low maturity is bound to be ripe with inefficiencies and therefore our group has positioned ourselves to uncover and capitalize on them. We are a team of analytics and data science professionals in a sports analytics company, GAME CHANGER who have come together to create a new analytical tool prototype to generate positive returns on online sports betting.

The team seeks to achieve the following:

1. Identify team and game characteristics (i.e. features) that rank the most impact to scoring metrics.
2. Build a classification model which would analyze NBA games and classify specific contests as recommended bets based on user-defined confidence thresholds.
3. Devise an optimal betting strategy for the most popular wager types in NBA games.

Problem/Opportunity

Many online sports betting websites have promotions to entice players to bet more considerably and more often. Most have playthrough requirements to their bonuses, which are the number and amount of bets required before you can withdraw. The NBA is a fast-paced game where game flow and line changes occur within a few minutes.

We plan to model these moving lines, to see if at any point before or during the game, they shift enough to allow for confidently predicting a better than ~52.4% success rate. The specific value is over 50% (i.e. a coin flip) due to the need to overcome the “vigorish”, online sportsbook cut for handling the action on a sports wager; for the most typical issued odds of -110, bettors will need to win 52.38% of times or more to break even or profit.

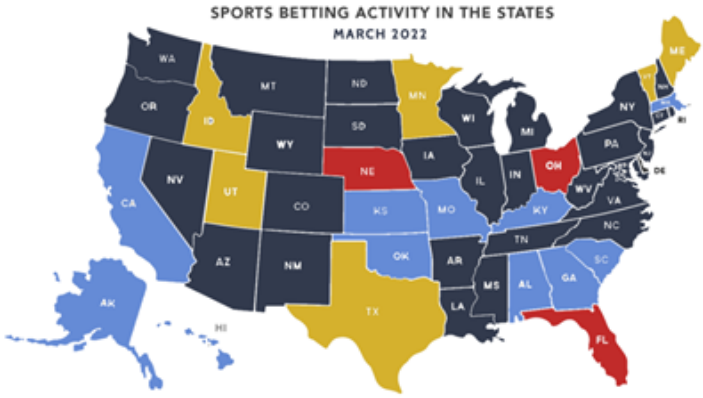
Proposed Solution

We seek to design and develop a new method of improving the accuracy (and returns) from predicting the outcomes from sporting events. We will focus this initial phase around the National Basketball Association (NBA), and focus specifically on the betting lines - which team will win, or betting point spread - the winning or losing team will cover the published margin of victory or defeat.

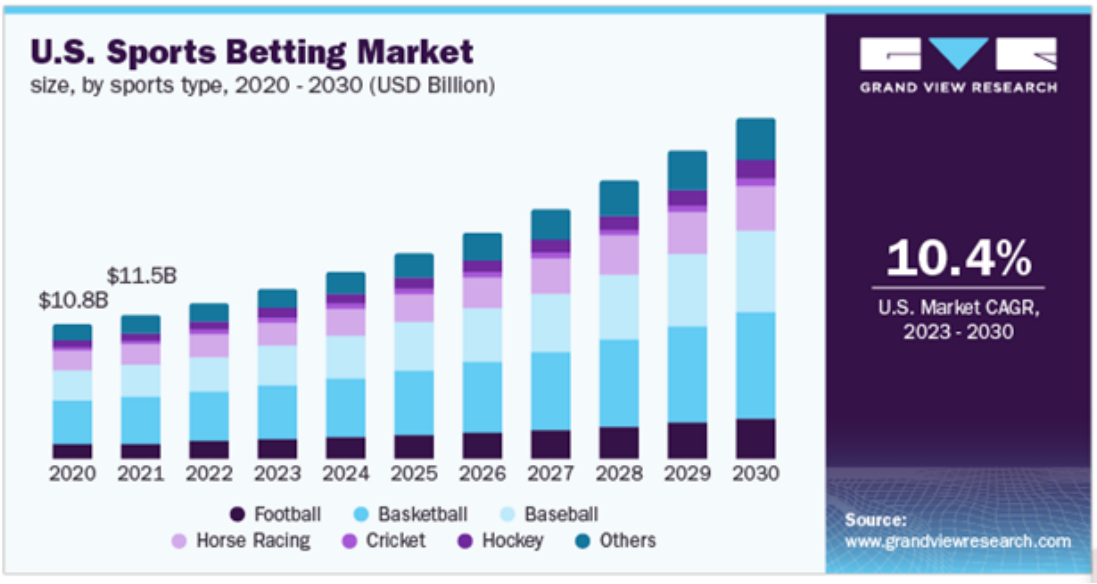
Market Analysis and Expansion Strategy

Since the May 2018 Supreme Court decision to allow states to set their own laws on sports betting, it is yet to be seen the difference that present opportunities to review which models, taxing and laws, are better than other states. Below maps show the status of each state compared to April 2019 (Bosch, 2022).

Depending on the industry, businesses may operate with payroll percentages in the 15-30% range. We utilised 20% as our payroll percentage to be conservative. Taxes are based on federal government excise tax of 25% and Nevada state income tax of 6.75%. We decided upon Nevada because it had one of the lowest taxes for this industry.

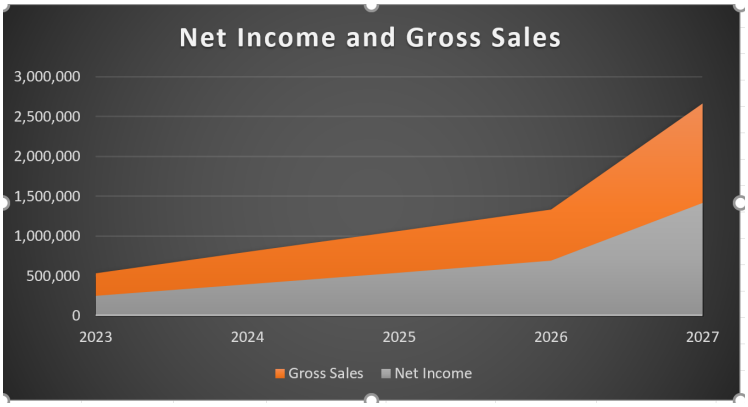


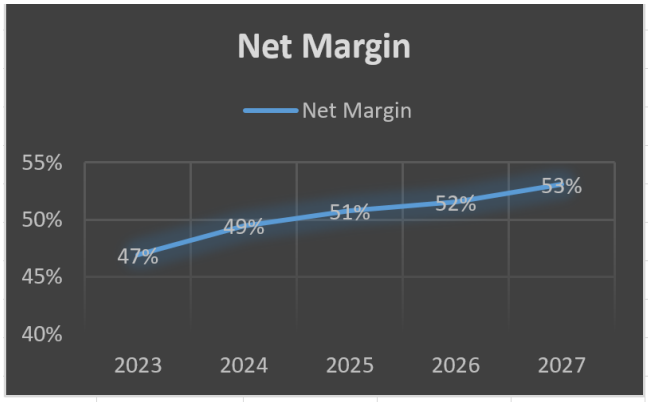
According to the figure below, the U.S. sports betting market was already worth USD $10.8 billion in 2020 at an average 10.4% CAGR through 2030. The same figure also shows that basketball has been the dominant sport the American public wagers online. The world of basketball sports betting offers a wide variety of wager types, among which the most popular are “Moneyline”, “Point Spread”, and “Point Total”.



Value

The Go-To-Market plan is simple and straight-forward: $20-30$ per wager for the first year based on confidence, and expect a 55-70% rate of capitalisation on invested funds. Assumptions made show a progression of increased wagers based on increased capital per year, with an average of 60% optimization rate. Wagers are based on an NBA season consisting of 1,230 games, and wagers averaging +-3 per game.





Year	Wagers	$ per Wager	Monthly Gross Sales	Projected Monthly Payroll	Cloud and API Hosting	Monthly Net Rev	Annual Net Rev	Net Income
1	3700	$20	$44,400	$8,880	$5,000	$30,520	$366,240	$249,959
2	3700	$30	$66,600	$13,320	$5,000	$48,280	$579,360	$395,413
3	3700	$40	$88,800	$17,760	$5,000	$66,040	$792,480	$540,868
4	3700	$50	$111,000	$22,200	$5,000	$83,800	$1,005,600	$686,322
5	3700	$100	$222,000	$44,400	$5,000	$172,600	$2,071,200	$1,413,594

Dashboard and Next Steps:

The main features we have identified are:

- Web Application and Dashboard
  - User Interface (UI/UX)
  - Back-end development (connected to data/models)
  - Live/real-time connections to APIs
- Distribution, as a part of Long-Term Strategic Financial Plan (LRP)
- Marketing and Sales, as a part of Long-Term Strategic Financial Plan (LRP)

INITIAL FINDINGS

Problems and Challenges

One of the first things the Game Changer team realized was the size and scope of the current project. With more than a dozen US based sports book makers, creating odds and lines on dozens of betting categories (over/under, money lines, spread, totals, parlays, etc…), across the 4 major US sports leagues (NBA, NFL, MLB, NHL), effectively solving this problem would be near impossible.

Considering our 'start-up' status, we needed to "crawl before we walked" and started out by first considering our organizational needs. In the book *Secrets of Analytical Leaders*, Eckerson wrote about the Analytics Maturity Model:

STAGE	DESCRIPTION
1. Analytically Impaired	"Flying blind" – Lacks data, analysts, and executive interest
2. Localized Analytics	Pockets of analytical activity, but no coordinated activity or strategy
3. Analytical Aspirations	A few strategic initiatives underway but progress is slow
4. Analytical Companies	Benefits from regular use of analytics, but it's not strategic
5. Analytical Competitors	Widespread use of analytics which delivers a competitive advantage

He goes on by writing that not long ago, "the Holy Grail for data management was an enterprise data warehouse" (Eckerson 2012). Leading organizations (think Google, Amazon, Apple), would find themselves at stage 5, where they have widespread use of analytics to drive a competitive advantage. At Transform 2019, Deborah Leff, CTO for data science and AI at IBM said that "If your competitors are applying AI, and they're finding insight that allow them to accelerate, they're going to peel away really, really quickly" (Venture Beat 2019). In the same Venture Beat article, they go on to question that even though "AI empirically provides a competitive edge, why do only 13% of data science projects, or just one out of every 10, actually make it into production?" (Venture Beat 2019). They suggest that these projects fail for reasons such as, (1) Falsely believing that simply throwing money at a problem can fix it, (2) Gaining access to the correct data, and (3) A lack of collaboration as these projects tend to span across silos. Their number one way to "avoid becoming one of the 87%" of failed analytics projects was to pick a small project to start with (Venture Beat 2019). In a 2016 article in the Harvard Business Review, the authors suggested that analytics projects tend to fail as the "efforts to adopt analytics upset the balance of power in the C-suite" (McShea 2016). They go on to say that poor leadership, lack of commitment, and no sense of ownership were the leading causes of failure.

With this information in mind, the Game Changer team identified specificity, scope, and simplicity as the prevailing themes of successful project implementations. The consulting firm Gartner, much like the leaders at IBM and Gap that spoke at Transform 2019, stresses the importance of defining a small project scope to ensure

simplicity. They mention that many organizations already struggle with their standard information strategy – "adding data [projects] to the mix increases complexity" (Buytendijk, Linden, & Laney, Big Data Strategy: Get Inspired, Get Going, Get Organized, 2015). Analytics projects without a specific problem to solve are **destined to fail**.

Scope Reduction and Pilot Selection

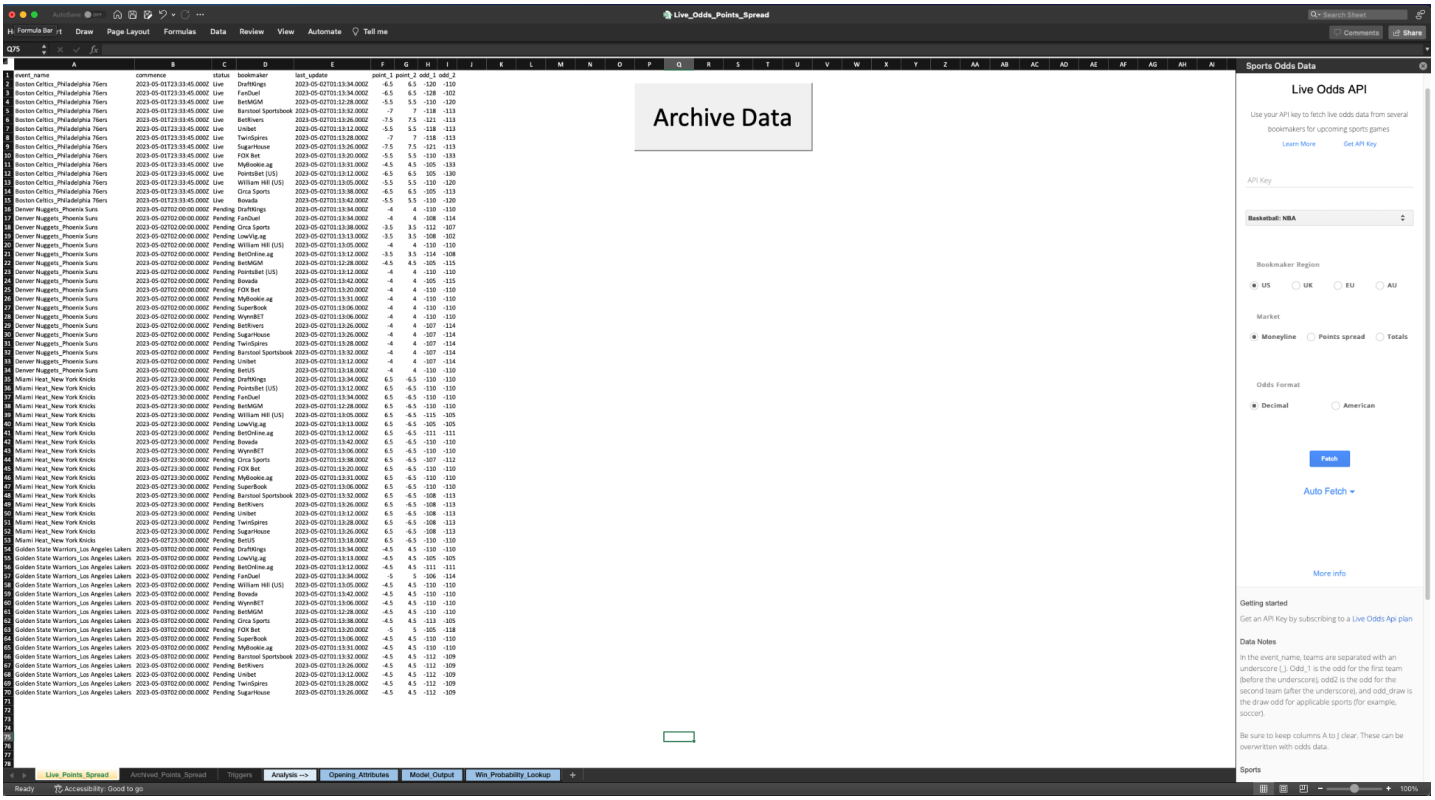
Considering that gaining access to the correct data was one of the leading sources of failed analytics projects, it was one of the driving factors behind our pilot selection. It also helped that all four of the Game Changer founders were avid NBA fans.

- **Specific**: Only model out the NBA win probabilities based on the moving Point Spread category, then only apply the strategy to a single sports book maker - the largest one, based on Las Vegas - MGMSports.
- **Small Scope**: This helps limit the data sources required and allows us to focus on building, testing, and tuning only a single fundamental model, which is applied in different ways.
- **Simple**: Deliver a simple, working prototype with a complimentary dashboard that helps answer a very specific question - can the real-time/in-game update of odds improve the probability of success for bettors.

Description of Data

For the purpose of the building the prediction model prototype , the input data was sourced from Kaggle.com (https://www.kaggle.com/datasets/erichqiu/nba-odds-and-scores) pre-compiled into 3 text files each from the 2012 season through 2018 season. The files represent NBA matchup data (described further below) and online betting data during the regular season and playoffs. Specifically the online betting data contain 3 major game bets information - moneyline, spreads, and over/under - from 5 different sportsbooks (Pinnacle, Bovada, Betonline, Heritage, and 5dimes).

On the other hand, the live in-game odds data will be acquired through an Excel add-on from the Odds API (https://the-odds-api.com/), an online sports betting API that is connected to multiple book makers. An Excel application (snapshot below) has been developed by our team solution architect, Christopher Kradjian**.** The application will be showcased in our team final presentation.



Overview of the Data

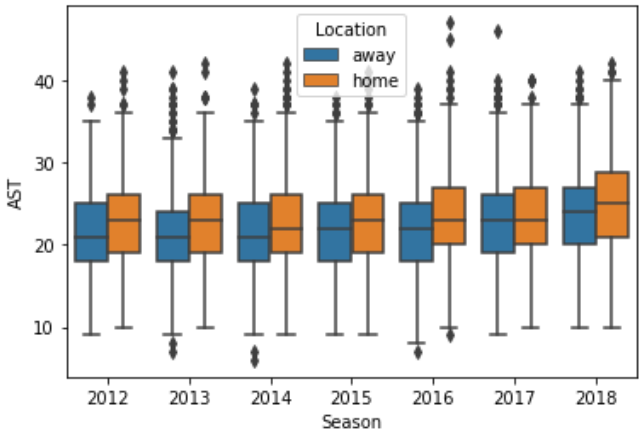
The data we used to build our predictive model came from 21 txt files we sourced from Kaggle. There are 3 files each from each season from the 2012-2013 through 2019-2019 NBA seasons The first of the 3 contains box score data for each team with their points scored by quarter and some summary statistics like field goal percentage, three point percentage, rebounds, assists, turnovers, etc. Each game since the 2012-2013 NBA season is represented by two data entries of the opposing teams sharing the same game ID. The matchup variables and statistics are listed in table 1 below.

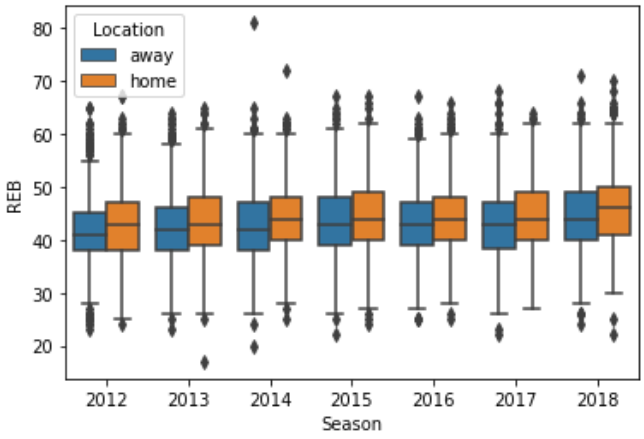
Table 1

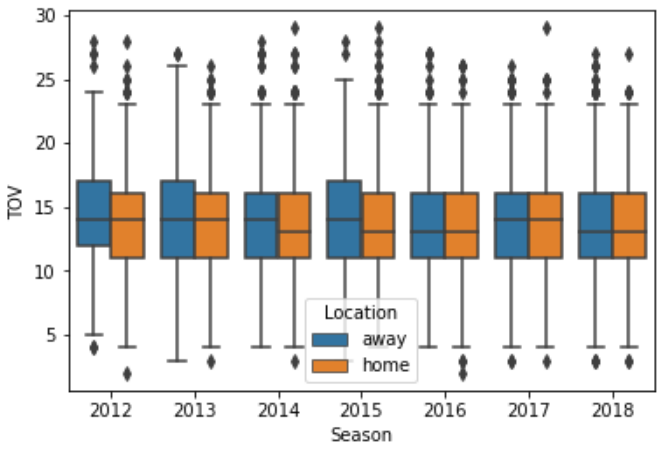
Game variables	Description	Score variables	Description	Performance variables	Description
Date	Date of match	PTS_QTR1 - PTS_QTR4	Quarter score	FG_PCT	Field goal percentage
GAME_SEQUENCE	Order of match of the day	PTS_OT1 - PTS_OT10	Overtime score	FT_PCT	Free throw percentage
GAMEId	Game identifier	PTS	Final score	FG3_PCT	Three point percentage
TEAMId	Team identifier	First_Half_Points	Half time score	AST	Number of assists
TEAM_ABBREVIATION	Three letters of	Second_Half_Points	Second half	REB	Number of

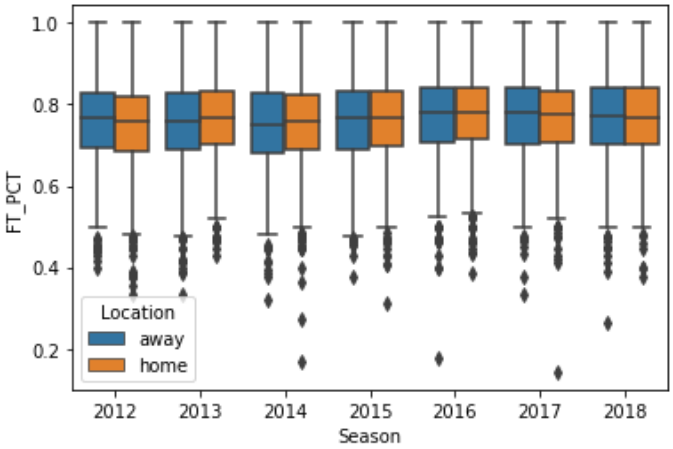
	NBA team (out of 30)		score (Final score minus half time score)		defensive and offensive rebounds
TEAM_CITY_NAME	City of NBA team			TOV	Number of turnovers
TEAM_WINS_LOSSES	Team wins-losses up till the time of record				
Location	Away or Home				
Season	Year of NBA season				

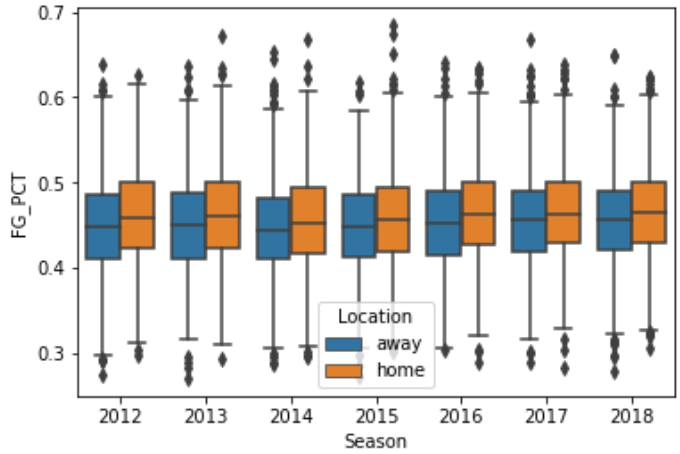
In the dataset we have verified the notion of “home court advantage” - that the team almost always performed better when playing on home court, as evidenced from the boxplot results below comparing away and home team statistics of every regular and playoff season since 2012.

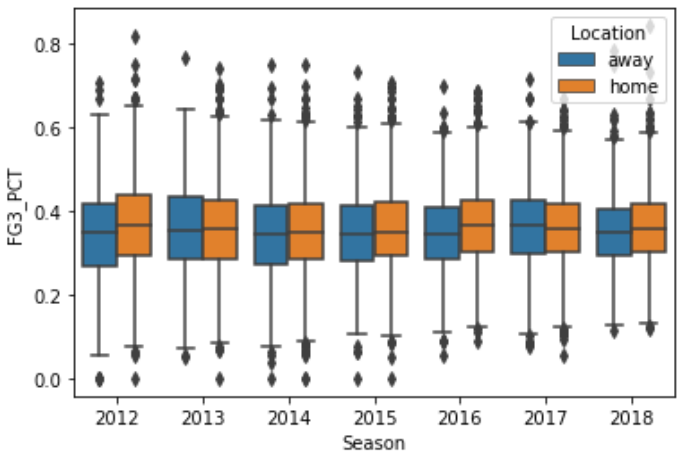




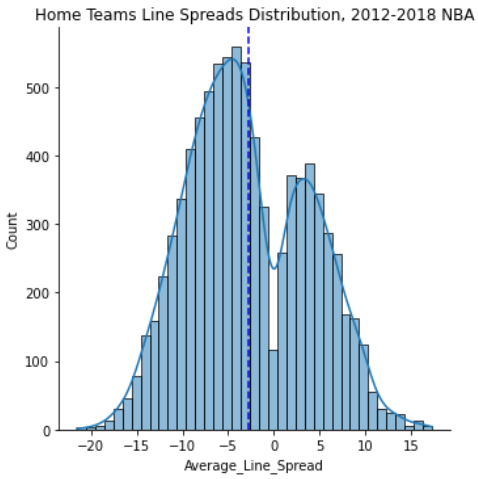


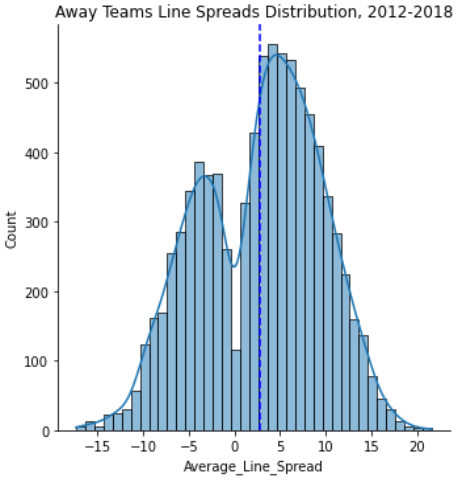






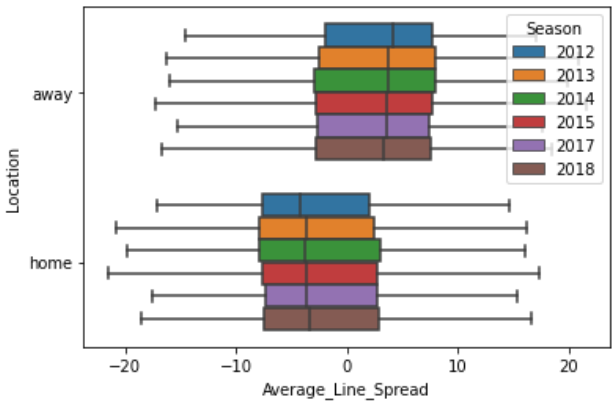
The same home advantage notion seems to be regarded by Sportsbooks, as evidenced by the average line spreads and moneyline set by prominent online sportsbooks over the 2012-2018 NBA seasons. The data roughly assumes a left-skewed normal distribution with a negative mean for home teams, and a right-skewed normal distribution with a positive mean for away teams, in other words home teams are more often viewed as the favorite while the away team as the underdog.

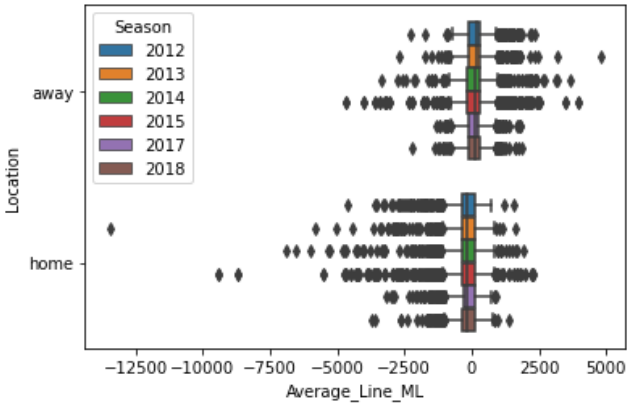




Note the aversion to setting the line at 0 points (i.e. the intentional dip at 0), as this is equivalent to simply picking the winner outright.

The average sportsbook line spreads and average moneyline boxplots further supports the separate perception of home vs away odds:





In addition, upon examination of the matchup data we found that the data are mostly pristine, with the only missing data identified as the April 16, 2013 playoff game between Boston Celtics and Indiana Pacers, which was cancelled in light of the Monday bombings during the Boston Marathon. As a result, those two data entries are dropped from this dataset. 6 additional data entries are dropped as they contain All-star games statistics which we feel contribute no value to the objectives of this project.

The second file contains wagering data for each match up including the location, point spread, moneyline, and over/under offered at opening and at tip off of each game from five separate online sportsbooks. The third file is similar to the first one, but only contains data for the playoff matchups. As mentioned earlier, we have a separate datasource that pulls in live information concerning contests currently in progress that we will be applying our model to and sending us alerts when our model predicts wagering opportunities to be exploited. When the datasets are cleaned, concatenated, and normalised our data set for training the model includes 8604 records, each representing the results of a game played.

Description of Transformation of Data

The team has decided to approach the data transformation, model development and analysis of this project from two different angles.

In our first approach, game momentum data are engineered as the predictive model features. The objective for the model is to predict the final score of both home and away teams. This approach further analyze live game odds data to boost model confidence.

In our second approach, cumulative average team statistics and matchup statistics are engineered as the predictive model features. Both regression and classification models are explored to predict 1) game spread, 2) whether or not a game can cover the spread, 3) whether or not the total game will score over/under the preset

threshold, and 4) which team will win outright. This approach is more comprehensive in comparing different model types and analyzing feature importance.

Approach 1

Since each record in our initial dataset is for a team for a game, and a pair of teams participating in a game, we have to join the box score and gambling datasets by identifiers in each that would allow us to identify two records as opposing teams in the same matchup. We do this by separating the box scores dataset into separate dataset for home and away teams, then joining them on fields (date & location of game) and joining the datasets on those fields. We then create momentum features for each team's offensive and defensive statistics and gambling trends (ie points scored in last 5, 10, 15 matchups, turnovers averaged last 5, 10, 15 matchups, 3 point percentage last 5, 10, 15 matchups, etc.) as well as similar wagering metrics like teams point spread for the last 5, 10, and 15 contests. We also create feature for each of the same statistics for each home teams last 5, 10, and 15 games at home, and each away teams last 5, 10, and 15 games while away to help us identify teams that may benefit or be harmed in an anomalous way while playing at home or on the road.

Approach 2

Feature Separation and Importance

The feature engineering endeavor is driven by our desire to find out "which statistics are the most important in predicting X result for an NBA game?" The easiest and most direct way to do this is to use a model which has a feature importance attribute. Decision tree methods do, and so served as the model backbone of this project. However, if any two features are closely related, the importance for either one will be largely negated by the existence of the other. For example, say we have the two following statistics in this database: assists in the first three quarters and assists for the entire game. When the stat for the entire game is removed and the model measures how accurate its predictions are, it will still have a feature included that provides three-quarters of the removed entire game stat's information, and the prediction accuracy won't be severely impacted. As a result, the assists for the entire game statistic will be reported as not being a very important feature. However, the reality is that this would indeed be an important statistic, but having a correlated or, in this case, partially duplicated feature clouds our ability to determine its true importance.

Because of this fact, the feature set are split into three parts:

1. Raw statistics
2. Per-game statistics
3. Matchup deltas

Raw statistics are simply every team's statistics coming into the given game of interest. Per-game statistics are cumulative averages of raw statistics for each of 30 teams in the particular regular season progressing as games take place. As for Matchup deltas, these are the differences between the away and home teams in these respective per-game statistics. For example, the home team might average seven more assists per game, but two more turnovers per game. Since the matchup stats are technically derived from the raw statistics, proper

evaluation of feature importances is key to ensure them being optionally separated. Table 2 captures the engineered statistics features - per game statistics and matchup deltas:

Table 2

Per-game variables	Description	Matchup variables	Description
meanFG_PCT	Cumulative average field goal percentage (per team per season)	deltaFG_PCT	Difference in meanFG_PCT between opposing teams
meanFT_PCT	Cumulative average free throw percentage (per team per season)	deltaFT_PCT	Difference in meanFT_PCT between opposing teams
meanFG3_PCT	Cumulative average three point percentage (per team per season)	deltaFG3_PCT	Difference in meanFG3_PCT between opposing teams
meanAST	Cumulative average number of assists (per team per season)	deltaAST	Difference in meanAST between opposing teams
meanREB	Cumulative average number of defensive and offensive rebounds (per game per season)	deltaREB	Difference in meanREB between opposing teams
meanTOV	Cumulative average number of turnovers (per game per season)	deltaTOV	Difference in meanTOV between opposing teams

Furthermore, the classifier model targets have been comprehended in the following manner:

"Cover" = Team score margin + Books line spread > 0
Explanation: Dummy coded. A point spread bet is won (i.e. covered) = 1 whether betting the favorite or underdog. Not covered = 0

"Over" = Total game score > Books over/under line
Explanation: Dummy coded. An over bet is won when the target = 1. An under bet is won when the target = 0.

"Result" = Winner or loser
Explanation: Dummy coded. A moneyline bet is won when the target = 1.

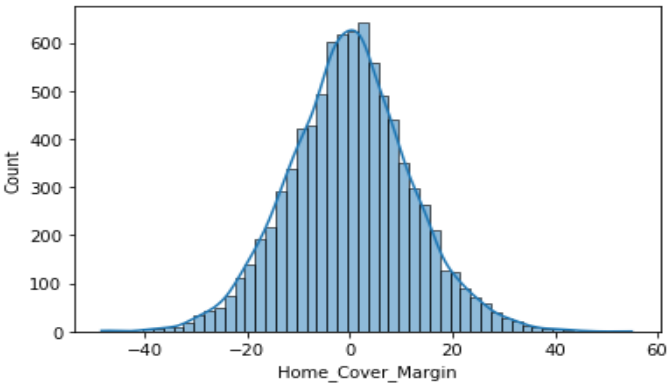
Prior to model input preprocessing, both NBA game data and betting data were joined together and had nulls dropped (5 out of 17207 data entries). Raw statistics, features that are not available pre-game such as quarter and final scores, as well as features that are arbitrary were dropped from the merged dataset.

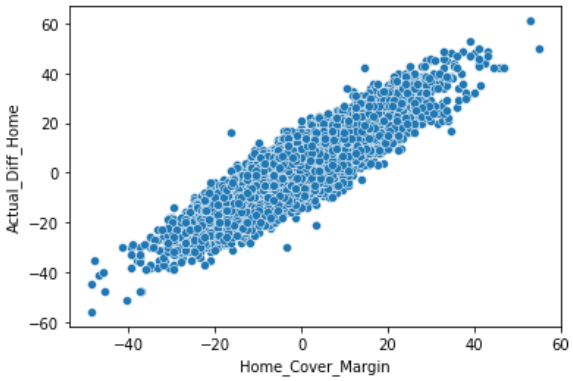
PREDICTIVE ANALYTICS

Analysis of Data

Approach 1

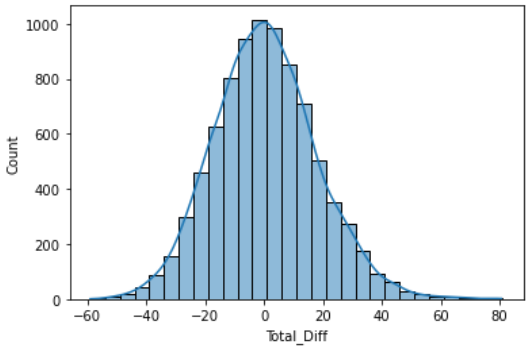
A central tenet of our thesis is that over the long run the gambling lines at game time are accurate, and over the long run, no matter what your strategy was in picking games, you would win roughly 50% of your picks and lose 50%, and the casinos additional 10% tax on losing bets gives them an edge of roughly 52.5% to 47.5% in their favour (ie if you bet a dollar on a thousand games you would win $500 and lose $550 putting your return at -$50/$1000 or -5%). Our idea is that with the advent of sports betting applications that allow you to bet from home and while the game is taking place, you can wager on lines that have deviated from the opening line in a beneficial way, and by default since you are getting a more favourable line than the opening one where one would win roughly 50% of the time, you can only improve your results. For example, if wagering on an underdog team while they are +2 has a win probability of 50%, waiting until the game commences and betting the same team +3 or + 4 can only improve your win probability. Clearly this idea is only valid if you can expect to win 50% of your wagers. To test this we programmatically created an array that assigned a wager on each game to either the home team OR the away team and the point total over OR the point total under for each game over the course of the 7 seasons in our dataset. This program assigned 17,208 wagers at random and it picked 8,629 winners and 8,579 losers for a winning percentage of 50.15%. Pretty close to our 50% hypothesis, so assuming the biases of a human do not cause dramatic underperformance to randomly assigned picks we are in good shape. Secondly, we need to verify that the point spreads available at the start of the game are reasonably predictive of the final outcome. If the point spread is -2 at the time of the start of a given contest, but the favoured team wins by 20, waiting for an extra few points deviation from the opening spread wouldn't necessarily present you with an advantage nor an ability to improve on your 50/50 baseline number. To do this we compared the difference between the winning margin of the favoured team and the opening point spread, in the hope the differences would have a mean of 0 and a relatively normal shape. We were not disappointed:

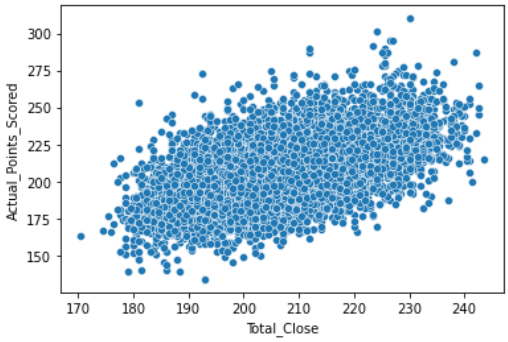




The average difference between the outcome and the point spread was roughly one tenth of one point over all games from those 7 seasons and the point spread was a remarkable predictor for the actual outcome (the correlation coefficient between point spread and game outcome over this time frame was .872). This again bodes well for our hypothesis, and we now know that a) a 50% baseline winning percentage is a reasonable

assumption for a betting strategy, and that b) the outcomes of the wagers are indeed sensitive to the opening line and it may be possible to capitalise on deviations from these spreads. Point total distributions were also distributed very normally, but there are outliers in our dataset regarding the point total data that warrant further investigation before model development:





The difference between the final point totals and the pregame point total mirrored our findings for the lines, with the difference being normally distributed with a mean of less than one half of one point (.464), suggesting that the point totals bet outcomes are also sensitive to small deviations from the opening point total line, and we should be able to capitalise on deviations from the opening point total via in game betting lines, with the caveat that the total points scored and the pregame total line have a significantly lower correlation coefficient (.558) than the the point spread and game result, and a higher standard deviation (17.58 points).

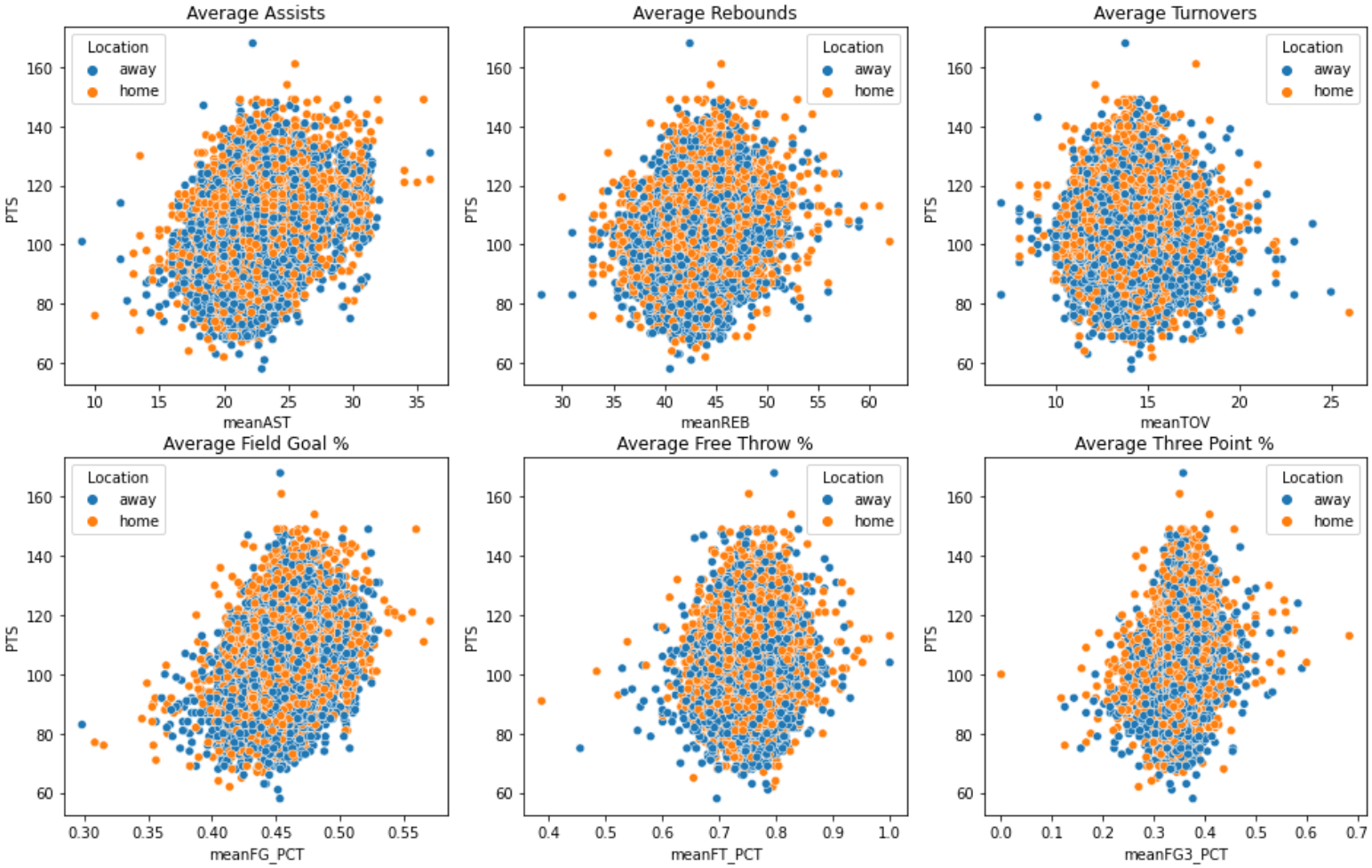
One potential source of concern is that a standard deviation of the difference between the outcome of a contest and the point spread is 12.02 points. Our aim is to capitalise on small deviations from the opening line, and having to wait for such large deviations from the opening line would either a)not occur often enough to make us consistently profitable or b) put us into a disproportionate amount of losing bets since a team would have to really be underperforming to warrant a sportsbook moving from the opening line that much. The standard deviation of the difference between the actual points scored was also quite high (19.632), but this number should be taken with a grain of salt until we can properly remediate these outlier values. The standard deviation will certainly be lower upon doing so.

Of the 8604 games in the dataset 12.85% (1106), the outcome of the game was determined by less than 2 points from the opening spread and an additional 3.67% (316) were determined by exactly 2 points from the opening spread, meaning that if we are able to wait until the game opens and wager on the in game spread once it moves two points, which anecdotally seems to happen in the vast majority of games, you would win your wager regardless of which of the two teams you selected in those 1106 contests.

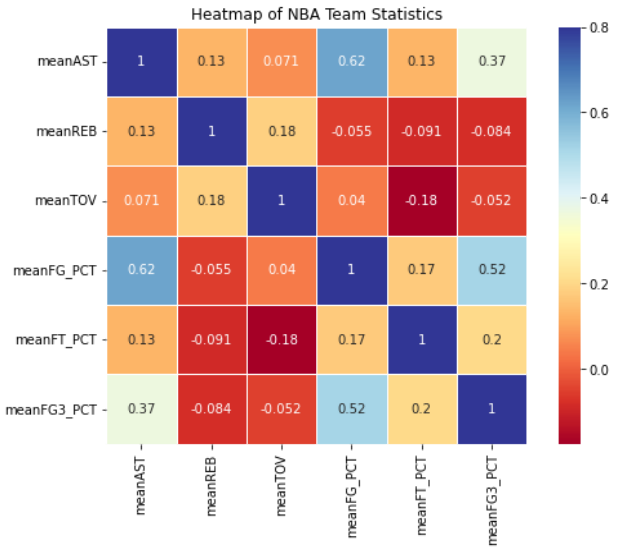
To tie this back into the random bet making program we discussed earlier, assume now that it made all the same wagers it did the prior iteration, but made them in game when the line had deviated 2 points in its favour from the open. If we reasonably assume that half (553) of those 1106 games it had picked ended up as winning wagers at the original line, and half as losing wagers, all of the losing wagers now become winning wagers and its performance moves from 8629-8579 (50.15% win rate), to 9182-8026 (53.36% win rate). That increase clears the profitability break even win rate of 52.5% and would give the firm an expected value of slightly over .02c return per dollar wagered. By the same logic if you waited for 3 additional points to the opening spread the win rate would rise to 55.23% expected return on a dollar wagered would nearly triple to .059c and you would have over double the edge on the house than they currently enjoy over their customers (10.46% vs 5%, assuming a 10% vig).

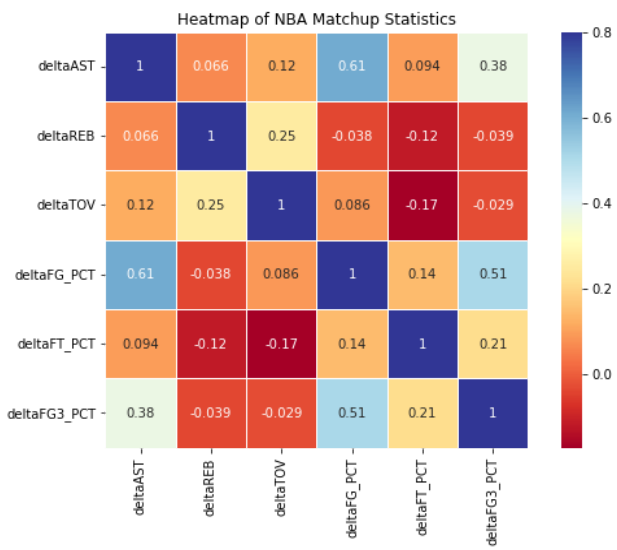
Approach 2

A preliminary examination of the relationship between final team score and each per-game statistics is shown below:



A very minor discernible difference is observed between away team and home team clusters, and a rough positive correlation exists between final score and average assists/average field goal %/average three point %. The correlation heatmaps among the per-game statistics and matchup deltas further supports this observation as the averages and corresponding deltas of assists, field goal % and three point % display the strongest correlation.





The regression “Spread” target is summarized as follows:

Target	**Min. (absolute)**	**Max. (absolute)**	**Range**
Spread	1	61	60

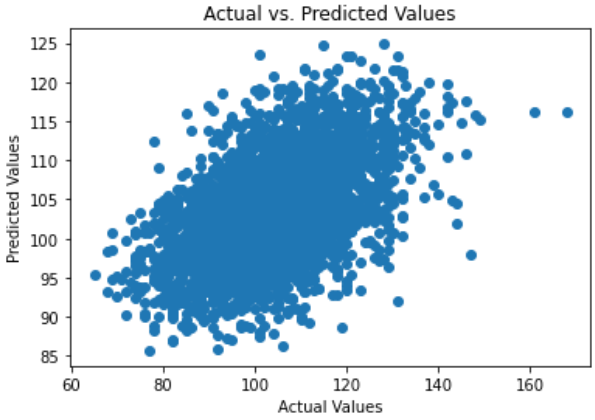
The classification targets are summarized as follows:

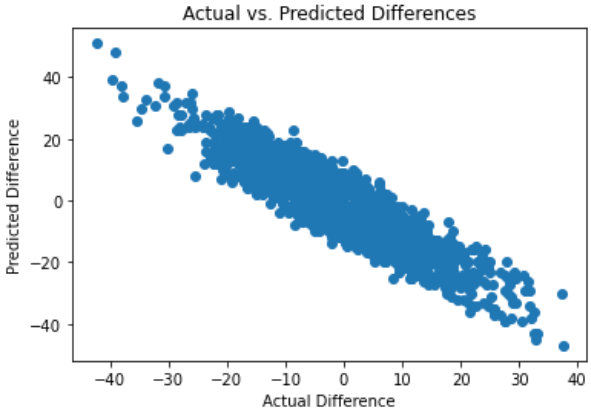
Target	**Majority Class**	**Minority Class**	**Majority %**	**Minority %**	**Counts**	**Total**
Cover	No	Yes	50.2%	49.8%	8572	17202
Over	No	Yes	50.1%	49.9%	4290	8601
Home team Win	Yes	No	58.7%	41.3%	5053	8601

Predictive Model and Results

Approach 1

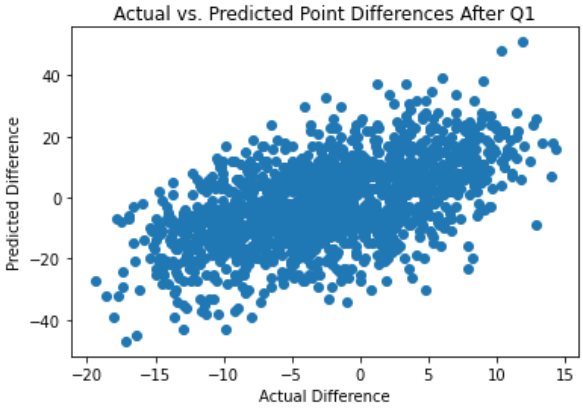
We fit several machine learning regression models to our data in an attempt to pinpoint the one that would deliver the best outcomes. We fit a multivariate linear regression model, a support vector regression, a random forest regression, and a gradient boosted regression to our dataset, and in the end the random forest regression model produced the best results. A random forest regressor is an ensemble machine learning algorithm that combines multiple decision trees to perform regression tasks. It works by using an ensemble of decision trees where each tree is trained on a random subset of the data. The randomness of the model is derived from both the sampling of data points and the selection of features considered at each split in the decision tree. A few benefits of using a random forest regressor for the task at hand is that these particular models are capable of capturing non linear relationships between features and the target, they can handle high dimensional datasets effectively (ours has over 165 features), and they are fairly easy to implement with libraries like Python's scikit-learn. Our model had two target variables (the home team final score and visiting team final score) and between 166 and 169 features depending on which point in the game you are running it (ie points scored in the first quarter would be a feature when it became a known, after the first quarter ended, but prior to the game it would not be used as a feature for obvious reasons). If our predictive model proved to be accurate, that is more accurate than the book maker's line, we plan to capitalise on it by running it throughout the game and sending a push alert to users notifying them that there is a deviation between our predictive score and the oddsmakers line. When running the predictive model on the features we engineered prior to the game starting, the results upon first glance appear to be unimpressive with an R^2 of only .258 between our predicted scores for the home and away team vs their actual final scores:

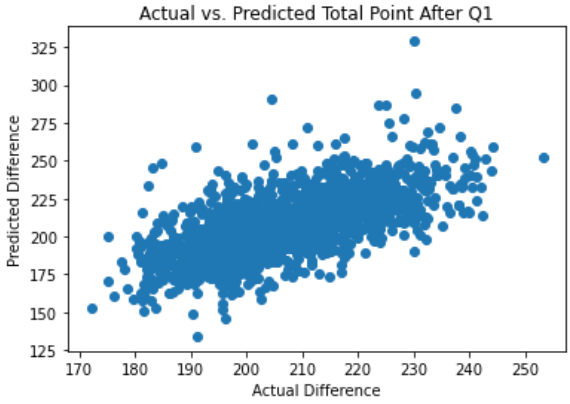




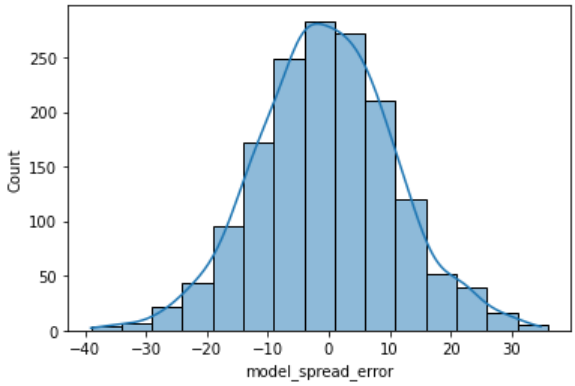
But there is a very important distinction to note. Odds makers before the game are only trying to predict the difference in the opposing teams final score and the total of the two teams scores, while our model is trying to take on the more difficult task of predicting the exact number of points each team will score. The graphic on the left shows a scatterplot of our models predicted points scored for each team run prior to the start of the game, which doesn't appear to extremely accurate (and the R^2 verifies this), but when viewed as the difference between our projected points scored vs the final score (the graphic on the right), our model provides a more accurate prediction of the final point differential than the actual spread (it is worth noting that the actual spread was a feature in our model, as it should be since it widely available before the game). Running our model prior to the start of the game gives an average difference between our predictions and the actual final score of .464 points with a standard deviation of 11.95 points. The pregame spread had a mean difference of .528 points with a standard deviation of 11.89 points. Not a significant enough gap for us to reliably profit

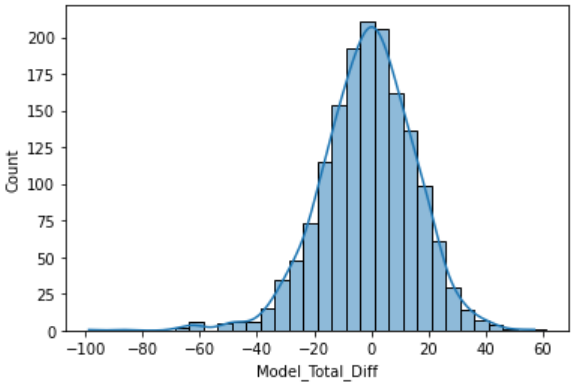
from, but worth noting that even before the games start, using a random forest regressor on the features we engineered gives a slightly more accurate prediction of the outcome of the game than the Las Vegas spread does. The same can be said for our model's predictions about total points scored in the game. Our model was more able to more accurately predict the total number with a difference an average difference of 1.15 points between our model's prediction and the actual result (with a standard deviation of 18.18 points) vs the pregame total's difference of 1.53 points and a standard deviation of 17.97 points. Again, not a significant enough margin to consistently profit from, but we were able to reduce the average error of the pregame total by by 24.8% at the expense of slightly higher variability. The results are even more encouraging when we use our model to start making predictions once the game has already started and we can take advantage of the fact that we are now legally able to place wagers intergame. The R^2 on the test set in predicting the amount of points each team will score increases from .258 to .378 and the root mean squared error decreases by about a point from 10.89 points to 9.97 points. The relationship between our predictions for both the spread and total points scored, and the actual results starts appearing more linear in nature:





While retaining the shape of distributions that are fairly normal:





The predictive value of our model at this point in the game is still better than that of the vegas line, but on average not by a wide enough margin for us to get above the 52.5% break even threshold reliably, although at this point in the game our predictive model still gives us a better prediction of the final point difference and final total amount of points scored than the vegas line, and at this point has a lower standard deviation for for both classes of wagers. By the end of the second quarter our model continues to improve on the accuracy of its predictions both for the spread (.448 points mean difference from actual result, with a standard deviation of 9.72) and the total (.47 points mean difference from final total with a standard deviation of 14.98). Our final run of our model occurs at the end of the 3rd quarter. At this point, as you would expect the model improves

again, and at this point the R^2 of the predicted scores of each team vs their actual results is up to .7198 and the average predicted spread error is reduced to .12 points. A ~77% reduction in the error of the pregame spread vs actual final spread, with a standard deviation of 7.31 points, 38.5% improvement on the pregame number. The model is even better at predicting point totals than point differentials at this point as well. It predicts the final point total in our test set within .03 points of the actual point total (a 93% improvement on the pregame point total error) with a standard deviation of 11.81 (a 34% improvement from the opening total standard deviation). The model gets more and more accurate as the game goes along, and will provide us with opportunity to place wagers via the live lines with a much higher degree of accuracy as the games progress.

Approach 2

Regression model selection and training

In order to choose the models which performed best the regression models were measured for the mean absolute error (MAE). Compared to the root mean squared error (RMSE), the MAE is consistent across ranges of errors and doesn't 'flare' up in response to larger residuals. For evaluating how many points a predicted NBA game's spread is from the actual spread, there is no disproportionately harsher penalty for being five points away than there is for being four points away. Using the absolute error ensures an easily interpretable metric for evaluating model accuracy with this data: a MAE of x means we have an average error of x points.

The two regression models trained and tested were a Random Forest, a Support Vector Machine (Epsilon-Support Vector Regression), and Gradient Boosting Tree (eXtreme Gradient Boosting). All models were wrapped in a SKlearn pipeline preceded by a One hot encoder for categorical variable preprocessing and min-max scaling. The dataset was split into 80:20 between training and test subset before passinging into the tree models pipeline. On the contrary, the SVM pipeline underwent a ten-fold cross-validation with reshuffling.

For the prediction of point spreads, both the RF and SVM models did not yield a promising result at its current state without further hyperparameter tuning. RF was the better model with a lower MAE and a higher R^2 value. The outcome are presented in the table below:

Regression outcome

Target	**Model**	**Metrics**	**Score**
Point Spread	RF	MAE R^2	7.94 0.383
Point Spread	SVM	MAE R^2	9.18 (mean) 0.268 (mean)
Point Spread	XGB	MAE R^2	8.76 0.295

At best the RF model can predict the game spread to within 7.94 points but it is beyond the average sportsbooks published line spread for both betting with or against the spread. This does not strike us as a long term profitable strategy.

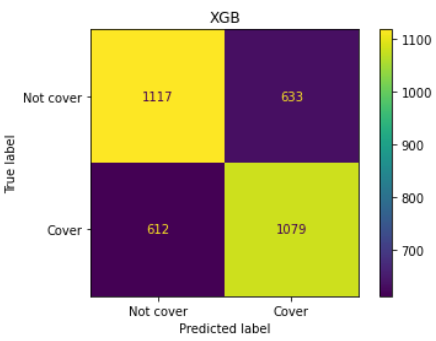
Classification model selection and training

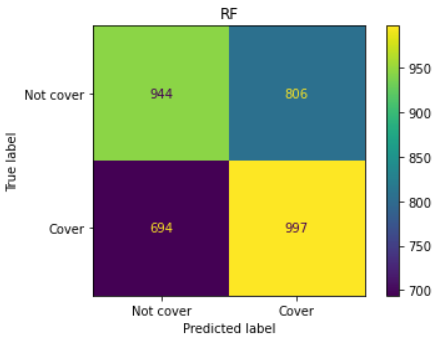
Three models were used for each classification target: Gradient Boosting Tree (eXtreme Gradient Boosting), Random Forest, and Random forest embedding paired with logistic regression (drawn from Feature transformations with ensembles of trees). By setting a medium tree depth (6) and high number of trees (500), the best performing model in all classification tasks was the XGB. Out of all classification targets, the XGB model has outperformed the other models. To our surprise, the XGB model can attain an accuracy well exceeding 60-70% for each of the classification target, which presents itself as an opportunity for us to capitalize on the long term. By weighing the percentage of winning bets against the odds, we can further calculate the expected profit value. The outcome are summarized in the table below:

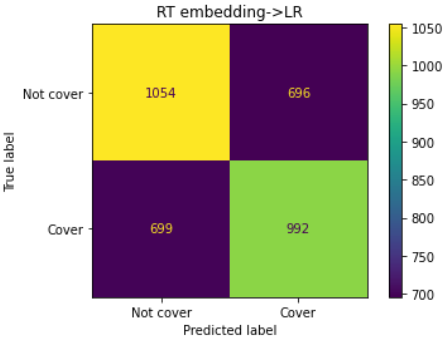
Classification outcome

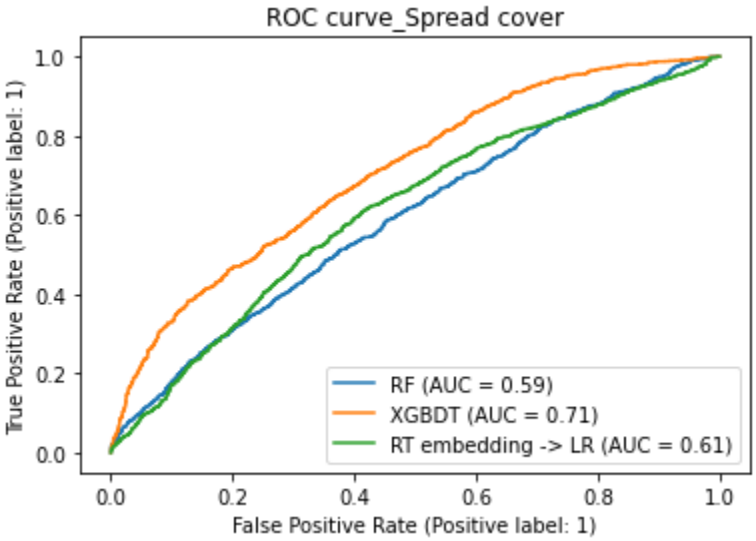
Target	**Model**	**Metrics**	**Score**
Cover Spread	XGB	Accuracy AUC	63.82% 0.71
Cover Spread	RF	Accuracy AUC	56.41% 0.59
Cover Spread	RT + LR	Accuracy AUC	59.46% 0.61
Win/Loss	XGB	Accuracy AUC	74.69% 0.83
Win/Loss	RF	Accuracy AUC	71.03% 0.78
Win/Loss	RT + LR	Accuracy AUC	69.20% 0.75
Over/Under result	XGB	Accuracy AUC	64.25% 0.72
Over/Under result	RF	Accuracy AUC	54.90% 0.58
Over/Under result	RT + LR	Accuracy AUC	59.37% 0.62

Model results - Cover Spread

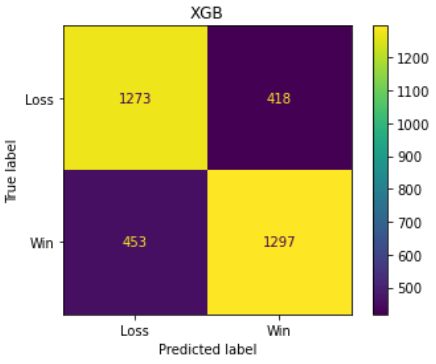


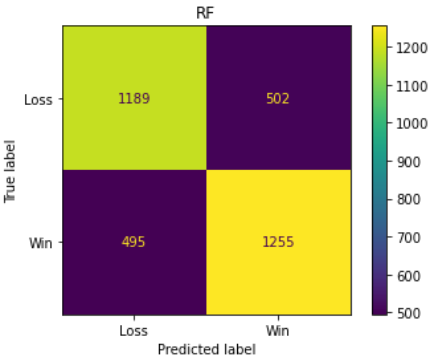


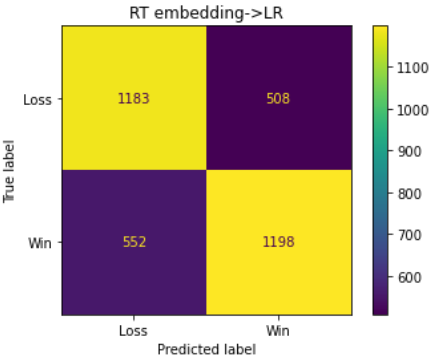


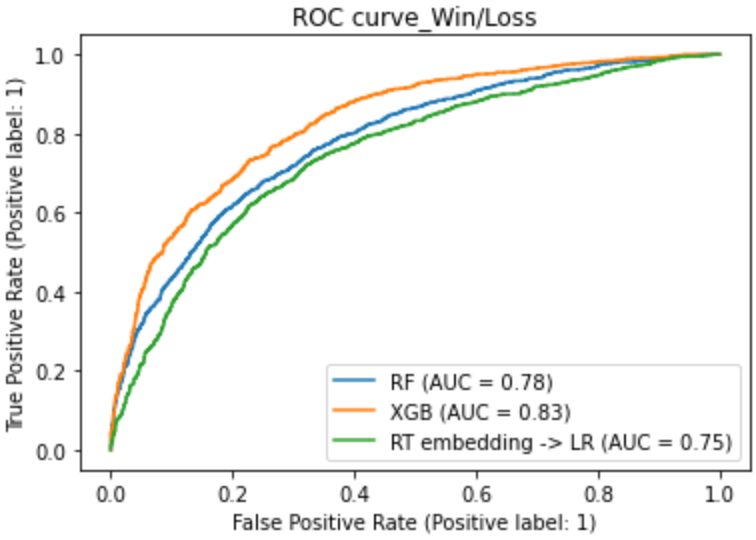


Model results - Win/Loss

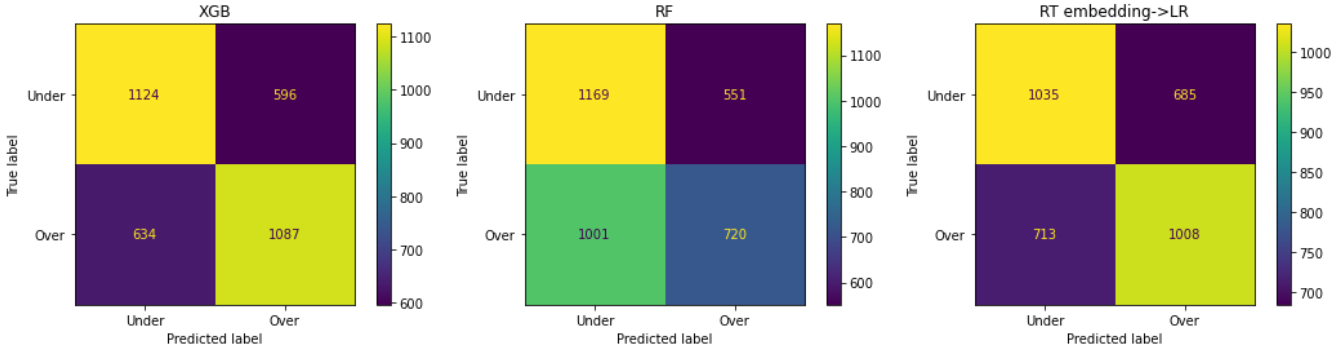


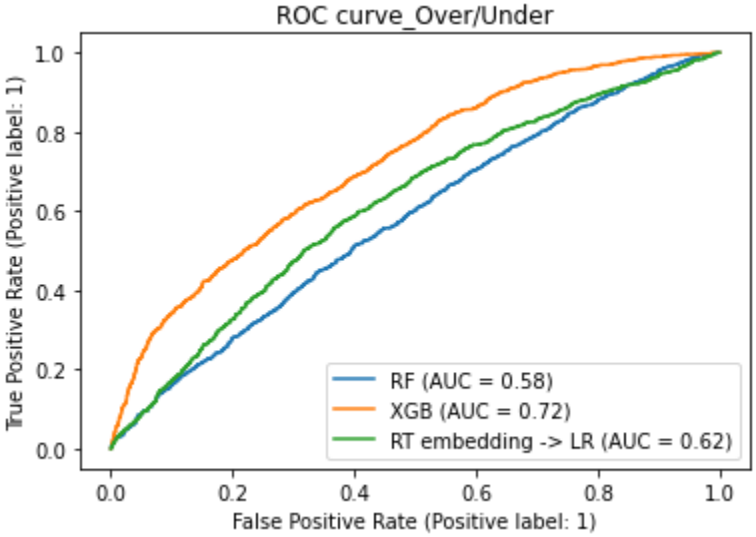






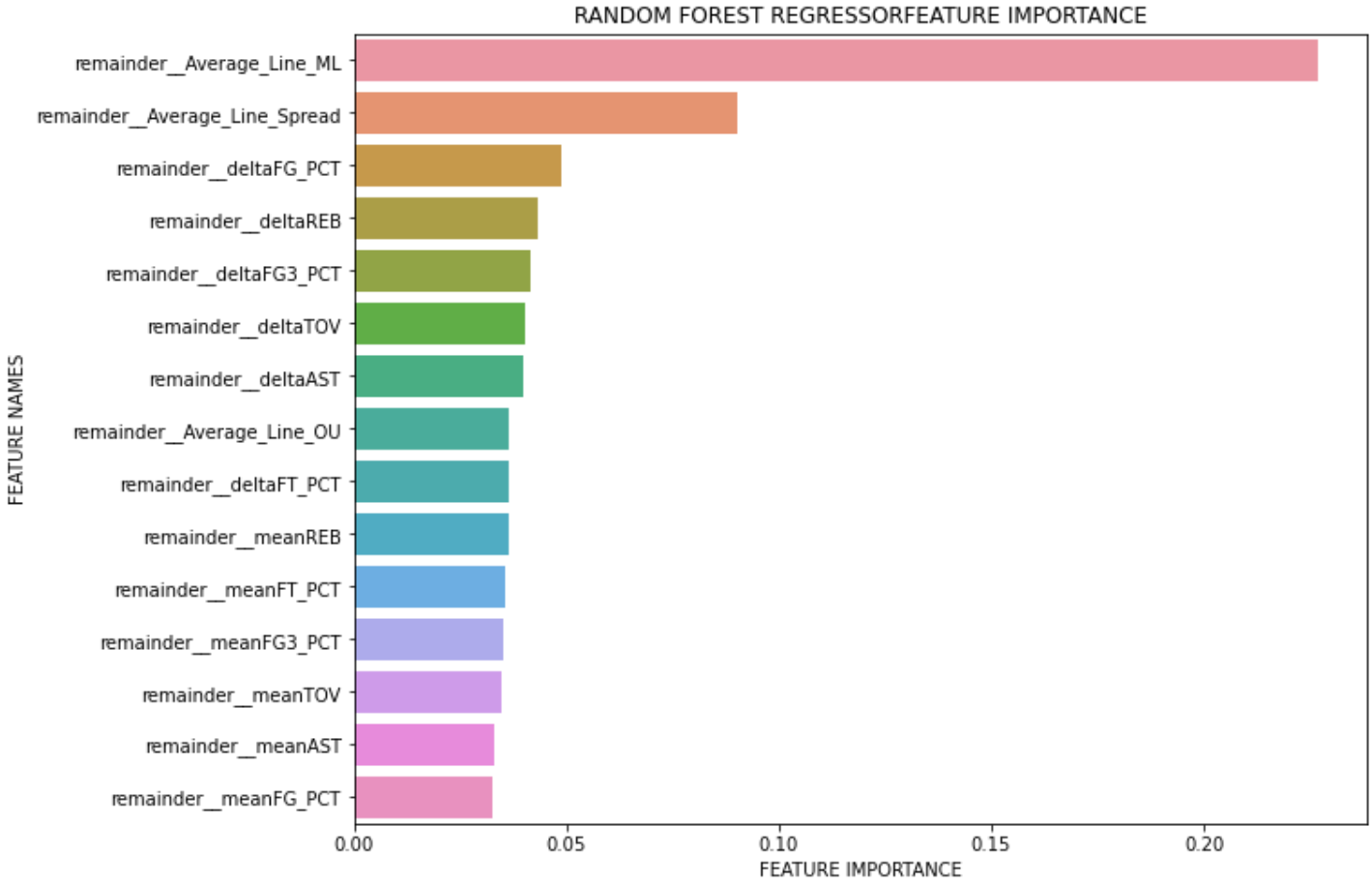
Model results - Over/Under





Regression Feature Importance

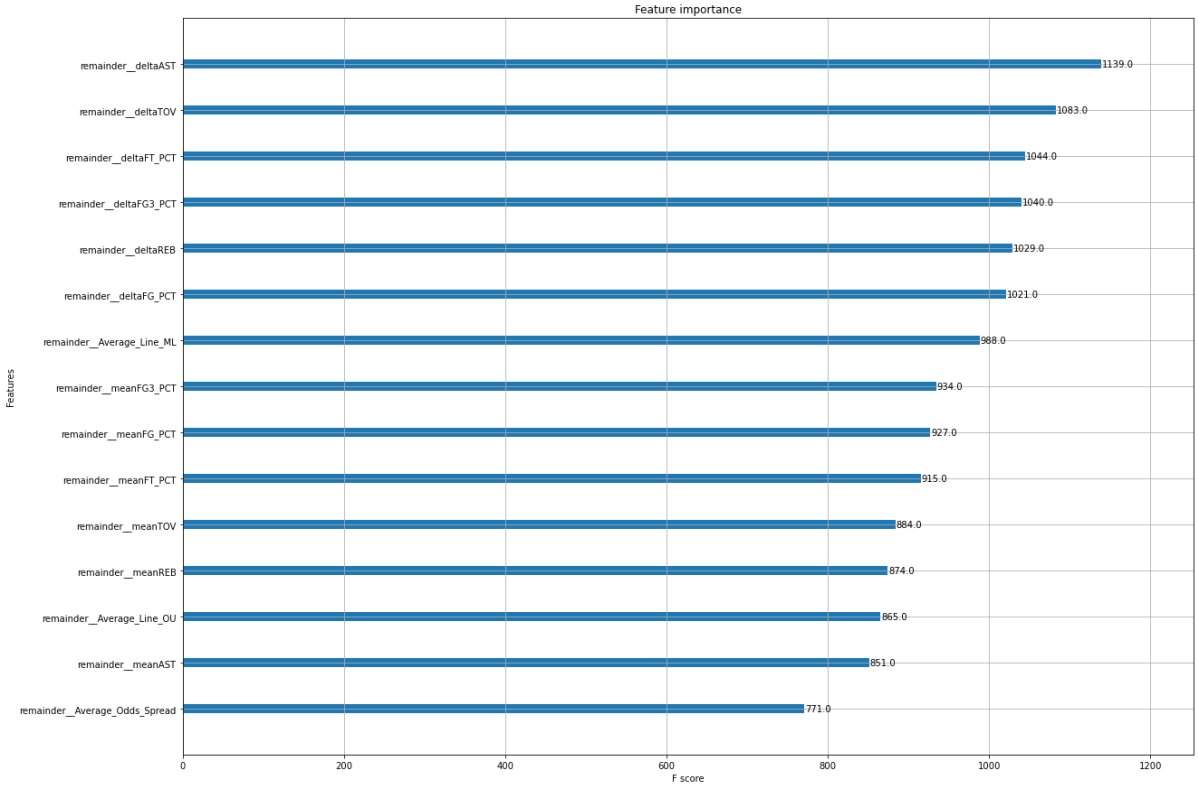
Mapping the top 15 features of the better performing RF model, the most important features are found to be the sportsbooks pre-game published moneyline and line spread. Given resources of sportsbooks to leverage supercomputers in determining the optimal line and odds to maintain balanced bets on both sides, this is not a surprising result at all. Note that all the matchup deltas outrank the per-game statistics as model features. These tell the difference between the two teams in a game in a given metric. A logical reasoning was that the raw value of a statistic would matter little for prediction of one team's superiority if the other team also had a high value in that metric. If one team was significantly better than the other team in a given metric we would be able to make more accurate predictions. This finding is shown consistently in this project as the matchup features rank high in importance.



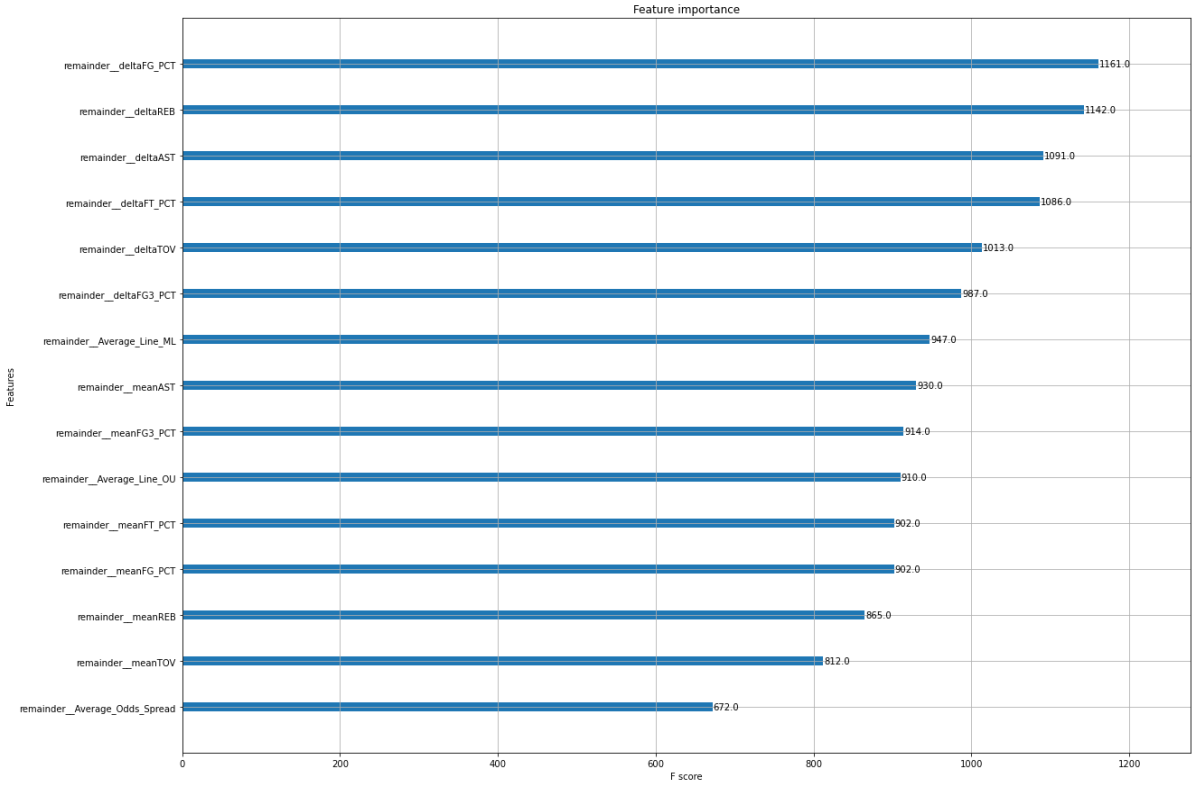
Classification Feature Importance

Similarly the top 15 features are mapped for each of the best XGB classifiers. The most important features in weight, which is the number of times a feature appears in a tree, are illustrated in the charts below. Again the top features appear to be either the matchup deltas or the sportsbook published lines.

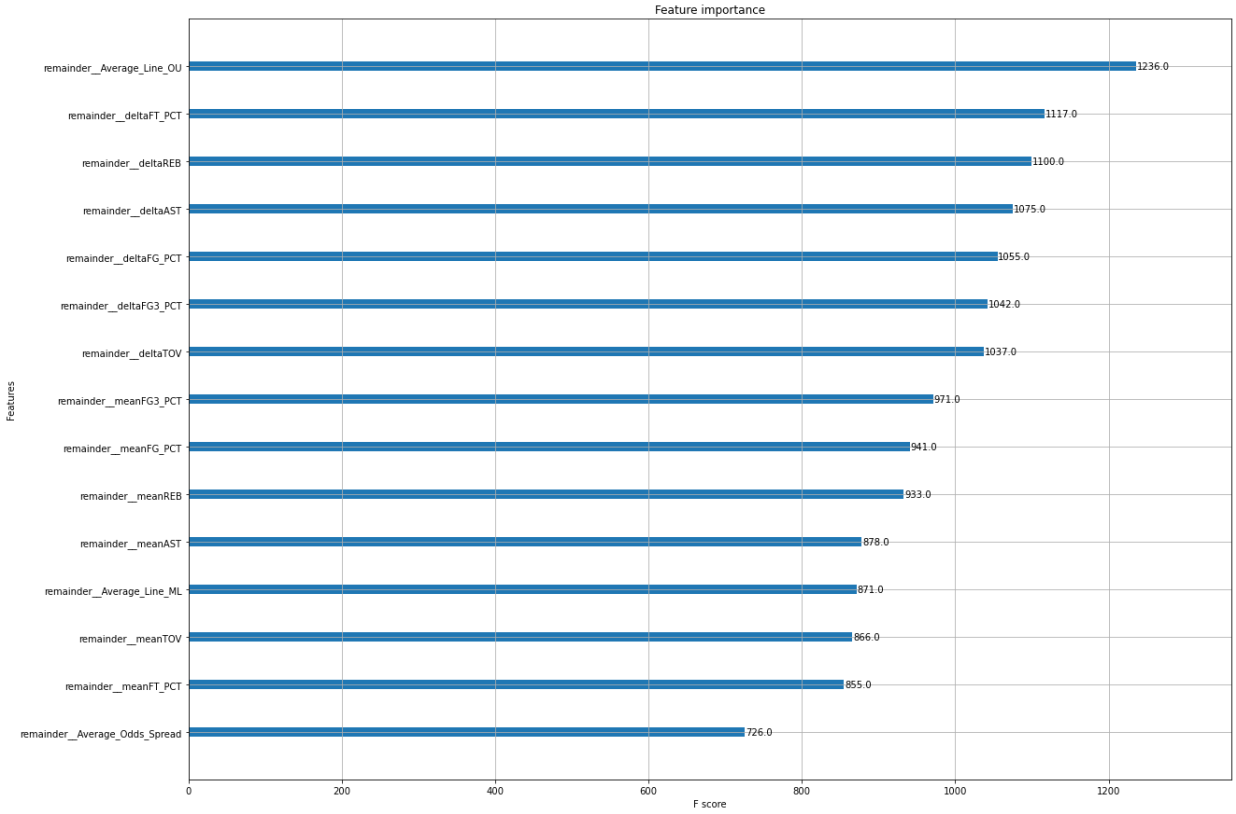
Cover Spread XGB model



Win/Loss XGB model



Over/Under XGB model



Conclusions

Our initial analysis of the data validates that our strategy to profitability is supported by the historical data. The opening point spreads are near perfect predictors of the final results of a game, and even small deviations from them via the in game betting lines, which have only become available recently due to the legalisation of mobile sports gambling (in 26 states) with the advent of sports wagering applications present the opportunity to capture significant theoretical edge over the house. We have also created an abundance of features to train our model on that will help us identify the most lucrative opportunities.

All in all, our team has demonstrated from our two-prong approach that there are viable prediction pathways to devise a profitable strategy to partake in the most popular betting types in NBA sports wagering. For the purpose of a proof of concept and prototype creation, we have achieved what we have originally outlined in our project charter.

The immediate next steps will be integrating the historical in-game odds data to our dataset and integrating the prediction model of choice to the front end live odds application. Short terms goals to improve our analysis will be to expand the dataset to capture both older and more recent NBA seasons data, plus building a model for playoffs as our current focus is on regular season. From the modelling perspective we have the opportunity to conduct further data manipulation and model hyperparameter tuning to strengthen the model performances.

From Analytics to Prototype

With some promising results from our team's modelling efforts, we have established a solid foundation. Even though we've been "burning money at an alarming rate" the model's results suggest a substantial improvement in win probability through the moving in-game odds, which should help ease some of the concerns from the Finance and Accounting teams. Gaining a better understanding on why data analytics projects fail, has helped us focus on success. As such, we've clearly defined what our expected final product will be - a working prototype that can be deployed into production, that specifically focuses on NBA Point Spread betting.

Now that we have completed the predictive modelling, compared alternatives, and evaluated, configured, and tuned the most optimal model, we move our focus into building out the prototype proof of concept (PoC) product, which connects the back-end (data analysis/models) to the front-end (user interface).

PROOF OF CONCEPT

Functional Requirements

We started off by first identifying the functional requirements for the pilot.

- Must be able to ingest the output from our Python predictive model
- Must be able to fetch live, in-game odds
- Must be able to work with NBA Points Spread
- Must be able to access the live odds from MGMSports
- Must be able to dynamically refresh and update into Domo
- Must have a live connection into Domo
- Must have a Domo dashboard
- Must have a triggered alerting mechanism for SMS, app pushes, and emails

Design Considerations

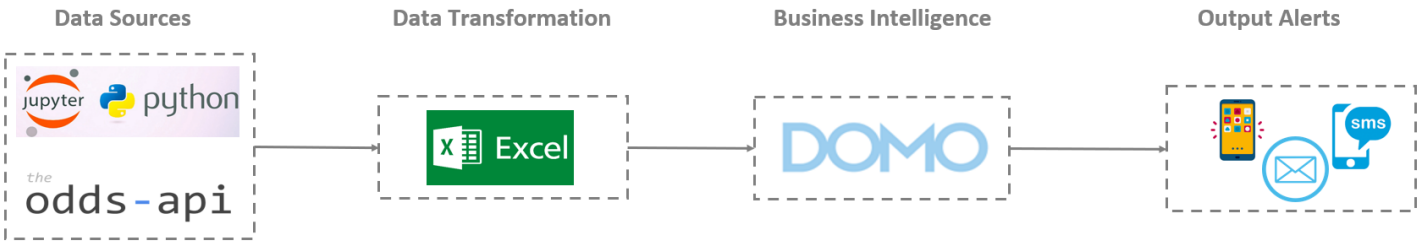
For the pilot, we needed 4 distinct components - Live odds, a way to translate the data, a cloud data repository, and an alerting mechanism. The table below outlines the selected tools and reasons.

Tool	Selected Vendor	Competitors Considered	Reasons for Selection
Live in-game Odds	the odds-api	• OddsJam • APILayer • SportRadar	• They offered a free plan that allowed 500 calls per month (however a Starter plan was eventually purchased) • They provided better support through online tutorials • They had recently released both a Python and R library, so they offered better scalability • They had a working Excel add-in which could be leveraged for the prototype
Data Extraction and Transformation	Excel	• R Studio • Python • PowerBI	• Considering the Odd API used MS Excel, it made more sense to stay in Excel • Future state will migrate into R/Python

Cloud Data Repository	DOMO	• Tableau • Azure	• Free version of Domo was more accessible • Smaller learning curve • Direct data connectors for MS Excel
Alerting Mechanism	DOMO	• R/Twilio • Python/SMS	• Competitors were very limited • Solutions through R and Python had limited volumes per month • Domo's mobile app also allowed for push notifications, in addition to SMS Texts and Emails

End-to-End Demonstration

Here is a visual representation of the systems architecture:



Step 1: Data Sources

The process starts with our Python predictive model. We are currently evaluating the following 9 scenarios:

- all_spread_bet
- home_dogs
- home_fav_between_zero_and_five
- home_fav_between_five_and_ten
- home_fav_ten_plus
- road_favorites
- road_dog_zero_to_five
- road_dog_five_to_ten
- road_dog_ten_plus

Then these scenarios are evaluated across the following 10 changes:

- standard_spread
- plus_one_live_points
- plus_two_live_points
- plus_three_live_points
- plus_four_live_points
- plus_five_live_points
- plus_six_live_points
- plus_seven_live_points
- plus_eight_live_points
- plus_nine_live_points

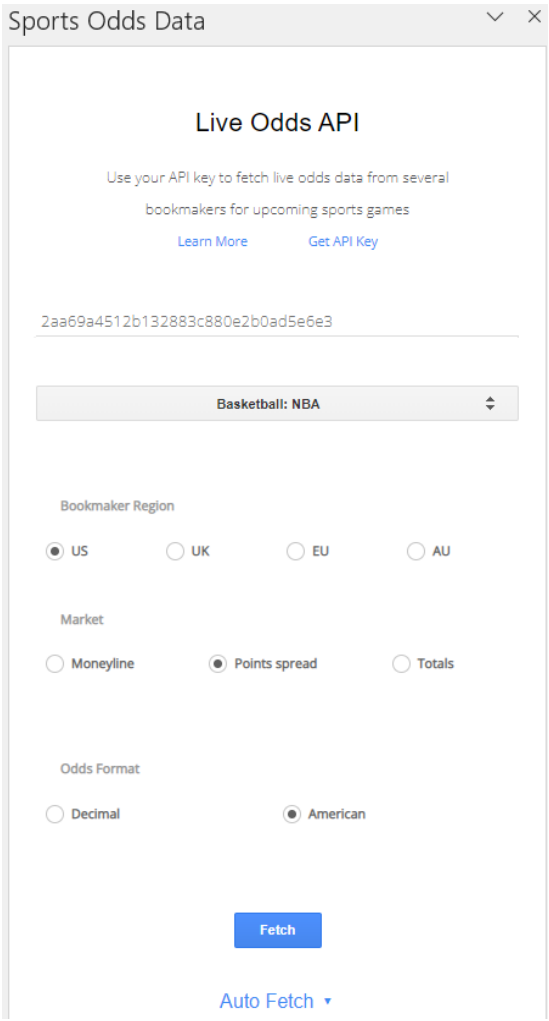
We ultimately get the following data matrix output from our Python model:

Criteria	standard_spread	plus_one_live_points	plus_two_live_points	plus_three_live_points	plus_four_live_points	plus_five_live_points	plus_six_live_points	plus_seven_live_points	plus_eight_live_points	plus_nine_live_points
2012_all_spread_bet	0.50	0.53	0.58	0.61	0.64	0.67	0.69	0.73	0.75	0.78
2012_home_dogs	0.48	0.50	0.56	0.63	0.68	0.70	0.73	0.75	0.78	0.80
2012_home_fav_between_zero_and_five	0.48	0.52	0.57	0.57	0.59	0.61	0.63	0.68	0.72	0.75
2012_home_fav_between_five_and_ten	0.51	0.59	0.61	0.61	0.63	0.73	0.76	0.80	0.83	0.88
2012_home_fav_ten_plus	0.51	0.59	0.61	0.61	0.63	0.73	0.76	0.80	0.83	0.88
2012_road_favorites	0.51	0.53	0.57	0.59	0.60	0.61	0.63	0.65	0.69	0.71
2012_road_dog_zero_to_five	0.53	0.57	0.62	0.63	0.67	0.71	0.74	0.80	0.81	0.81
2012_road_dog_five_to_ten	0.48	0.50	0.55	0.59	0.61	0.66	0.69	0.72	0.75	0.79
2012_road_dog_ten_plus	0.52	0.55	0.59	0.68	0.71	0.71	0.74	0.77	0.78	0.81
2013_all_spread_bet	0.51	0.54	0.58	0.61	0.65	0.68	0.72	0.74	0.77	0.79
2013_home_dogs	0.53	0.55	0.57	0.60	0.66	0.69	0.73	0.77	0.80	0.82
2013_home_fav_between_zero_and_five	0.45	0.48	0.51	0.53	0.55	0.57	0.59	0.63	0.67	0.70
2013_home_fav_between_five_and_ten	0.44	0.45	0.49	0.56	0.62	0.68	0.73	0.75	0.78	0.81
2013_home_fav_ten_plus	0.44	0.45	0.49	0.56	0.62	0.68	0.73	0.75	0.78	0.81
2013_road_favorites	0.57	0.61	0.63	0.65	0.68	0.70	0.73	0.75	0.78	0.79
2013_road_dog_zero_to_five	0.53	0.57	[illegible]	0.66	0.71	0.74	0.77	0.79	[illegible]	0.86
[illegible]	[illegible]	[illegible]	0.52	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2017_home_fav_between_zero_and_five	0.54	0.59	0.63	0.64	0.66	0.66	0.69	0.71	0.75	0.77
2017_home_fav_between_five_and_ten	0.39	0.45	0.51	0.54	0.62	0.68	0.71	0.77	0.82	0.84
2017_home_fav_ten_plus	0.39	0.45	0.51	0.54	0.62	0.68	0.71	0.77	0.82	0.84
2017_road_favorites	0.49	0.53	0.56	0.60	0.62	0.66	0.69	0.72	0.75	0.78
2017_road_dog_zero_to_five	0.47	0.54	0.58	0.63	0.65	0.67	0.69	0.72	0.76	0.76
2017_road_dog_five_to_ten	0.53	0.56	0.58	0.61	0.67	0.71	0.72	0.74	0.75	0.77
2017_road_dog_ten_plus	0.57	0.64	0.68	0.70	0.77	0.80	0.83	0.85	0.86	0.86
2018_all_spread_bet	0.48	0.50	0.54	0.57	0.60	0.63	0.66	0.69	0.72	0.75
2018_home_dogs	0.48	0.50	0.53	0.56	0.59	0.63	0.68	0.69	0.72	0.76
2018_home_fav_between_zero_and_five	0.50	0.51	0.54	0.55	0.57	0.58	0.64	0.66	0.69	0.74
2018_home_fav_between_five_and_ten	0.49	0.50	0.57	0.60	0.63	0.66	0.71	0.76	0.76	0.77
2018_home_fav_ten_plus	0.49	0.50	0.57	0.60	0.63	0.66	0.71	0.76	0.76	0.77
2018_road_favorites	0.48	0.52	0.52	0.56	0.58	0.62	0.64	0.69	0.72	0.75
2018_road_dog_zero_to_five	0.46	0.48	0.52	0.56	0.60	0.64	0.67	0.70	0.75	0.78
2018_road_dog_five_to_ten	0.48	0.51	0.56	0.60	0.62	0.63	0.66	0.69	0.71	0.74
2018_road_dog_ten_plus	0.51	0.52	0.53	0.54	0.60	0.64	0.68	0.70	0.72	0.74

It contains the 9 Criteria, per year, and the updated win probabilities based on each of the 10 potential changes.
Once this is output from Python, we load it into MS Excel, thus completing the first step for our prototype.

Step 2: Data Transformation

At this step, we move into MS Excel. We have already downloaded, installed, and configured the Live-Odds-API excel add-on. At the right, is a screenshot of its UI.



Once a unique API Key has been procured, the add-on allows you to fetch odds for dozens of sports, including all 4 major professional US sports (NBA, NFL, MLB, NHL), as well Golf, MLS, and every soccer league in the world.

You can then select various parameters, Bookmaker Region and then specify the specific game type.

This add-on also allows you to set-up an automated data fetching interval, however the free licence only allows 500 monthly pings.

Fetching new data results in the following data table:

event_name	commence	status	bookmaker	last_update	point_1	point_2	odd_1	odd_2
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	DraftKings	2023-05-22T01:59:36.000Z	15.5	-15.5	-125	-105
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	PointsBet (US)	2023-05-22T01:58:57.000Z	13.5	-13.5	-105	-120
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	FanDuel	2023-05-22T01:59:36.000Z	12.5	-12.5	-113	-115
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	William Hill (US)	2023-05-22T01:59:36.000Z	14.5	-14.5	-110	-120
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	BetMGM	2023-05-22T01:58:59.000Z	13.5	-13.5	-110	-120
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	Circa Sports	2023-05-22T01:58:04.000Z	15	-15	-109	-109
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	Bovada	2023-05-22T01:59:47.000Z	13.5	-13.5	-105	-125
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	FOX Bet	2023-05-22T01:59:28.000Z	15.5	-15.5	-120	-120
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	Barstool Sportsbook	2023-05-22T01:59:44.000Z	14	-14	-112	-120
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	Unibet	2023-05-22T01:59:36.000Z	13.5	-13.5	-113	-118
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	BetRivers	2023-05-22T01:59:16.000Z	12	-12	-117	-117
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	SugarHouse	2023-05-22T01:59:20.000Z	12	-12	-117	-117
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	TwinSpires	2023-05-22T01:59:20.000Z	12	-12	-115	-115
Boston Celtics_Miami Heat	2023-05-22T00:30:00.000Z	Live	MyBookie.ag	2023-05-22T01:59:36.000Z	10.5	-10.5	-111	-125
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	FanDuel	2023-05-22T01:59:36.000Z	3.5	-3.5	-108	-112
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	BetMGM	2023-05-22T01:58:59.000Z	3.5	-3.5	-115	-105
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	DraftKings	2023-05-22T01:59:36.000Z	3	-3	-110	-110
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	PointsBet (US)	2023-05-22T01:58:57.000Z	3	-3	-110	-110
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	Circa Sports	2023-05-22T01:59:28.000Z	3	-3	-109	-109
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	William Hill (US)	2023-05-22T01:59:36.000Z	3	-3	-110	-110
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	LowVig.ag	2023-05-22T01:59:27.000Z	3	-3	-105	-105
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	BetOnline.ag	2023-05-22T01:59:38.000Z	3	-3	-111	-111
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	BetRivers	2023-05-22T01:59:16.000Z	3	-3	-109	-112
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	SugarHouse	2023-05-22T01:59:20.000Z	3	-3	-109	-112
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	TwinSpires	2023-05-22T01:59:20.000Z	3	-3	-109	-112
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	Barstool Sportsbook	2023-05-22T01:59:44.000Z	3	-3	-109	-112
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	Unibet	2023-05-22T01:59:36.000Z	3	-3	-109	-112
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	MyBookie.ag	2023-05-22T01:59:36.000Z	3	-3	-110	-110
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	FOX Bet	2023-05-22T01:59:28.000Z	3.5	-3.5	-110	-110
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	Bovada	2023-05-22T01:59:47.000Z	3	-3	-110	-110
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	SuperBook	2023-05-22T01:59:34.000Z	3	-3	-110	-110
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	WynnBET	2023-05-22T01:59:36.000Z	3	-3	-110	-110
Denver Nuggets_Los Angeles Lakers	2023-05-23T00:30:00.000Z	Pending	BetUS	2023-05-22T01:59:42.000Z	3	-3	-110	-110

It will bring in data for the upcoming 7 calendar days, including the odds from 15-20 of the world's leading book makers. A critical component for these odds is the status column, as that provides information related to the game itself. A 'Live' status means the game is currently in-progress, while a 'Pending' status means the game has not yet started.

An important aspect of our prototype was to ensure we archived all of the live, in-game odds as they were updated. We wrote a custom Visual Basic for Applications (VBA) script to automatically archive all of the odds after they had been retrieved.

Once we have input the information for all upcoming games, we need to set the Opening Lines. This was again completed in the custom MS Excel model.

Here is a screenshot of how we set the Opening Lines:

event_name	date	bookmaker	opening_line_1	opening_line_2	home_fav	home_dogs	road_fav	road_dog	classification	home	road
Miami Heat_New York Knicks|04/30/23	04/30/23	BetMGM	4.5	-4.5	Yes	No	No	Yes	home_fav_between_zero_and_five	home_fav	road_dog
Golden State Warriors_Sacramento Kings|04/30/23	04/30/23	BetMGM	1.5	-1.5	Yes	No	No	Yes	home_fav_between_zero_and_five	home_fav	road_dog
Boston Celtics_Philadelphia 76ers|05/01/23	05/01/23	BetMGM	-9.5	9.5	Yes	No	No	Yes	home_fav_between_five_and_ten	home_fav	road_dog
Denver Nuggets_Phoenix Suns|05/01/23	05/01/23	BetMGM	-4.5	4.5	Yes	No	No	Yes	home_fav_between_zero_and_five	home_fav	road_dog
Miami Heat_New York Knicks|05/02/23	05/02/23	BetMGM	6.5	-6.5	Yes	No	No	Yes	home_fav_between_five_and_ten	home_fav	road_dog
Golden State Warriors_Los Angeles Lakers|05/02/23	05/02/23	BetMGM	-4.5	4.5	Yes	No	No	Yes	home_fav_between_zero_and_five	home_fav	road_dog
Denver Nuggets_Los Angeles Lakers|05/16/23	05/16/23	BetMGM	-5.5	5.5	Yes	No	No	Yes	home_fav_between_five_and_ten	home_fav	road_dog
Boston Celtics_Miami Heat|05/17/23	05/17/23	BetMGM	-7.5	7.5	Yes	No	No	Yes	home_fav_between_five_and_ten	home_fav	road_dog
Boston Celtics_Miami Heat|05/21/23	05/21/23	BetMGM	-2.5	2.5	No	Yes	Yes	No	road_dog_zero_to_five	home_dogs	road_fav
Denver Nuggets_Los Angeles Lakers|05/22/23	05/22/23	BetMGM	3.5	-3.5	Yes	No	No	Yes	home_fav_between_zero_and_five	home_fav	road_dog

For our pilot POC, we only used the MGM sports book. The Live Odds API provides the opening lines for both teams. We use these lines to define the classification for every game prior to event start. The classifications we use are the same 9 evaluation scenarios that our Python model used (all_spread_bet, home_dogs, home_fav_between_zero_and_five, etc…).

Once the opening classifications for each game were set, we simply needed to wait for the start of each game. Using the MS Excel model, we would update the live, in-game odds at various periods. Then we cross-reference the changes in the odds across the Python predictions for the classified game type.

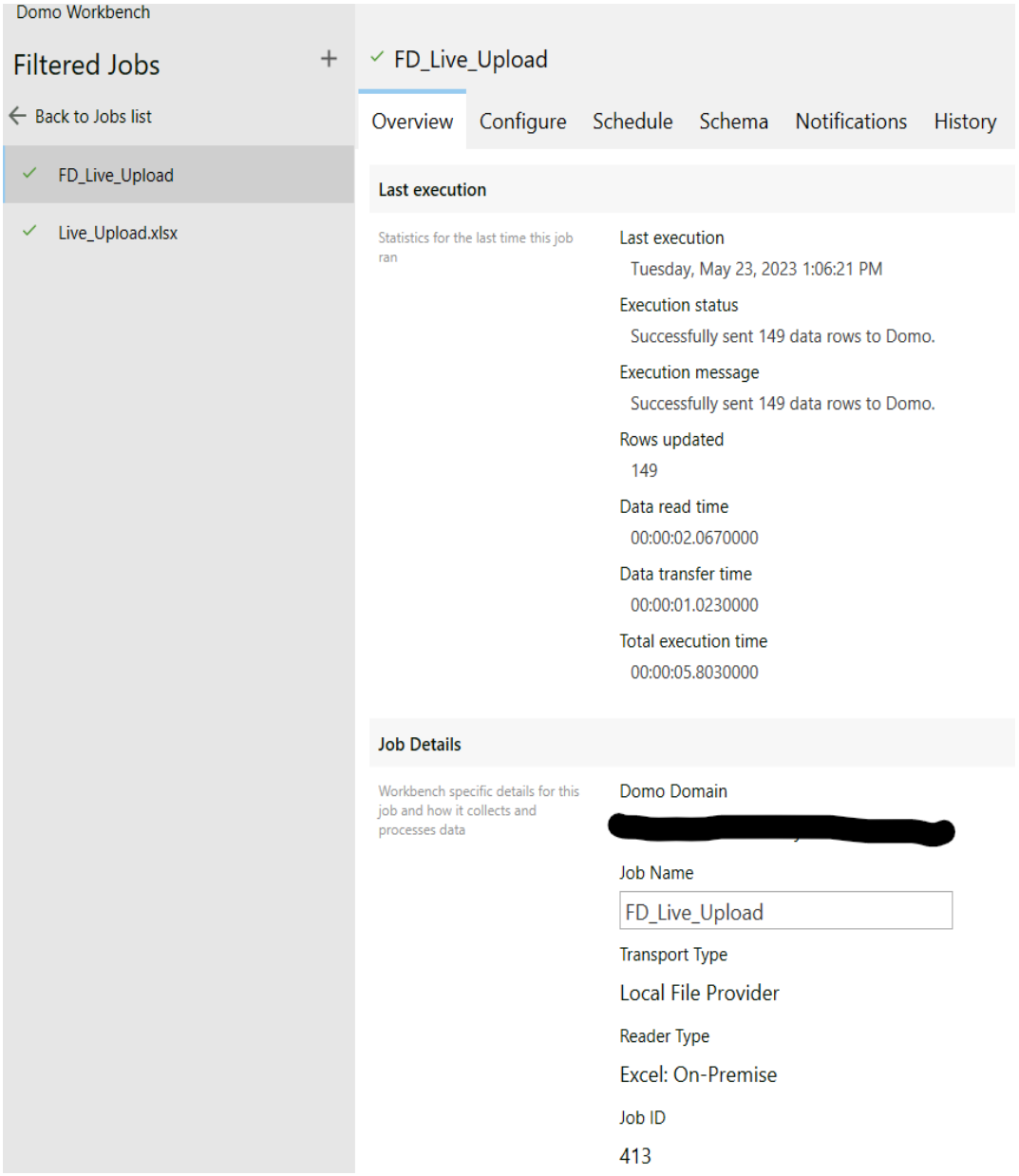
event_name	bookmaker	opening_line_1	opening_line_2	opening_classification	live_line_1	live_line_2	point_change	Win Probability
Miami Heat_New York Knicks|04/30/23	BetMGM	4.5	-4.5	home_fav_between_zero_and_five	-8.5	8.5	-13.0	74%
Golden State Warriors_Sacramento Kings|04/30/23	BetMGM	1.5	-1.5	home_fav_between_zero_and_five	0.0	0.0	-1.5	53%
Boston Celtics_Philadelphia 76ers|05/01/23	BetMGM	-9.5	9.5	home_fav_between_five_and_ten	-5.5	5.5	4.0	64%
Boston Celtics_Miami Heat|05/21/23	BetMGM	-2.5	2.5	road_dog_zero_to_five	13.5	-13.5	16.0	79%
Denver Nuggets_Los Angeles Lakers|05/22/23	BetMGM	3.5	-3.5	home_fav_between_zero_and_five	0.0	0.0	-3.5	58%

Step 3: Cloud-Based BI Platform

A critical component was connecting our MS Excel tool to our SaaS BI platform, Domo. We installed and configured an application called Domo Workbench to establish a live connection between the MS Excel file and the Domo cloud UI. A screenshot from the Workbench application can be found on the right.

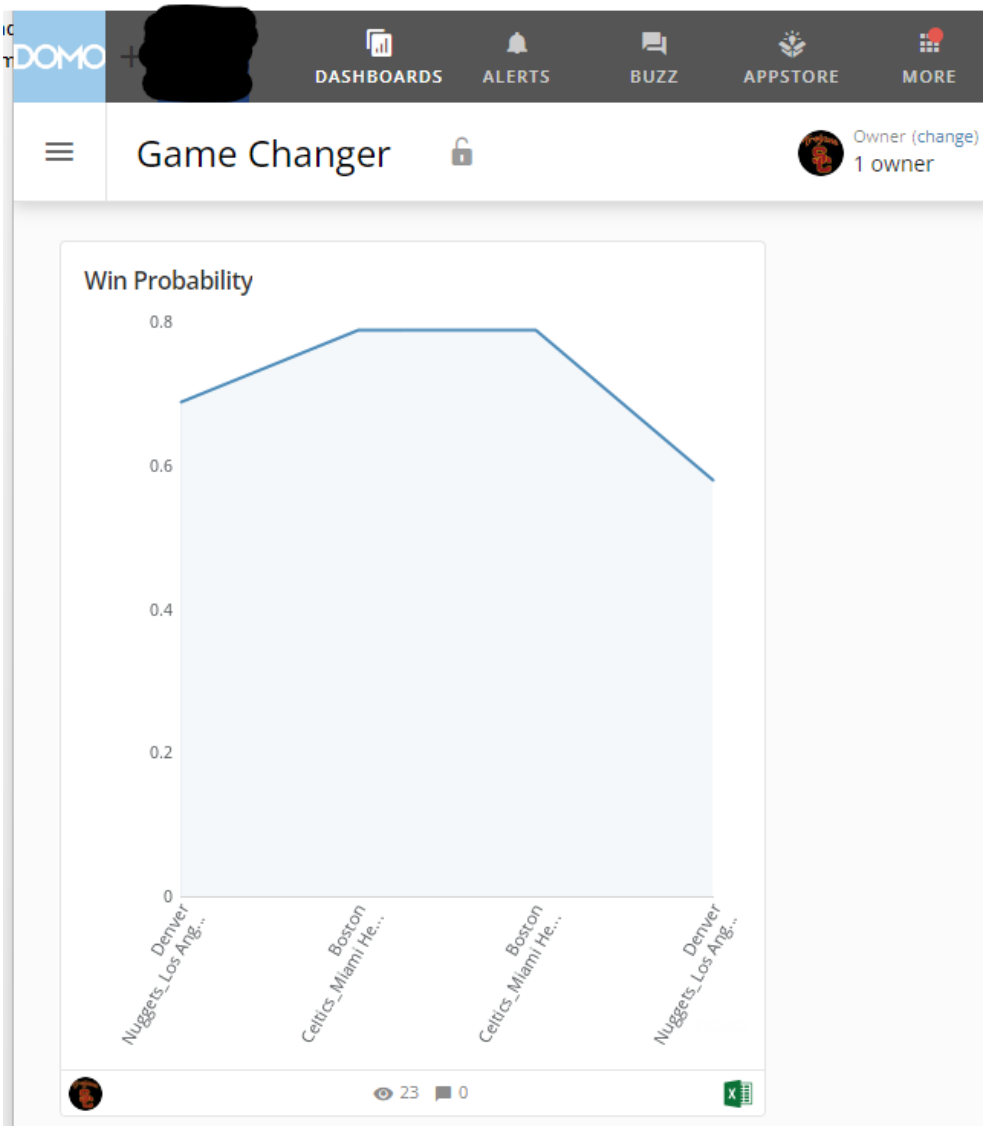
Using the Workbench's native Excel On-Premise data connector, we created a connection to the MS Excel model. There were dozens of configuration options and very precise conditions that needed to be met for the live upload process to work as expected.

This Domo Workbench job would effectively monitor the saved MS Excel model. As soon as its 'Last Modified' state was updated, Workbench would automatically start its upload job - normally within seconds.

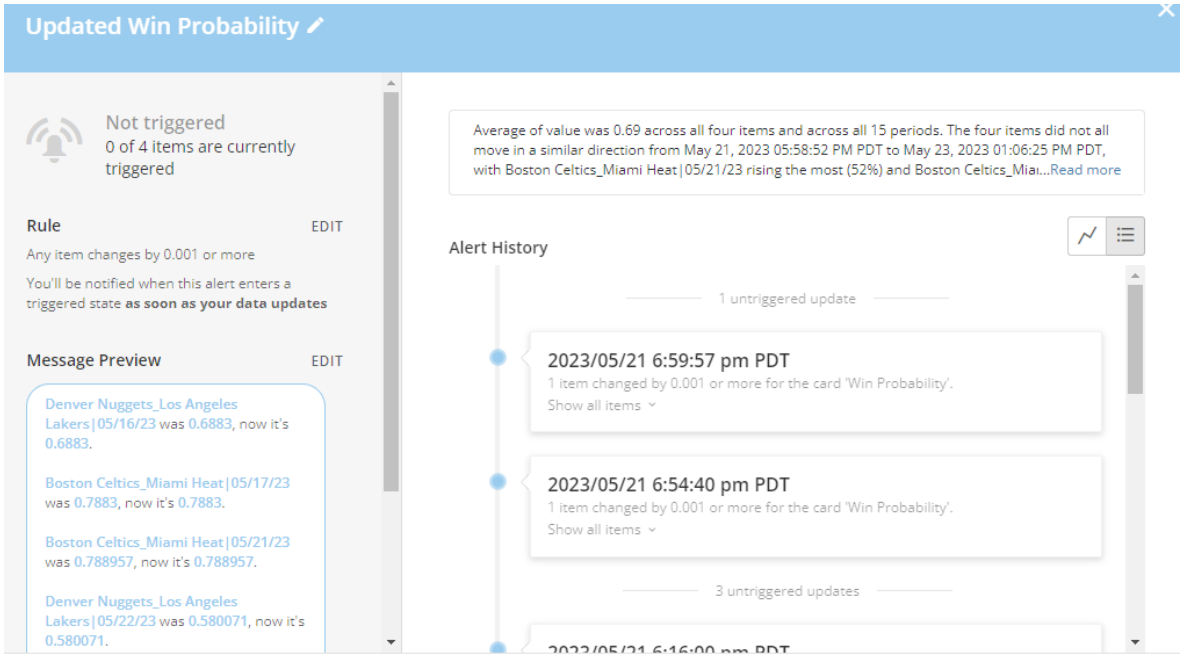


Step 4: Output Alerts

Once the source data was uploaded into the Domo UI, it would automatically update into a custom dashboard (see screenshot below).

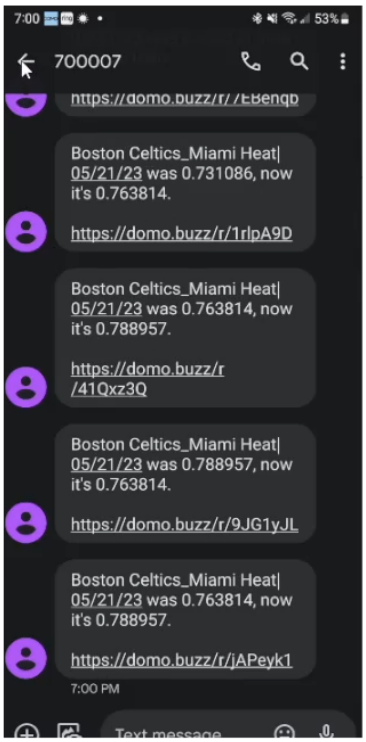


This dashboard was configured with the following set of alerts:

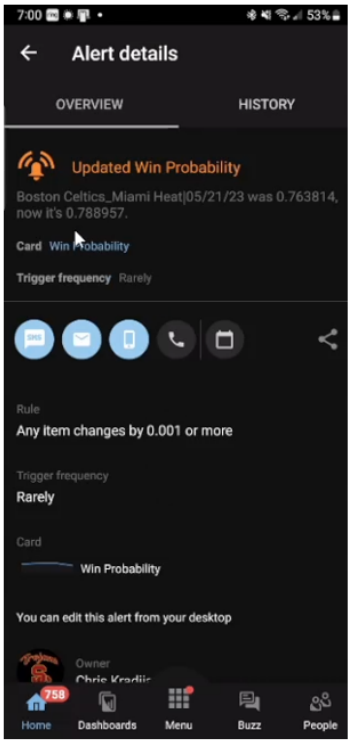


Effectively, any time the opening lines would change by 0.1%, it would trigger an alert. We configured Domo to send SMS texts, mobile app push notifications, and emails.

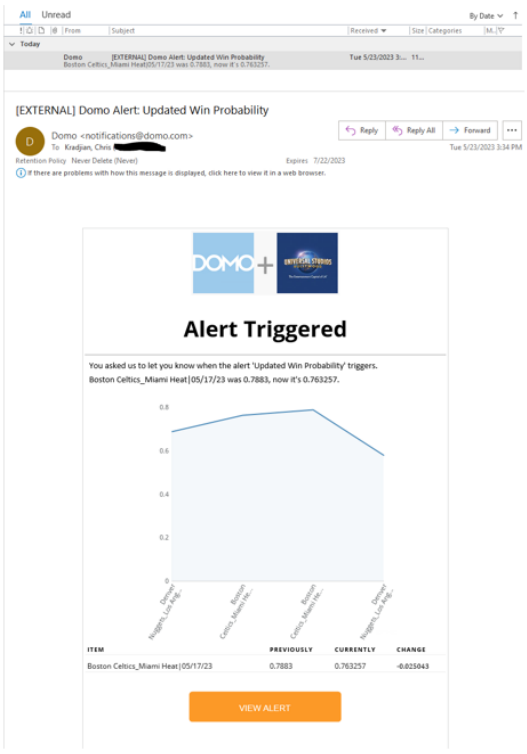
SMS Text



App Push Notification



Email



Future State

Our pilot demonstrated the feasibility of this project, and even came with a working prototype. Future iterations will need to be scalable and more automated. With live, in-game odds updates still a relatively new aspect of sports wagering, there aren't too many players in the space. The Live Odds API site has a very user-friendly MS Excel add-on, which we leveraged for our POC; however they also recently released a library for R, and offer some (albeit) limited documentation on configuring their APIs through Python. Programmatically extracting the live odds would vastly improve on this project's scalability. Domo also has native R connectors, which allows users to write data directly to Domo datasets. A programmatic connection to the odds APIs would effectively enable a nearly fully automated, end-to-end, customer facing product.

PROJECT TEAM

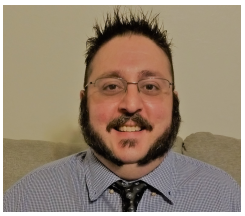
Timothy Steed

Tim S. is a researcher and aspiring data scientist at the Financial Industry Regulatory Authority (FINRA). He has over a decade of experience in derivative modelling and trading roles at various banks and hedge funds in his hometown of Chicago.

Timothy Chiu

Tim C. had prior experience in academic and industrial R&D as well as industrial manufacturing after obtaining a B.S. in Material Science and Engineering from U of I. Currently he is in a hybridized design development and project planning/management role within the operations organization of health diagnostics division in Abbott Laboratories' US headquarter. He is pursuing the MSDS program at Northwestern to pivot to an analytics management role and aspires to transform business operations through advanced analytics.

Christopher Kradjian



Chris has been working in analytics for over a decade now, after receiving his BS in Industrial and Systems Engineering from USC. He received his first Masters degree in Information Management Systems from Harvard Extension School and expects to obtain his second in Data Science from Northwestern in 2023. He currently leads the Revenue Management team at Universal Studios Hollywood.

Garrett Lynch

Garrett is a Cost Estimating and Pricing Analyst working in the defense/aerospace industry for 4 years. He is an aspiring data scientist, pivoting after he completed his MBA at UC Irvine and specialisation in digital transformation.. He has experience with risk analysis and project management of efforts in the space sector.

REFERENCES

Sports Betting Market Size & Share Analysis Report, 2030. (n.d.). https://www.grandviewresearch.com/industry-analysis/sports-betting-market-report

Jagger, Emma. “Best Sports APIs.” AbstractAPI.com. Retrieved April 15, 2023 from, https://www.abstractapi.com/guides/best-sports-apis#what-is-a-sports-api

Boeson, Ulrik. “Large Spread in Tax Treatment of Sports Betting Operators.” Tax Foundation. 9 Feb 2022. Retrieved April 15, 2023. https://taxfoundation.org/sports-betting-tax-treatment/

Buytendijk, F., Linden, A., & Laney, D. (2015, Sep 2). Big Data Strategy: Get Inspired, Get Going, Get Organized. Retrieved from Gartner: https://www.gartner.com/document/3123117

Eckerson, Wayne W. 2012. Secrets of Analytical Leaders. New Jersey: Technics Publications, LLC.

2021. PM Partners. June 23. Accessed April 16, 2023. https://www.pm-partners.com.au/the-agile-journey-a-scrum-overview/.

IBM. 2012. IBM SPSS Modeler CRISP-DM Guide.

McGrath, Tanner. 2022. Forbes.com. July 28. Accessed April 28, 2023. https://www.forbes.com/betting/sports-betting/how-sports-betting-odds-work.

Appelbaum, Josh. 2021. ActionNetwork.com. December 9. Accessed April 16, 2023. https://www.actionnetwork.com/education/how-do-betting-lines-work-vegas.

Metcalf, Matt. n.d. CircaResortCasino. Accessed April 16, 2023. https://www.circalasvegas.com/blog/sportsbook/from-the-experts/article/how-lines-are-set-an-oddsmakers-perspective/.

McShea, Chris, Dan Oakley, and Chris Mazzei. 2016. "The Reason So Many Analytics Efforts Fall Short." Harvard Business Review 5.

Venture Beat Staff. 2019. Venture Beat. July 19. Accessed May 7, 2023. https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-intoproduction/.