Aalto University
School of Business

PREDICTING THE RESULTS OF NFL GAMES USING MACHINE LEARNING

Master's Thesis
Sebastian Juuri
Aalto University School of Business
Information and Service Management
Fall 2023

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author Sebastian Juuri

Title of thesis PREDICTING THE RESULTS OF NFL GAMES USING MACHINE LEARNING

Degree Master of Science in Economics and Business Administration

Degree programme Information and Service Management

Thesis advisor(s) Pekka Malo

Year of approval 2023 **Number of pages** 49 **Language** English

Abstract

The NFL is the most popular sport in terms of viewership among the leagues in the United States, and is an entertainment business worth billions of dollars. Being able to predict the outcome of games is a popular topic among bettors, academics, but also among analysts who want to determine which game events most influence games in the long run. Additionally, predictions can serve as infotainment to provide context to spectators, and machine learning models can be utilized for other beneficial aspects such as injury prevention among athletes.

This thesis aims to take data from the most recent NFL seasons to find out if machine learning models can be built to accurately predict the outcome of games, and additionally see which statistics and in-game events weigh most heavily on the outcomes.

The literature review of the thesis covers previous research of machine learning models in American football, and also in other professional sports. The data is sourced from Pro Football Focus' website and includes the scores and team statistics from two recent NFL seasons. Three different machine learning models are applied on the dataset to compare their performances and results. All three models reached quite high accuracy scores, with the lowest model having 85% accuracy in predictions.

In conclusion, it seems possible to apply machine learning to predict the outcome of NFL games. There is room for further research by taking more seasons into account, or possibly comparing the results of the models on a dataset of college American football games for example.

Keywords NFL, American football, Machine Learning, Prediction Model

Aalto-yliopisto
Kauppakorkeakoulu

Aalto-yliopisto, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Maisterintutkinnon tutkielman tiivistelmä

Tekijä Sebastian Juuri		
Työn nimi NFL-PELIEN TULOSTEN ENNUSTAMINEN KONEOPPIMISTA KÄYTTÄMÄLLÄ		
Tutkinto Kauppatieteiden Maisteri		
Koulutusohjelma Tieto- ja palvelujohtaminen		
Työn ohjaaja(t) Pekka Malo		
Hyväksymisvuosi 2023	**Sivumäärä** 49	**Kieli** Englanti

Tiivistelmä

NFL on Yhdysvaltojen suosituin urheilulaji katsojamäärän perusteella, ja on vuosittain miljardien dollarien arvoinen urheiluliiga. Kyky ennustaa otteluiden tuloksia on suosittu aihe vedonlyöjien, akateemikkojen ja analyytikoiden keskuudessa, jotka haluavat selvittää, mitkä ottelutapahtumat vaikuttavat eniten pitkällä aikavälillä. Lisäksi ennusteet voivat toimia ns. infotainmentina, opastaen katsojia pelin tilanteesta. Koneoppimismalleja voidaan hyödyntää myös muihin hyödyttäviin käyttötarkoituksiin, muun muassa ennustamaan pelaajien loukkaantumisia.

Tämän väitöskirjan tavoitteena on ottaa dataa viimeisimmiltä NFL-kausilta selvittääkseen, voidaanko koneoppimismalleja rakentaa ennustamaan tarkasti otteluiden tulokset ja lisäksi selvittää, mitkä tilastot ja ottelutapahtumat vaikuttavat eniten lopputuloksiin. Kirjallisuuskatsaus keskittyy aiempaan tutkimukseen koneoppimismalleilla amerikkalaisessa jalkapallossa, sekä muissa urheilulajeissa. Data on peräisin Pro Football Focus -verkkosivustolta ja sisältää pisteet ja joukkueiden tilastot kahdelta viimeiseltä NFL-kaudelta. Kolme eri koneoppimismallia sovelletaan tietojoukkoon niiden suorituksen ja tulosten vertailua varten. Kaikki kolme mallia saavuttivat varsin korkean tarkkuuspistemäärän, ja heikoinkin malli pystyi ennustamaan 85 prosenttia tarkkuudella.

Johtopäätöksenä voidaan todeta, että koneoppimisen soveltaminen NFL-otteluiden ennustamiseen vaikuttaa mahdolliselta. Tutkimusaiheesta voi jatkaa eteenpäin ottamalla huomioon useampia kausia tai vertailemalla esimerkiksi korkeakouluamerikkalaisen jalkapallon tietojoukkoon perustuvien mallien tuloksia.

Avainsanat NFL, amerikkalainen jalkapallo, koneoppiminen, ennustava malli

Acknowledgements

I’d like to thank my friend Samuel for the inspiration on the thesis topic that allowed me to start writing the thesis.

Table of Contents

List of Tables

List of Figures

1 Introduction

The NFL is the most popular American sport in terms of viewership, and reported revenues of over 17 billion USD in 2021, being a massive entertainment business.

According to Nielsen, an American global measurement and data analytics company, In 2020, the NFL regular season had an average of 15.4 million viewers per game, and playoffs had an average of 28.6 million viewers per game. Furthermore, The Super Bowl is consistently the highest-rated television program in the United States each year. Super Bowl LIV (2020) had an estimated 100.7 million viewers, marking the tenth time in 11 years that the Super Bowl drew at least 100 million viewers. As for the Attendance figures for the NFL, In the 2019-2020 season, the NFL reported an average of 67,591 fans in attendance per game.

These numbers are considered to be among the highest in the world for any sports league. The league consists of 32 teams, playing a combined 272 regular season games and 13 playoff games. Like other sports, the NFL is a popular betting sport, being a huge business. Due to the nature of the game, there are many different targets to bet on than just the outcomes of games: bets can be placed on the amount of touchdowns scored, how many yards a certain player has carried the ball, and many others. Betting is done on even more obscure facts of the game, such as the color of Gatorade traditionally poured on the head coach of the winning team (Forbes, 2022).

The NFL itself is utilizing Machine Learning for purposes such as injury prevention and other safety improvements, and has partnered with Amazon Web Services towards this goal (NFL, 2019).

1.1 Research problem and objectives, structure

The aims of this thesis are to use multiple machine learning models suitable for a binary classification problem (win or no win in this case) to learn influential variables for the success of teams in NFL games, especially now with the regular season extended to have 17 games for each team since 2021. Most recent NFL seasons are also most suitable as rule changes and the evolution of playbook schemes has affected the dynamics of the sports over the years, and the NFL has become a more pass-heavy sport over the years. As a result, the models should provide accurate results to be considered dependable and to indicate statistics that truly matter for the outcome of games.

The question the thesis is answering is do the prediction and predictive power of Machine Learning models differ for the sport of American Football compared to other sports and if so, which models may be most suitable for this particular purpose. Also of interest is to see which available data points are most influential: what events affect the outcome the most? The research questions are as follows:

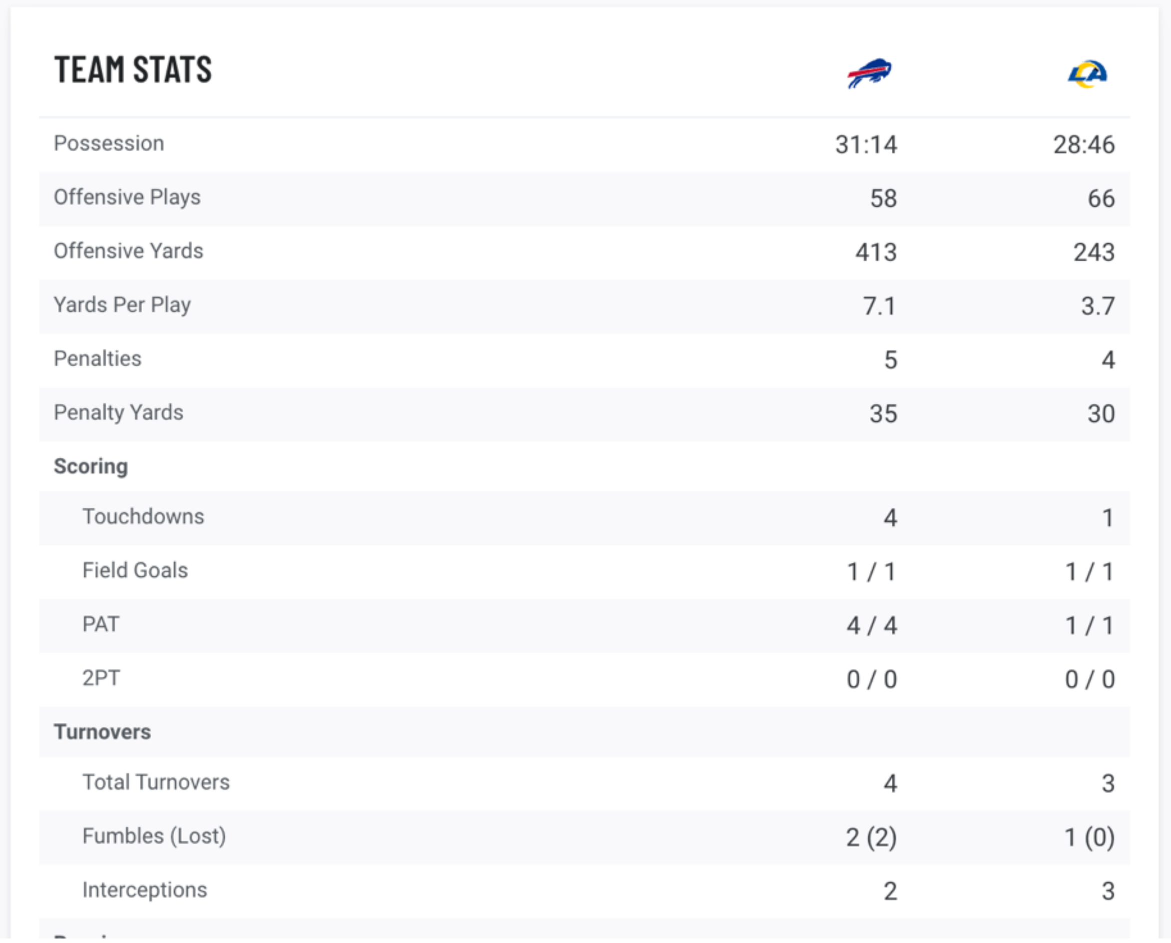
- Can the outcome of NFL games be predicted using Machine Learning models?
- Which factors are determined to be the most important for the outcome of games according to different models?

Following the introduction, Section 2 will dive into previous literature of predictive analytics and machine learning in the world of sports from many different angles. Following that, the third section will describe the data source, and the necessary engineering completed to create a dataframe that the machine learning models can interpret successfully. After that, the results will be analyzed, the models are compared for their differences in results and also performance. Finally, there will be a conclusions sections along with limitations and future research opportunities described in brief.

2 Literature Review

2.1 General information about American Football

American Football is a team sport with 22 players simultaneously on the field, played on a rectangular field. The teams consist of an offense and defense, with roles swapped based on which team is in possession of the oval-shaped ball according to rules. The task of the offense is to move the ball towards the other team’s end zone, and has 4 opportunities (called downs in American football) to advance 10 yards, and they gain another 4 tries if they do reach the 10-yard distance. The primary way to score points is to advance the ball into the end zone at the end of the field (a touchdown, worth 6 points), and also by kicking through the uprights at the end of the field (a field goal, worth 3 points). After a successful touchdown, the team also gains the opportunity to attempt an extra point: by kicking the ball, they can gain one point, and by attempting a attempting a single play to reach the end zone from a 2-yard distance, that awards them two points.



TEAM STATS		LA
Possession	31:14	28:46
Offensive Plays	58	66
Offensive Yards	413	243
Yards Per Play	7.1	3.7
Penalties	5	4
Penalty Yards	35	30
Scoring		
Touchdowns	4	1
Field Goals	1 / 1	1 / 1
PAT	4 / 4	1 / 1
2PT	0 / 0	0 / 0
Turnovers		
Total Turnovers	4	3
Fumbles (Lost)	2 (2)	1 (0)
Interceptions	2	3

Figure 1: Example from Pro Football Focus on statistics that are tracked for American Football matches

The team with more points at the end of the match is the winner. There are also what are called special team plays, which consist of kickoff, field goal, and punting which occur under specific circumstances throughout a game of American football.

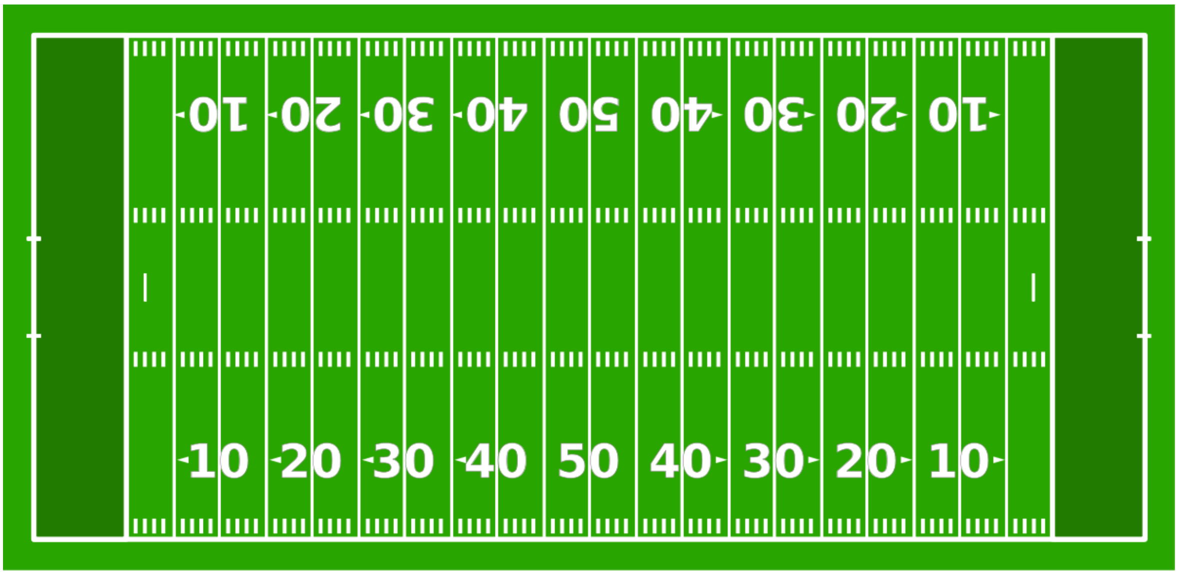


Figure 2: The playing field in American football. The dark green areas on the ends indicate the endzone, where a touchdown is scored

Advancing the ball, and also defensive efforts to stop the advancement make for plenty of opportunities to numerically measure the performance of the teams and individuals playing, such as yards gained or pass completion percentage. Thus, American Football games offer plenty of numerical statistics from games that can be utilized in predicting the outcomes of games.

For the purposes of explaining key terms that are important for interpreting the variables that are used in the data of this thesis, and to clarify dynamics of a sport that is not as widely spectated in Europe, this part of the thesis includes a table with explanations for different terms.

Table 1: Glossary of American football terms

Terminology	Explanation
Rushing	The ball is advanced by a player without throwing the ball
Passing	The ball is thrown to another player on the offense as an attempt to advance
Penalty	A player on the field did something against the rules. Different penalties can involve moving the ball a number of yards, costing the team in field position

Sack	The quarterback is tackled while attempting a passing play.
Turnover	The offense loses the ball to the defense during a live play, considered to be a very negative result for a play.
3rd down	The second last attempt an offense has to attempt to reach the 10 yards to gain another set of downs, attempts in other words. Converting on the 3rd down is considered an important metric of offensive performance.
4th down	The last attempt. Often if the field position is not favorable and reaching the required yards is not likely, teams will often opt to kick the ball to the opposing team instead of playing an offensive play.

2.2 Statistics used to measure performance in American Football

The field in American Football is made up of 100 yards of playing field, with the end zone on each end of the field. The yardage is marked throughout the field, and this allows a lot of statistics to be gathered from each play, for example the amount of yards gained or lost, who tackled the ball carrier, was the ball thrown or handed off to a player, among other statistics. A lot of these statistics can be turned to an average or a total for the game that can indicate the overall performance of a team in a certain game. Some of the most common measures of performance for an offense is average yards per attempt, time of possession, and also total yards gained on offence. On defense, important statistics are points allowed per game, and yards allowed per game.

Pro Football Focus (PFF) is one of the most trusted sources of grading for a players performance on a graded scale based on a formula. PFF takes into account events that can be considered a successful effort by a player, such as a good throw by a quarterback that ended up being dropped by the intended receiver, where one player executed perfectly, but it does not end up converting into a positive statistic in the game stats. Players are graded on each play according to an expected "normal" performance on the play for their position, with negative or positive impact on the grade based on their contribution on the specific play. This style of measuring also isolates the performance of a single player: a quarterback can make a throw to a receiver, and the receiver may either run it for a touchdown or get tackled after two yards. In raw statistics, this is counted towards a quarterback's passing yards and touchdowns, even though the quarterback performed to the same exact level. This is argued by PFF as an advantage of their measuring system.

Football Outsiders (FO) is another organization providing analytics and performance ratings based on NFL games. Football Outsiders was founded in 2003, and they provide analysis and commentary on the NFL and college football through their website and various publications. They also produce annual publications, such as the Football Outsiders Almanac, which includes detailed statistical analysis and projections for the upcoming season. They are known for their advanced metrics, such as DVOA (Defense-adjusted Value Over Average), which measures a team's performance relative to the league average, adjusted for the strength of their opponents (footballoutsiders.com, 2023).

DVOA stands for Defense-adjusted Value Over Average, and it is a metric used to measure a team's performance on a per-play basis, relative to the league average, while adjusting for the strength of their opponents. It is calculated by comparing the success of each play to a baseline of expected success, based on the down, distance, and field position of the play (Football Outsiders, 2023).

To calculate a team's overall DVOA, the individual DVOA values for each play are aggregated and adjusted for the strength of the team's opponents. This adjustment is based on the average DVOA of the teams a team has played against, with greater weight given to more recent games. The result is a percentage that indicates how much better or worse a team is than the league average. A positive DVOA indicates that a team is better than average, while a negative DVOA indicates that a team is worse than average.

DVOA is considered a more accurate measure of a team's performance than traditional box score statistics, as it takes into account factors such as strength of schedule and situational context. It is used by Football Outsiders and other NFL analysts and fans as a way to evaluate team performance and predict future outcomes.

Limitations of DVOA include that the weights assigned to different performances still involve some subjectivity from the part of the analysts, and because DVOA is calculated on a per-play basis, it can be subject to a small sample size. In some cases, a single play or a small number of plays can have a significant impact on a team's DVOA, which may not be reflective of their overall performance.

2.3 Previous research

Statistics have been used for predicting the results of upcoming games for a long time in professional sports, for various purposes from a reporting perspective to creating odds for sports betting. Predictions have ranged from a combination of expert opinions to statistics to include mathematical models during the last decades. With the development of technology and methods to track in-game events, the amount of data available has increased. At the same time, the development of machine learning and computing power has allowed for more possibilities to apply algorithms and models to predictions. The more data is available, the more data points can be considered to have an effect, and also enables to find more factors that impact game results.

2.3.1 Literature on machine learning and predictive models in various sports

Using statistics and machine learning models to predict outcomes and events in sports games has been a popular topic in other sports than American football as well. The historical performance from previous seasons has been used to predict Top 10 finishes in the PGA tour using logistic regression and decision tree models (Korpimies, 2020).

Bai et al. (2022) investigated the impact of game and team statistics on the outcome of soccer matches in three major soccer leagues: English Premier League (EPL), Spain's La Liga, and Major League Soccer (MLS) in the USA, using game data and different machine learning models on historical data to predict the outcome of a win, loss or tie. The difference between the number of shots on target, and whether the team were the home/away team were found to be the most significant variables affecting the outcome of the game.

Tabtah et al. (2019) applied machine learning models such as Logistic Regression and Support Vector Machines to predict the most significant factors of success in the NBA. The authors used data of NBA final games between 1980 and 2017 for this study, and found the most significant factor to be defensive rebounds, followed by three-point percentage, and number of free throws. One of the limitations of this study to consider are that the NBA finals represent a playoff scenario, which can not only influence the strategy of teams and psychologically be a different type of game, but is also played after a full regular season. Fatigue, lingering injuries and other factors could have an effect on which variables become most significant in this scenario compared to the NBA regular season games.

2.3.2 Applications of machine learning in sports in other contexts than statistic performance

The use of machine learning for sports is an intriguing field from many perspectives, and not just for purely predicting the results of games. Academics, sports organizations and leagues are interested in finding ways to for example prevent injuries, understand the impact of different conditions to games, and other areas that have become possible to measure thanks to the advancement of tracking and data available from sports games and training.

Majumdar et al. (2022) conducted a study on the use of machine learning for injury prediction in football. The authors noted that attempts to better understand the relationship between training and competition load and injury in football are essential for helping to understand adaptation to training programmes, assessing fatigue and recovery, and minimizing the risk of injury and illness. To this end, technological advancements have enabled the collection of multiple points of data for use in analysis and injury prediction.

The study found that while machine learning has the potential to bring new insight to our understanding of injury prediction in football, there are several limitations and challenges associated with its use. One major concern is that the studies examined in the article are based on data collected across a single season, which may limit their generalizability to other seasons or contexts. Additionally, there is considerable variability in study design and analysis, which can affect the accuracy of injury prediction models. Finally, machine learning algorithms require large amounts of high-quality data to be effective, which can be challenging to obtain in practice.

2.3.3 Status of the betting market for the NFL and possible inefficiencies that could be supplemented with Machine Learning

Plenty of research has been done on the betting markets of significant sports leagues that attract a lot of bettors and present a significant business. Betting markets make an interesting topic of research since the motivation of financial gain is likely to incentivize the creation of as accurate as possible predictions and expertise, and one could argue that this would drive the betting market to be efficient. Several researchers have looked at the NFL from different points of view to see if this is the case.

For example, Borghesi (2007) looked into the effect of weather conditions on games. Weather and temperature has a significant impact in American football especially on how easy it is to throw and catch the ball. In this article, the author looked into if any team can be considered to gain an advantage from weather. The study found that there is a mispriced acclimatization advantage associated with games played in extreme temperature conditions, and that bets on cold-acclimatized home teams playing in their element are significantly underpriced. These findings suggest that the NFL betting market is not fully efficient, and that there may be opportunities for profitable betting strategies based on weather-related biases. However, this might have adjusted from the time of the study in the betting market.

The same author had taken a different perspective in another study. Borghesi (2007) examined the home-underdog effect in the NFL point-spread betting market and identifies a persistent increase in bias magnitude during the final few weeks of each season as its primary cause. The authors use several statistical tests and regression analysis to confirm the existence of this phenomenon and analyze its causes. They demonstrate that NFL bet prices persistently deviate from their underlying true values at regularly occurring intervals over a 20-year period studied. The authors propose several regression models that can be used to identify inefficiencies and predict mispricing before they occur. The model differs from those in prior literature by taking into account both short-term and aggregate biases. Overall, this article offers a new approach to identifying mispricing in the NFL point-spread betting market and implementing profitable betting strategies.

A more recent study into the implications of the atmosphere during games was done by Paul (2017). The author found that atmospheric conditions such as temperature, humidity, precipitation, barometric pressure, wind speed, and altitude all have significant effects on scoring in NFL games. Number of turnovers and offensive statistics were heavily impacted for example by precipitation, and it seemed that the betting market did not fully adjust for this in the study. The author states that there are multiple reasons for this and perhaps the betting market is not liquid enough for accounting such nuanced factors. The accurate prediction of exact conditions and for example how long rain lasts during a game can also be a factor that it is too unpredictable to factor in.

Mohsin & Gebhardt (2022) conducted a simulation study to assess the stability of the model parameters, which showed that the estimated values of the model parameters become close to true values as sample sizes increase. They also applied their proposed model to a real dataset of NFL games and compared it with other extant distributions.

The stochastic model Mohsin and Gebhardt (2022) developed was a Bivariate Affine-Linear Exponential (BALE) distribution. This distribution is used to model the margin of victory in NFL games, which is the difference between the scores of the favorite and underdog teams. The BALE distribution is a subtle distribution that is suitable for modeling low and moderate negative dependence between scores by the favorite and underdog teams. It has four parameters: two location parameters, a scale parameter, and a correlation parameter. The authors use maximum likelihood estimation to estimate these parameters from observed data.

The results from Mohsin and Gebhardt (2022) showed that their proposed model provides a better fit than other distributions, such as normal, t, skew-normal, skew-t, and generalized hyperbolic distributions. The authors also computed quantiles and assessed point spread using their proposed model.

Overall, the authors conclude that their proposed stochastic model based on BALE distribution provides a good fit for modeling margin of victory in NFL games and can be used for point spread assessment in sports betting markets.

Shank (2019) also examined the biases that exist in NFL betting markets and their impact on profitability. Shank found that bettors have a preference for betting against line movement and a preference for betting on favorites, which can create opportunities for profitable betting strategies using a simple regression model. However, these biases can also be exploited by bookmakers who set point spreads and over/under lines to take advantage of these biases. The article also presents results for betting behavior in the over/under market, showing that bettors prefer to bet on the over and are more likely to do so if the home team has covered the over in more recent games. Overall, Shank's review highlights the importance of understanding biased decisions in sports betting markets to make more rational decisions and maximize profits. In another study by Shank (2018), he found that there are statistical inefficiencies in both points spread and total markets, with home teams that are substantial underdogs more likely to cover the spread and games with low or high posted totals more likely to cover the over. Additionally, games are more likely to cover the over when the home team has a "hot hand."

2.3.4 Variables and statistical predictions considered in previous literature for American Football

Past research has focused on prediction of NFL games and sports in general. One example is the use of in-game situation to predict whether the next play results in a turnover (the game ball is forcibly turned over to the opposing team during play) which is a significant occurrence in terms of an outcome for a game (Bock, 2016). The author trained a Gradient Boosting Method model (GBM) on NFL play-by-play data for 32 teams spanning seven full seasons to predict the likelihood of observing a turnover on the next play from scrimmage during a football game. Factors of the game situation such as time remaining and the score differential were found to be the most significant factors, which make intuitive sense as they involve a situation where a team is likely to take more risk with offensive play calling.

An overtime occurs in a game of American football when the teams are evenly scored by the end of the four quarters. Because of the nature of the game, one team will start on defense and one on offense, and a point of discussion has been whether one or the other is advantageous to start at. This of course depends on the rules of the overtime, but also the historical performance of the team running their offense near the goal zone is a significant variable (Wilson, 2020).

Logistic regression models have also been applied on late-game field goal kicking situations, another event that occurs often near the end of the game when the score is tied or the score is very even where 3 points awarded for a successful field goal can change the lead. Interestingly, a common practice of the opposing team calling a timeout right when the kicker would be ready to mess with their routine was found to actually help kickers according to a logistic regression model that used play-by-play data from 2000 to 2017 from the NFL (Hsu et al., 2019).

Gifford and Bayrak (2020) developed a decision tree model to determine which team statistics had the most significant influence on matches. The decision tree is used to determine which attributes a team should possess in order to have the best chance at winning a game. By understanding the most significant influences for winning, they can predict the outcome of matches. They found that turnovers were the most significant factor in winning games. According to their decision tree model, turning the ball over at least twice on offence had a heavy impact, with only 32% of such cases still resulting in a win.

Lock's research (2022) that focused also on the NHL, NBA, but also on NFL found as well a way to predict the accuracy of NFL field goal kickers using a Bayesian approach.

Field goals can often decide games in the final minutes in the NFL, so the model can have significance on the prediction of game outcomes.

Hill (2022) applied predictive models to predict the result of Canadian Football games, which is close to American football with a few rule changes, such as both teams having 12 players on the field. Hill's dataset consisted of play-by-play and wagering data for games from the CFL, with statistics of matches between the 2015-2019 seasons to develop win probability models for the sport. Two types of machine learning models were developed: a gradient boosting model, and a logistic regression one.

Logistic regression is a statistical method that estimates the probability of an event occurring based on one or more predictor variables. In this study, logistic regression was used to estimate the probability of a team winning a game based on various factors such as time score differential, field position, down and distance, remaining, and other situational variables.

Gradient boosting is a machine learning technique that builds an ensemble of decision trees to make predictions. In this study, gradient boosting was used to create a more accurate model by combining multiple decision trees. The model was trained on historical data and then tested on new data to evaluate its accuracy.

Hill (2022) also incorporated pregame spread and total (over/under) data into these models. Pregame spread refers to the predicted point differential between two teams before a game starts. Total refers to the predicted total number of points scored by both teams in a game. These variables were included in the models as predictors to improve their accuracy.

The accuracy of these models was evaluated using classification accuracy, which measures how well the model correctly predicts whether a team will win or lose based on its estimated win probability. The classification accuracy of these models was found to be 0.737 (73.7%), assuming a probability threshold of 0.5.

2.4 Conceptual Framework

The previous research on predicting the outcome of NFL games and of sports matches in general has focused on various different variables having an effect on the outcome, and has included many angles from reviewing the efficiency of betting markets to also focusing on individual events in games such as the predictability of a successful field goal. The following framework lays out how previous literature is utilized and taken into account in this thesis.

One of the significant differences compared to some of the previous literature is the use of play-by-play data, not just pure statistics from games. The other difference to previous literature is the incorporation of data that is included from seasons when the National Football League installed a 17-game regular season, extending each team's schedule by one game each season. In a league where every game has significant impact for playoff and draft outcomes compared to say, NBA's 82-game regular season, the addition of one game is likely to have impact to outcome of games.

NFL play-by-play data will be used primarily for this research, however some of the variables will likely need to be dropped or combined to tackle possible issues in the model. Some variables in the data do not directly relate to performance of teams, but may factor into specific teams winning games, such as whether the game is played on Sunday, Monday or Thursday, among other variables. The measure of success in this thesis is easy to choose as we will be looking at the outcome of games. Games end either in a win, loss or tie. Ties are rare occasions in American Football, but still occur from time to time. For the purposes of having a binary variable, ties can be counted as "non-wins" along with losses.

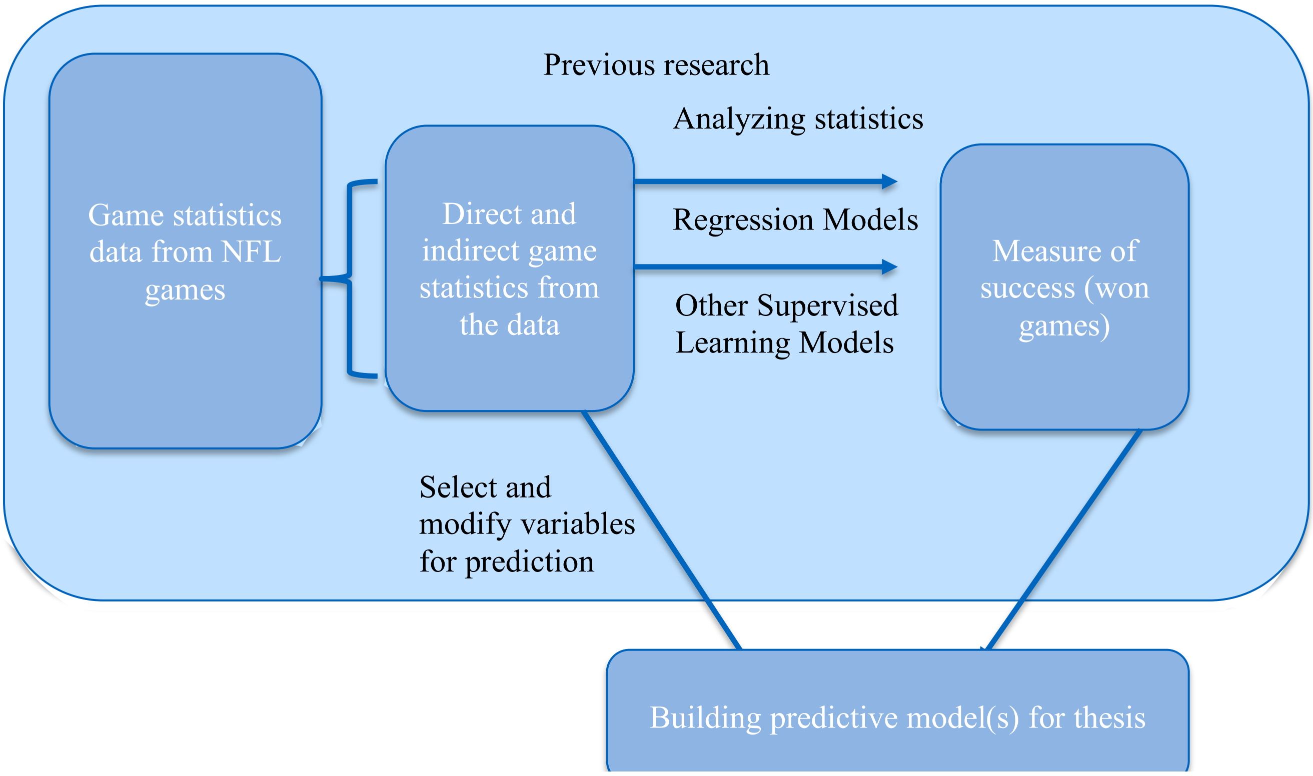

Previous research
Game statistics data from NFL games
Direct and indirect game statistics from the data
Analyzing statistics
Regression Models
Other Supervised Learning Models
Measure of success (won games)
Select and modify variables for prediction
Building predictive model(s) for thesis

Figure 3: Conceptual Framework

3 Data

Before the appropriate predictive models can be applied to the NFL game data, the data must first be selected, formatted, cleaned up and select the correct variables in a process often referred to as data engineering. This chapter aims to explain the process from the data sources selected to this particular research, a description of how the data was formatted, combined and engineered to be ready for running machine learning models, and why particular variables are selected or not selected in the context of American football.

3.1 Data Selection

The data selected for this thesis research originates from Pro Football Reference's website (https://www.pro-football-reference.com/https://www.pro-football-reference.com/) that is accessed using the nflscraPy package. The Pro Football Reference website is upheld by the company Sports Reference, which has been keeping records of current and past NFL seasons since 2004 (sports-reference.com). The website lists data for individual player statistics, seasons, and box scores of games. For this research, we are using the scoring statistics together with box score data of games from NFL seasons. For the purposes of this research, we are particularly interested in the latest NFL seasons which implemented an additional regular season game for each team, bringing the length of the regular season to 17 games. Thus, the data for seasons 2021 and 2022 are selected for the model.

3.2 Modeling

Predicting the outcome of sports games and other phenomena around sports in general has changed significantly with the evolution of machine learning and availability of analysis tools. The capability to track and record more events during games has also contributed to the rise of using machine learning models outside of purely predicting the performance of players and teams.

Table 2: Variables in analysis. Each variable is included with prefix team1 and team2 to represent statistics of both teams in the game

Independent variable	**Explanation**
team_rush_att	Rush attempts (running without throwing the ball)
team_rush_yds	Yards gained by rushing
team_rush_tds	Touchdowns scored by rushing
team_pass_cmp	Completed passes (successful throws)
team_pass_att	Attempted passes
team_pass_cmp_pct	Completion as a percentage of attempts
team_pass_yds	Yards gained by passing plays
team_pass_tds	Touchdowns scored by passing
team_pass_int	Interceptions thrown (defending team catches the ball)
team_passer_rating	
team_net_pass_yds	
team_total_yds	Total yards team gained on offense
team_times_sacked	Times Quarterback was tackled while attempting a pass
team_yds_sacked_for	Negative yards suffered because of a sack
team_fumbles	Number of times offense fumbled the ball
team_fumbles_lost	Number of times the fumble resulted in turning the ball over
team_turnovers	Total number of turnovers
team_penalties	Number of penalties suffered
team_penalty_yds	Yards lost because of penalties
team_first_downs	Number of first downs gained during the game
team_third_down_conv	Number of times the team achieved a first down from a play on the third down
team_third_down_att	Number of attempts on a third down
team_third_down_conv_pct	Percentage of third downs converted to first downs
team_fourth_down_conv	Number of times the team achieved a first down from a play on the fourth down
team_fourth_down_att	Number of attempts on a fourth down
team_fourth_down_conv_pct	Percentage of fourth downs converted to first downs
team_time_of_possession	Total time of posession the offense had with the ball

The variables represent many indicators that reflect the performance of an American football team. It is of note that the statistics are from the viewpoint of an offense: converting downs, passing and rushing yards are all offensive statistics. Offensive statistics are easier to quantify than defensive statistics, and offensive plays also are responsible for most of the points scored within games, so it makes practical sense to use offensive statistics, but also the convenience of the numbers being easily available plays a factor. Nevertheless, defensive performance is often reflected in the fact that when a particular team's defense is playing well, then the opponent offensive statistics are impacted: they are likely to have less yards gained, and more turnovers for example.

Table 3: Summary statistics representing 'Team 1'

	mean	std	min	25 %	50 %	75 %	max
team1_rush_att	26.595	7.459	8.000	21.000	26.000	32.000	48.000
team1_rush_yds	117.109	50.522	1.000	78.000	109.500	148.000	320.000
team1_rush_tds	0.889	0.897	0.000	0.000	1.000	1.000	4.000
team1_pass_cmp	21.935	6.231	5.000	18.000	22.000	26.000	43.000
team1_pass_att	34.120	8.334	11.000	28.000	34.000	39.000	68.000
team1_pass_yds	239.599	72.845	43.000	189.750	235.000	287.000	446.000
team1_pass_tds	1.474	1.164	0.000	1.000	1.000	2.000	6.000
team1_pass_int	0.836	0.922	0.000	0.000	1.000	1.000	4.000
team1_passer_rating	89.166	25.025	2.778	73.195	89.028	106.250	157.583
team1_net_pass_yds	222.801	75.218	1.000	171.000	217.000	272.000	445.000
team1_total_yds	339.907	82.872	47.000	284.000	334.000	397.000	576.000
team1_times_sacked	2.509	1.762	0.000	1.000	2.000	3.000	9.000
team1_yds_sacked_for	16.798	13.122	0.000	7.000	15.000	24.000	68.000
team1_fumbles	1.199	1.126	0.000	0.000	1.000	2.000	6.000
team1_fumbles_lost	0.474	0.669	0.000	0.000	0.000	1.000	4.000
team1_turnovers	1.310	1.130	0.000	0.000	1.000	2.000	5.000
team1_penalties	5.702	2.576	1.000	4.000	5.000	7.000	14.000
team1_penalty_yds	47.694	24.856	4.000	30.000	45.000	63.000	166.000
team1_first_downs	19.891	4.979	6.000	17.000	20.000	23.000	36.000
team1_third_down_conv	5.086	2.234	0.000	4.000	5.000	6.000	14.000
team1_third_down_att	12.871	2.552	6.000	11.000	13.000	15.000	21.000
team1_third_down_conv_pct	0.389	0.144	0.000	0.300	0.385	0.500	0.900
team1_fourth_down_conv	0.780	0.944	0.000	0.000	1.000	1.000	6.000
team1_fourth_down_att	1.493	1.399	0.000	0.000	1.000	2.000	7.000
team1_fourth_down_conv_pct	0.373	0.415	0.000	0.000	0.250	0.750	1.000
team1_time_of_possession	1814.433	274.152	1105.000	1616.000	1814.000	2012.500	2764.000

Table 4: Summary statistics for variables representing opposing team

	mean	std	min	25 %	50 %	75 %	max
team2_rush_att	27.114	8.073	6.000	21.000	26.000	33.000	50.000
team2_rush_yds	118.437	53.963	3.000	77.000	107.000	153.000	363.000
team2_rush_tds	0.935	0.994	0.000	0.000	1.000	1.000	5.000
team2_pass_cmp	21.889	6.259	2.000	18.000	22.000	26.000	41.000
team2_pass_att	33.835	8.884	3.000	28.000	34.000	39.250	66.000
team2_pass_yds	240.539	75.774	19.000	190.000	235.500	290.000	525.000
team2_pass_tds	1.481	1.127	0.000	1.000	1.000	2.000	6.000
team2_pass_int	0.746	0.920	0.000	0.000	0.000	1.000	4.000
team2_passer_rating	91.578	25.590	5.303	74.138	91.453	107.633	156.090
team2_net_pass_yds	225.570	77.341	-6.000	174.000	223.000	275.250	498.000
team2_total_yds	344.007	80.111	53.000	297.750	342.000	396.000	575.000
team2_times_sacked	2.190	1.588	0.000	1.000	2.000	3.000	9.000
team2_yds_sacked_for	14.968	12.015	0.000	6.000	13.000	22.000	82.000
team2_fumbles	1.206	1.101	0.000	0.000	1.000	2.000	5.000
team2_fumbles_lost	0.546	0.709	0.000	0.000	0.000	1.000	4.000
team2_turnovers	1.292	1.195	0.000	0.000	1.000	2.000	5.000
team2_penalties	5.715	2.482	0.000	4.000	6.000	7.000	14.000
team2_penalty_yds	48.366	25.057	0.000	30.000	45.000	62.000	161.000
team2_first_downs	20.164	4.860	4.000	17.000	20.000	23.000	33.000
team2_third_down_conv	5.211	2.198	0.000	4.000	5.000	7.000	12.000
team2_third_down_att	12.827	2.502	6.000	11.000	13.000	14.000	22.000
team2_third_down_conv_pct	0.402	0.143	0.000	0.308	0.400	0.500	0.818
team2_fourth_down_conv	0.692	0.872	0.000	0.000	0.000	1.000	5.000
team2_fourth_down_att	1.317	1.277	0.000	0.000	1.000	2.000	7.000
team2_fourth_down_conv_pct	0.368	0.422	0.000	0.000	0.000	1.000	1.000
team2_time_of_possession	1818.331	273.552	1160.000	1612.500	1819.500	2014.250	2553.000

When inspecting the summary statistics of the game statistics in the dataset, the numbers are quite much as expected for a dataset of NFL games. Some variables seem to have outliers in the data, such as rushing and passing yards which have to be kept in mind when analyzing the data. The normality in the data seems to be quite evenly distributed, as the mean and median (50 %) are relatively close for all variables.

3.3 Data pre-processing and variable selection

First, we had two separate datasets, namely the "Season" dataset and the "Game statistics" dataset. The ’Seasons’ dataset contains the final scores, location and team data of each game played in a particular NFL season. The ‘Statistics’ dataset then contains the detailed statistics of each individual game, such as the amount of yards gained, attempted passes and other descriptive data of the performance of the teams playing. Current NFL seasons have a total of 284 games counting the playoffs and Super Bowl, and the ‘Statistics’ dataset had 568 rows for each season, as each individual game was represented by two rows: one describing the statistics for each team in the game. The ‘Statistics’ dataset had to be modified so that each game was on an individual row. Thus, the rows were combined and the data modified to have ‘team1’ and ‘team2’ prefixes to indicate the statistics for each team.

To combine these datasets, we needed a common identifier to link the data from both sources. In this case, we had a common column called "nano" that uniquely identified each game. We merged the datasets based on this common column using the appropriate join operation, ensuring that we retained all relevant information from both datasets.

By combining the datasets, a new dataset was created that contained information from both the "Season" and "Game statistics" datasets for each game. This combined dataset provided us with a comprehensive view of each game, including the points scored by each team along with other relevant statistics.

Next, a new column called "team1_win" was created that would serve as the dependent variable for machine learning predictions. The goal was to predict whether "team1" (the first team mentioned in the dataset) won the game or not.

To create this column, the scores of "team1" and "team2" were compared for each game. If "team1" had a higher score, a value of 1 was assigned to "team1_win" to indicate that "team1" won the game. Otherwise, a value of 0 was assigned to indicate that "team1" did not win.

This new column became the target variable for training and evaluating machine learning models. We could use the other columns in the dataset, such as the game statistics and scores, as features to predict the outcome of future games.

By following these steps, the datasets were successfully combined, a new dependent variable was engineered ("team1_win"), and the data was prepared for further analysis and machine learning modeling. This preprocessing process allows to leverage the available information and build predictive models to forecast game outcomes based on historical data.

Various variables that had no analytical purpose were removed from the data to make it ready for machine learning models. Amongst them were data points such as the names of the teams and their geographical location, for example ‘Miami’ and ‘Dolphins’. These do not contribute any statistical purpose or mattered for the outcome so they were removed. The scores of both teams were also removed because they represented a variable that outright states whether or not team1 were victorious in the game, and would overfit the data, and show up as the most important features in the machine learning models without providing valuable insight into what in-game events actually contribute to the outcome of games.

3.4 Dealing with multicollinearity issues

After the data was engineered, there were 52 variables in the model to be used for predicting the results. Because many of the statistics are measuring closely related numbers (passing yards are part of total yards, rushing attempts are correlated to rushing yards), checking for multicollinearity issues was important with the dataset.

For example, the correlation between variables "team1_rush_att" and "team1_rush_yds" is 0.75. The trend continues for similar situations especially in statistics related to passing. The Variance Inflation Factor (VIF) finds variables that correlate heavily with one or several other variables that exist in the dataset. A higher VIF-score indicates that a higher amount of explanatory variables are behind multicollinearity issues. In this particular dataset with 52 variables, some variables even came out to have an infinite value. In the below table, the VIF-scores for each variable are marked down before the dataset was engineered further to remove multicollinearity.

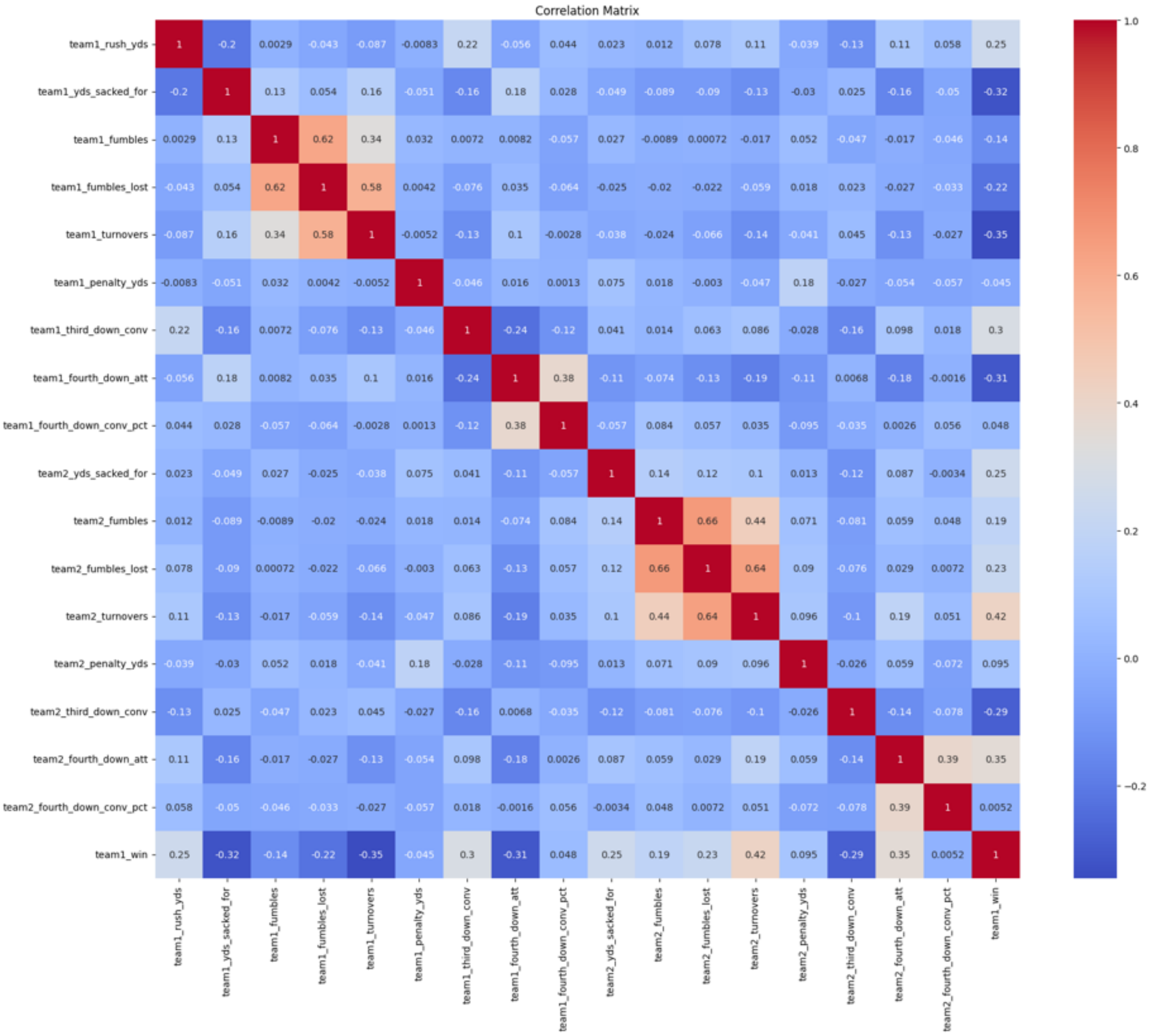
Table 5: VIF-scores before engineering variables

Variable	VIF
team1_rush_att	193
team1_rush_yds	2494026
team1_rush_tds	5
team1_pass_cmp	208
team1_pass_att	533
team1_pass_yds	inf
team1_pass_tds	17
team1_pass_int	inf
team1_passer_rating	257
team1_net_pass_yds	inf
team1_total_yds	18764260
team1_times_sacked	16
team1_yds_sacked_for	inf
team1_fumbles	4
team1_fumbles_lost	inf
team1_turnovers	inf
team1_penalties	24
team1_penalty_yds	19

team1_first_downs	215
team1_third_down_conv	219
team1_third_down_att	424
team1_third_down_conv_pct	179
team1_fourth_down_conv	12
team1_fourth_down_att	9
team1_fourth_down_conv_pct	6
team1_time_of_possession	340
team2_rush_att	237
team2_rush_yds	inf
team2_rush_tds	5
team2_pass_cmp	187
team2_pass_att	446
team2_pass_yds	inf
team2_pass_tds	15
team2_pass_int	inf
team2_passer_rating	185
team2_net_pass_yds	inf
team2_total_yds	inf
team2_times_sacked	19
team2_yds_sacked_for	inf
team2_fumbles	4
team2_fumbles_lost	inf
team2_turnovers	inf
team2_penalties	23
team2_penalty_yds	17
team2_first_downs	236
team2_third_down_conv	331
team2_third_down_att	596
team2_third_down_conv_pct	267
team2_fourth_down_conv	11
team2_fourth_down_att	8
team2_fourth_down_conv_pct	6
team2_time_of_possession	401

3.4.1 Dropping variables with high VIF-scores

One method of dealing with multicollinearity is to drop variables one at a time in the order of the one with the highest VIF-score. Dropping a variable will lower the VIF-score of other variables as well. The usually accepted VIF-score lies between the values 5 and 10. For this research, the threshold was set at 10 and a function was looped to remove the variable with the highest VIF-score, then the function recalculates scores for the remaining variables, and keep removing the highest scoring variable until all remaining variables are below the threshold of 10.



	team1_rush_yds	team1_yds_sacked_for	team1_fumbles	team1_fumbles_lost	team1_turnovers	team1_penalty_yds	team1_third_down_conv	team1_fourth_down_att	team1_fourth_down_conv_pct	team2_yds_sacked_for	team2_fumbles	team2_fumbles_lost	team2_turnovers	team2_penalty_yds	team2_third_down_conv	team2_fourth_down_att	team2_fourth_down_conv_pct	team1_win
team1_rush_yds	1	-0.2	0.0029	-0.043	-0.087	-0.0083	0.22	-0.056	0.044	0.023	0.012	0.078	0.11	-0.039	-0.13	0.11	0.058	0.25
team1_yds_sacked_for	-0.2	1	0.13	0.054	0.16	-0.051	-0.16	0.18	0.028	-0.049	-0.089	-0.09	-0.13	-0.03	0.025	-0.16	-0.05	-0.32
team1_fumbles	0.0029	0.13	1	0.62	0.34	0.032	0.0072	0.0082	-0.057	0.027	-0.0089	0.00072	-0.017	0.052	-0.047	-0.017	-0.046	-0.14
team1_fumbles_lost	-0.043	0.054	0.62	1	0.58	0.0042	-0.076	0.035	-0.064	-0.025	-0.02	-0.022	-0.059	0.018	0.023	-0.027	-0.033	-0.22
team1_turnovers	-0.087	0.16	0.34	0.58	1	-0.0052	-0.13	0.1	-0.0028	-0.038	-0.024	-0.066	-0.14	-0.041	0.045	-0.13	-0.027	-0.35
team1_penalty_yds	-0.0083	-0.051	0.032	0.0042	-0.0052	1	-0.046	0.016	0.0013	0.075	0.018	-0.003	-0.047	0.18	-0.027	-0.054	-0.057	-0.045
team1_third_down_conv	0.22	-0.16	0.0072	-0.076	-0.13	-0.046	1	-0.24	-0.12	0.041	0.014	0.063	0.086	-0.028	-0.16	0.098	0.018	0.3
team1_fourth_down_att	-0.056	0.18	0.0082	0.035	0.1	0.016	-0.24	1	0.38	-0.11	-0.074	-0.13	-0.19	-0.11	0.0068	-0.18	-0.0016	-0.31
team1_fourth_down_conv_pct	0.044	0.028	-0.057	-0.064	-0.0028	0.0013	-0.12	0.38	1	-0.057	0.084	0.057	0.035	-0.095	-0.035	0.0026	0.056	0.048
team2_yds_sacked_for	0.023	-0.049	0.027	-0.025	-0.038	0.075	0.041	-0.11	-0.057	1	0.14	0.12	0.1	0.013	-0.12	0.087	-0.0034	0.25
team2_fumbles	0.012	-0.089	-0.0089	-0.02	-0.024	0.018	0.014	-0.074	0.084	0.14	1	0.66	0.44	0.071	-0.081	0.059	0.048	0.19
team2_fumbles_lost	0.078	-0.09	0.00072	-0.022	-0.066	-0.003	0.063	-0.13	0.057	0.12	0.66	1	0.64	0.09	-0.076	0.029	0.0072	0.23
team2_turnovers	0.11	-0.13	-0.017	-0.059	-0.14	-0.047	0.086	-0.19	0.035	0.1	0.44	0.64	1	0.096	-0.1	0.19	0.051	0.42
team2_penalty_yds	-0.039	-0.03	0.052	0.018	-0.041	0.18	-0.028	-0.11	-0.095	0.013	0.071	0.09	0.096	1	-0.026	0.059	-0.072	0.095
team2_third_down_conv	-0.13	0.025	-0.047	0.023	0.045	-0.027	-0.16	0.0068	-0.035	-0.12	-0.081	-0.076	-0.1	-0.026	1	-0.14	-0.078	-0.29
team2_fourth_down_att	0.11	-0.16	-0.017	-0.027	-0.13	-0.054	0.098	-0.18	0.0026	0.087	0.059	0.029	0.19	0.059	-0.14	1	0.39	0.35
team2_fourth_down_conv_pct	0.058	-0.05	-0.046	-0.033	-0.027	-0.057	0.018	-0.0016	0.056	-0.0034	0.048	0.0072	0.051	-0.072	-0.078	0.39	1	0.0052
team1_win	0.25	-0.32	-0.14	-0.22	-0.35	-0.045	0.3	-0.31	0.048	0.25	0.19	0.23	0.42	0.095	-0.29	0.35	0.0052	1

Figure 4: Correlation matrix of variables at acceptable VIF-scores

After the data was engineered further, the dataset was left with 17 variables. After checking the correlation matrix as shown in Figure 3, the remaining variables do not have significant multicollinearity issues as before, but there are enough variables left to have a considerable selection for a machine learning model to analyze.

4 Results and analysis

4.1 Classification Models

The aim of this thesis is to see if the outcome of NFL games can be predicted with machine learning models, and which factors are most significant contributors to the outcomes of games. Due to the nature of the sport and the approach of the thesis, this can be treated as a binary problem: ‘Team 1’ either wins the game, or does not win (ties counted in the “not win” category, and there are a total of three tie games during the last two NFL seasons in the dataset). Due to the nature of the research, various classification models are suitable for predicting the outcome of games and feature importances in the dataset. The NFL game data is split into training and testing data.

Multiple machine learning models are suitable for the dataset of this thesis as it is a classification problem in its nature. For this particular research, the three chosen models were 1) Logistic Regression, 2) Random Forest Classification, and 3) Support Vector Machines.

4.2 Logistic Regression

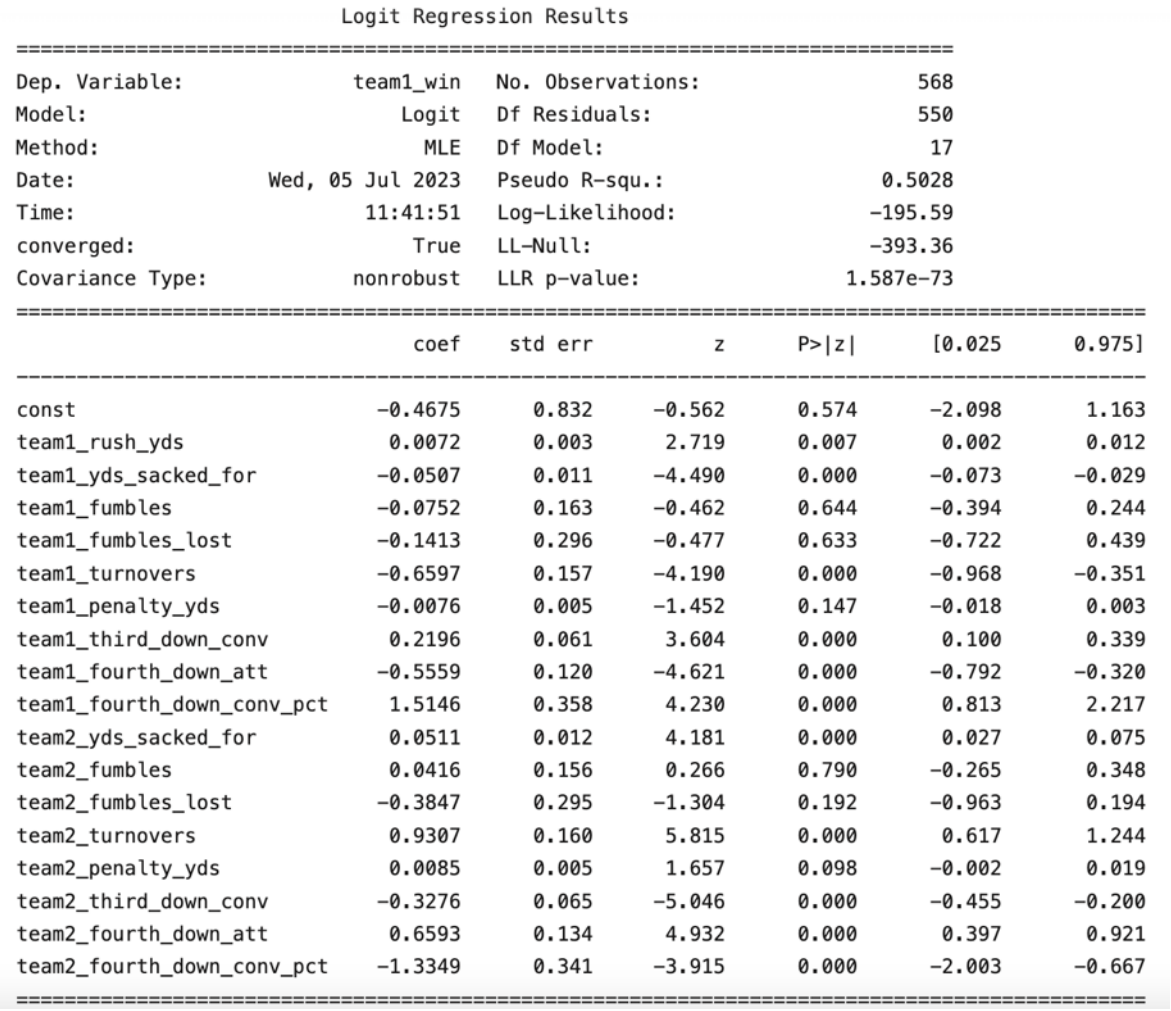
Logistic Regression is a widely adopted machine learning algorithm used for predictions. It is suitable for a classification problem, as linear regression is more suitable for the predicting a continuous value such as age, monetary value et cetera.

Logistic Regression has been applied for sports prediction previously in academic research as well. Egidi and Ntzoufras (2020) used logistic regression to predict the outcome of volleyball games. They used the data of the Italian SuperLiga 2017-2018 season matches, with a total of 182 matches and 14 teams included. The model was effective in predicting the outcome of matches, however the authors mention various limitations. Firstly, they mention that the model does not consider any further covariates to improve either the ability to interpret or the predictive power of the model. They acknowledge that the use of some team-specific covariates could be useful both to game explanation and to increase predictive accuracy.

Logistic Regression has been used further for betting purposes in a sport by Silverman and Suchard (2012), who used Conditional Logistic Regression to predict horse racing winners in Hong Kong. Historical data of 3681 horse races was used, and the data included 186 statistic points about each horse, such as days of rest since last race, change in weight of the rider, average speed in previous races, and whether the horse had changed their weight since the last race. 2994 races were used to train the regression model, and remaining 737 races (20% of the dataset) were used as the testing sample for the model. The authors achieved a maximum return on investment of 36.73% in betting with the model, so the model was quite successful in its use case.

Logistic Regression has also been used in the context of American football among other sports. As mentioned in the literature review of this thesis, Hsu et al. (2019) utilized logistic regression for predicting the outcome of field goals in late game situations, and Hill (2022) used logistic regression for outcome predictions in the CFL, a league comparable to NFL with some rule differences.

Seeing as there is previous literature that has used the same approach in other sports and also in gridiron football, logistic regression suits well for this thesis also from the viewpoint of having some previous research to reflect back on. After solving some of the multicollinearity issues pointed out earlier in this section of the thesis, the data was ready to be analyzed with a logistic regression model using Python's *"scikit-learn"* library.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:               team1_win   No. Observations:                  568
Model:                           Logit   Df Residuals:                      550
Method:                            MLE   Df Model:                           17
Date:                 Wed, 05 Jul 2023   Pseudo R-squ.:                  0.5028
Time:                         11:41:51   Log-Likelihood:                -195.59
converged:                        True   LL-Null:                       -393.36
Covariance Type:             nonrobust   LLR p-value:                 1.587e-73
==============================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         -0.4675      0.832     -0.562      0.574      -2.098       1.163
team1_rush_yds                 0.0072      0.003      2.719      0.007       0.002       0.012
team1_yds_sacked_for          -0.0507      0.011     -4.490      0.000      -0.073      -0.029
team1_fumbles                 -0.0752      0.163     -0.462      0.644      -0.394       0.244
team1_fumbles_lost            -0.1413      0.296     -0.477      0.633      -0.722       0.439
team1_turnovers               -0.6597      0.157     -4.190      0.000      -0.968      -0.351
team1_penalty_yds             -0.0076      0.005     -1.452      0.147      -0.018       0.003
team1_third_down_conv          0.2196      0.061      3.604      0.000       0.100       0.339
team1_fourth_down_att         -0.5559      0.120     -4.621      0.000      -0.792      -0.320
team1_fourth_down_conv_pct     1.5146      0.358      4.230      0.000       0.813       2.217
team2_yds_sacked_for           0.0511      0.012      4.181      0.000       0.027       0.075
team2_fumbles                  0.0416      0.156      0.266      0.790      -0.265       0.348
team2_fumbles_lost            -0.3847      0.295     -1.304      0.192      -0.963       0.194
team2_turnovers                0.9307      0.160      5.815      0.000       0.617       1.244
team2_penalty_yds              0.0085      0.005      1.657      0.098      -0.002       0.019
team2_third_down_conv         -0.3276      0.065     -5.046      0.000      -0.455      -0.200
team2_fourth_down_att          0.6593      0.134      4.932      0.000       0.397       0.921
team2_fourth_down_conv_pct    -1.3349      0.341     -3.915      0.000      -2.003      -0.667
==============================================================================================
```

Figure 5: Logistic Regression results for the dataset

After running the Logistic Regression model through the dataset, the model was given an accuracy score of 0.8596, with further results laid out in Figure 4. The Pseudo R-squared value of 0.5028 of suggests a moderate fit with the data, though comparison to the value in similar research should be considered to see if the value is good or bad. Six of the 17 coefficients also are shown to not be statistically significant according to the results summary. The variables were scaled to make sure the feature importance is not impacted by the relative numbers being of varying ranges in the variables. However, consideration should be taken if the factors are truly insignificant, or can be affected by example by the sample size of two NFL seasons. Fumbles, and turnovers in genral are considered to be critical in the sport, as a lost ball from a fumble essentially means forfeiting one opportunity to score points, and gives the opposing team a chance on the other hand.

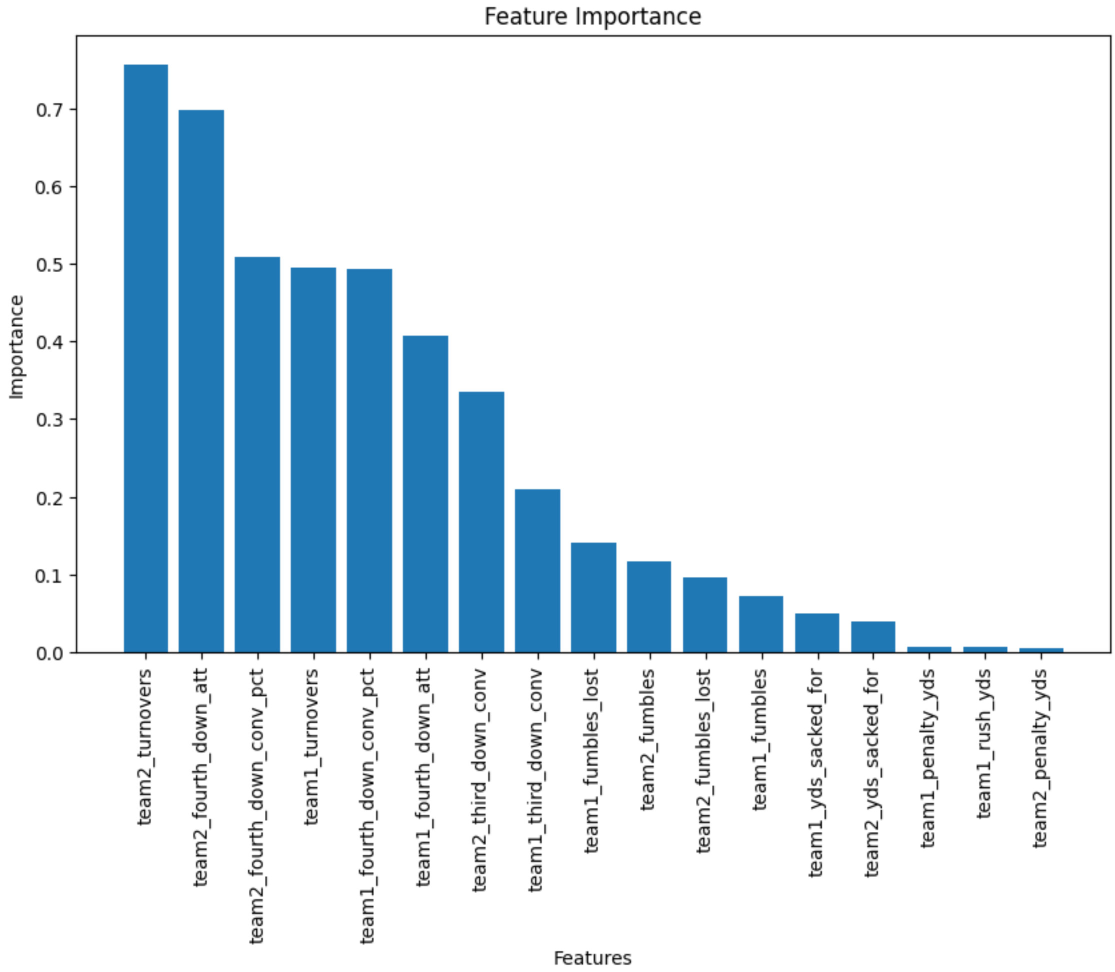


Figure 6: Feature Importances of Logistic Regression Model

The most significant variables affecting the result according to the model are the opposing team’s turnovers, fourth down attempts, followed by how well they convert, and also the team’s own turnovers. Overall, the model provides reasonable results and explains the relationship of the most significant variables involved, but it is reasonable to also utilize other suitable machine learning models to measure their performance, and then compare to the results of the logistic regression model.

4.3 Random Forest Classifier Model

The second model implemented in this thesis to predict the outcome of NFL matches is Breiman's Random Forest Classifier (RFC), introduced by Leo Breiman in a 2001 research paper. Since its introduction, RFC has been widely used by academia for classification problems. In the sports context, Jia et al. (2020) used a Random Forest Regression Model to predict the winner of Summer Olympics Games winners with data from the 1964-2016 Summer Olympics medal lists. Vistro et al. (2019) on the other hand used a RFC model with historic data of the 2008-2017 seasons of the Indian Premier League in cricket.

Random Forest Classifier model has an advantage over the Logistic Regression model in the aspect that multicollinearity issues do not have to be taken into account to the same extent. All of the original variables in the dataset can be included in the RFC model. Otherwise similar insights can be gathered from the model, as RFC is also able to output feature importances, describing which variables are more significant than others in determining the outcome of games.

The RFC model was also built using Python's *"scikit-learn"* library. There are a number of different options, including number of estimators (trees in the forest), minimum samples, and maximum tree depth. The model is run with default numbers, such as 100 estimators and tree depth not being limited. The test size is 20% of the dataset similar to the Logistic Regression model.

Table 6: Perfomance metrics for the RFC model

Accuracy	**Precision**	**Recall**	**F1-score**
0.9035	0.8889	0.9057	0.8972

Based on the performance metrics of the RFC model above, the results seem promising. The overall accuracy of the model in being able to make correct predictions is roughly 0.90, with the precision score in the same area, indicating the model is able to avoid false positives quite well. The values for Recall and the F1-score are also promising, and the model is able to make quite accurate predictions overall from the data, and not skewing towards false positives or negatives.

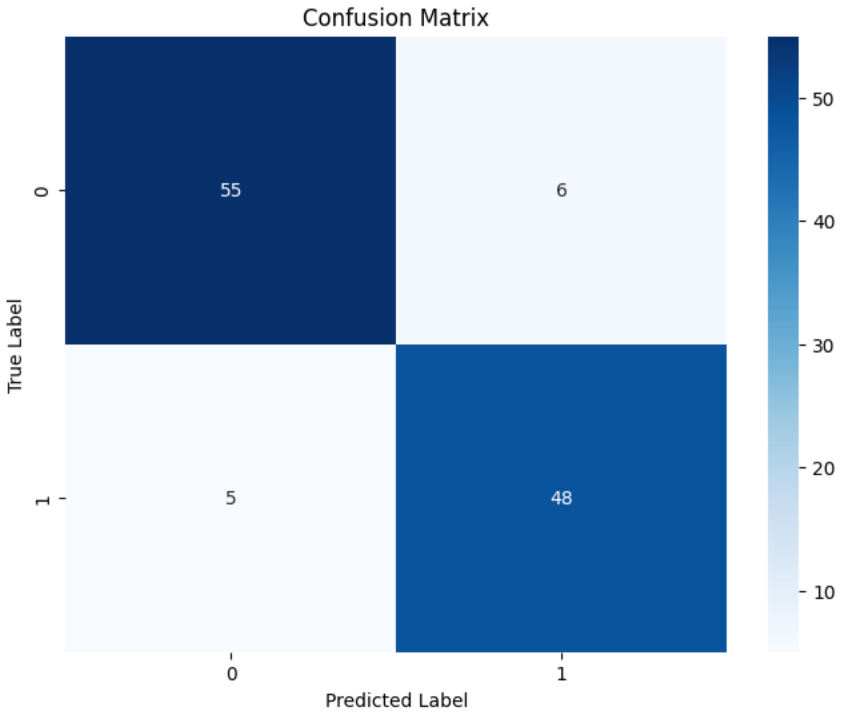


Figure 7: Confusion matrix of the RFC model

The confusion matrix in Figure 6 indicates that for the test dataset, the RFC model made 103 correct predictions, 5 false negative predictions, and 6 false positive predictions.

Another advantage of the RFC model is that the decision process can be visualized as a decision tree map directly in Python which can help in grasping the full process. With 57 variables, visually presenting the entire tree would be impossible, but below in Figure 7 there is a smaller section of the decision tree visualized with a maximum depth of 2. In this tree, the model would determine whether the number of rushing attempts of the opposing team was equal to or below 25.5 attempts, and then move on to the next layer which is divided to the first team’s pass attempts or first downs achieved during the game.

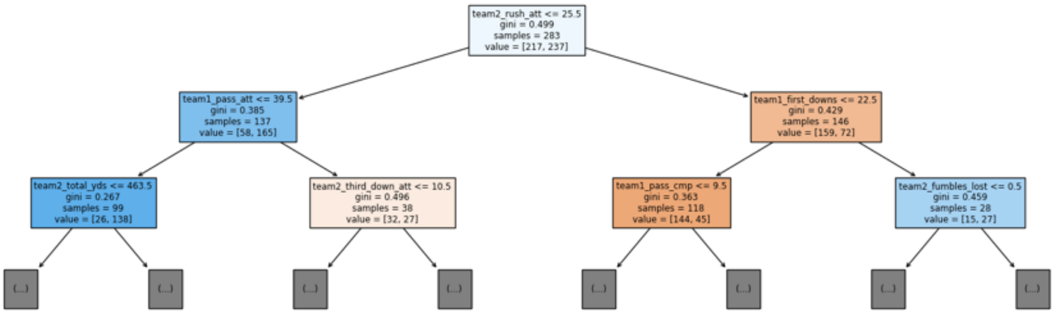


Figure 8: Snapshot of RFC model's decision tree

The RFC model is also able to indicate feature importances. The importances are visualized in Figure 8, and the most significant variables according to the RFC model seem to be: the opposing team’s passer rating, the number of rushing attempts by both teams as the second and third most important variables. The results are quite different from the Logistic Regression’s feature importance list, explained largely by the fact that many of the variables were removed to solve multicollinearity issues. The most important features in the RFC models were all removed in that analysis. The RFC model likely offers more explanatory power in this regard as it is able to retain more variables, and they also are important for indicating the outcome of games.

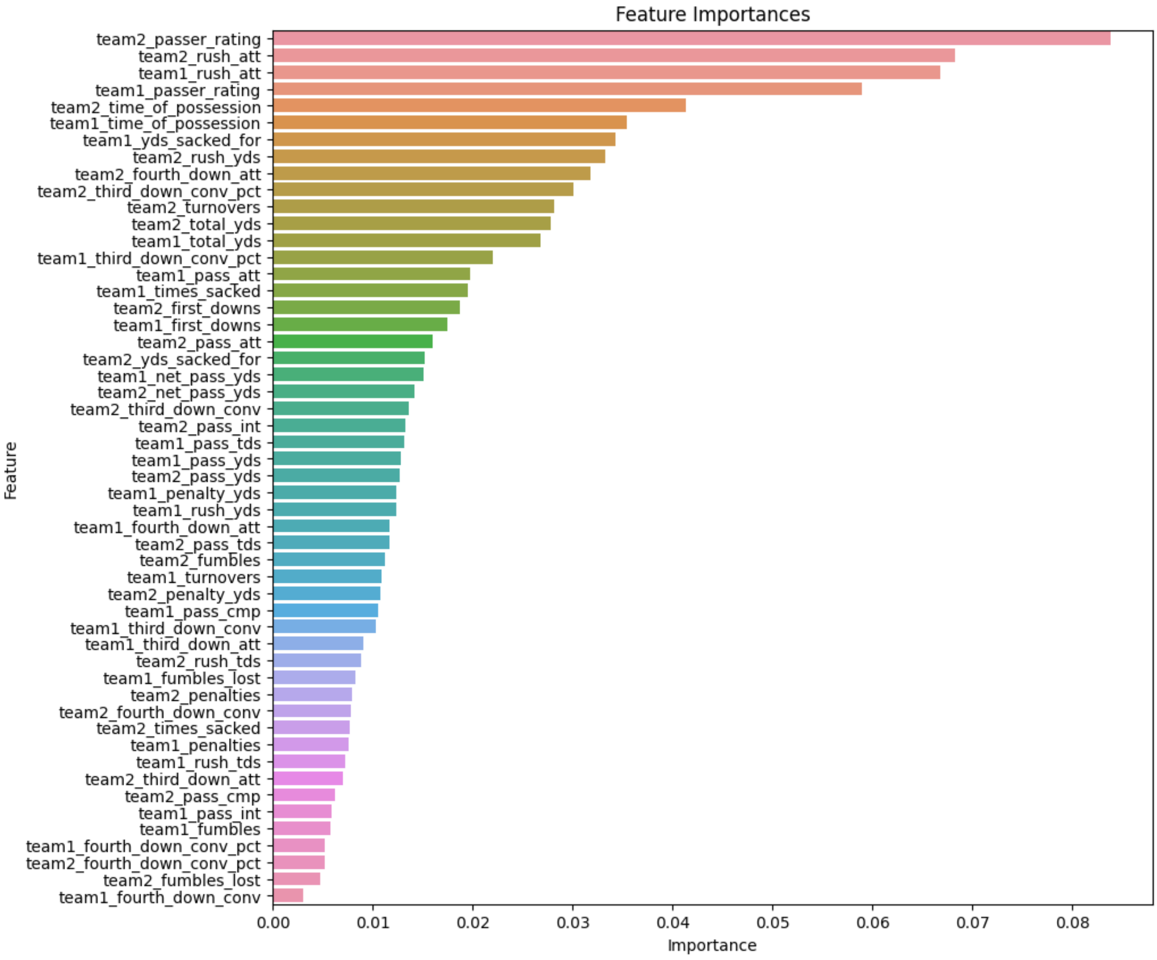


Figure 9: Feature importances of RFC model

4.4 Support Vector Machine

While in this thesis process two suitable machine learning models have been used to analyze the dataset, there are also other models remaining that are suitable for a classification problem such as sports game predictions. One such model is the Support Vector Machine algorithm (SVM). SVM is a popular classifier model based on a linear discriminant function, and is well suited for binary classification. (Murty & Raghava, 2016).

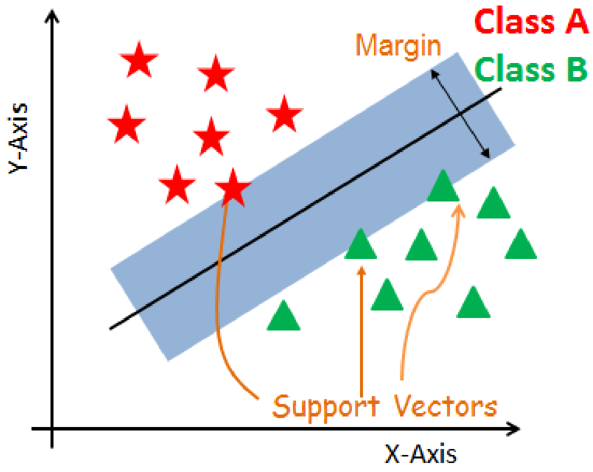


Figure 10: Visual represantion of how an SVM model works: it draws a separator line among the data to classify into binary categories (Datacamp, 2019).

An SVM model maps data into a high-dimensional space, where it is able to conclude a separator for classifying to which binary category the data point belongs to. The model then develops a hyperplane to separate the data, and is able to determine to which category any test data belongs to. (IBM, 2021) As in previous models utilized, the model is deployed by using Python's *"scikit-learn"* library. Similar parameters are used again for training the model, with 20% of the dataset serving as the test sample.

Table 7: Performance metrics of the SVM Classifier model

Accuracy	Precision	Recall	F1-score
0.9561	0.9138	1.0	0.9550

The performance metrics for the SVM model represented in Table 6 give promising results. Recall is exactly one, indicating that the model resulted in no false positive predictions. The overall accuracy of the model is also roughly 0.95, indicating an accurate model.

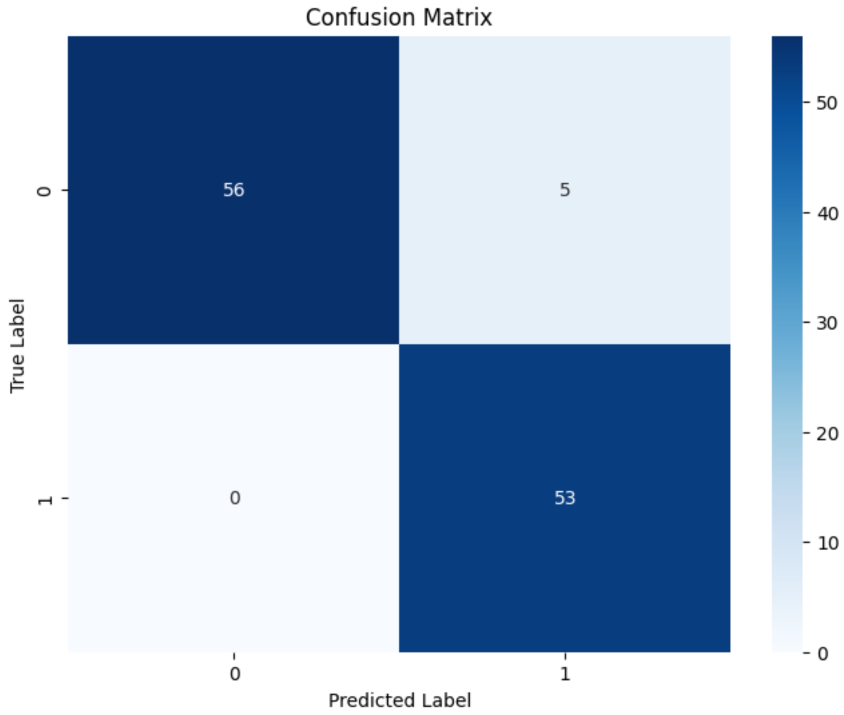


Figure 11: Confusion matrix of the SVM Classifier model

The SVM model cannot indicate feature importances in similar way as the Logistic Regression and RFC model can, but there are other methods to judge the impact of particular variables to the model. Analyzing the coefficients of the support vectors can indicate whether a variable has a positive or negative impact on the decision boundary. The variables were also scaled during the analysis process as the differing range of variables can impact the decision boundary in the SVM model. For example, passing yards are measured often in the hundreds, but passing attempts range between 20 and 50 for most teams.

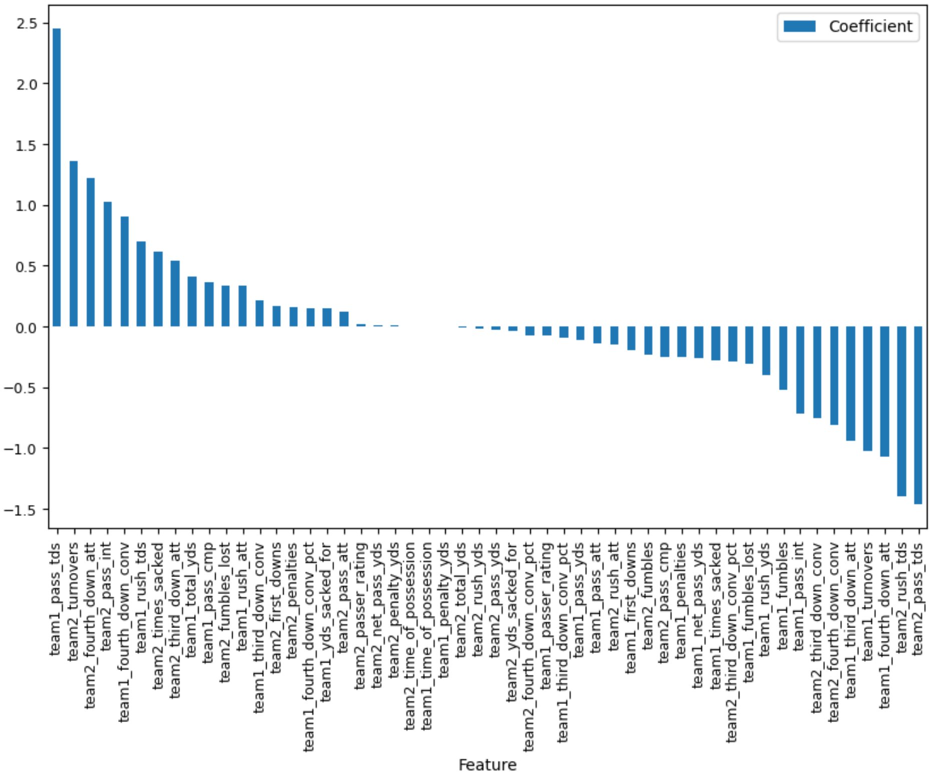


Figure 12: Coefficient analysis of the SVM Classifier model

The sign and magnitude of the coefficients in Figure 10 indicate the direction and strength of the influence of each feature on the decision boundaries. Positive coefficients indicate that an increase in the feature value leads to a higher probability or likelihood of belonging to a certain class, while negative coefficients indicate the opposite. Passing touchdowns and turnovers seem to be among the most significant variables responsible for impacting the judgment of the SVM model.

5 Discussion of results

This thesis has deployed three different machine learning models for predicting the outcome of NFL games. Each model has its own advantages and disadvantages, providing also slightly different results. All three performed relatively well, but there are some differences and implications in data engineering that also need to be considered when deciding which model best fits this purpose according to the results. It is interesting to note that each model considered different variables to be most influential to the predictions they make.

5.1 Comparison of the models

A stark difference between Logistic Regression and the two other models is that many variables were dropped from the dataset to solve multicollinearity issues. This makes the logistic regression model quite a bit simpler than the other two used in this research, however it can still provide useful information on what variables are most influential, and as a simpler model may be easier to implement on other datasets, especially as not as many variables are needed. The performance metrics also indicate that the LR model could predict results dependably, with an accuracy score of 0.8596. Considering the dataset includes the latest two seasons of NFL football, the models could be further evaluated by using previous NFL seasons that featured a 16-game regular season, and in the future the models can be tested against the upcoming NFL seasons.

Advantages of the RFC model include the fact that it is able to deal with multicollinearity better. All 57 variables are included in the RFC model, and as an ensemble method it is able to combine multiple decision trees to capture more complex relationships in the data. While the entire tree cannot be presented in this thesis, mapping it out can provide useful insights into the relationships of variables that are not presented in the other models.

The most accurate model in this thesis was the SVM Classifier model. Support Vector Machine model has also the advantage of dealing better with multicollinearity, but one particular benefit of the SVM model is its robustness against outliers. NFL games are prone to sometimes having some unique statistical results, such as one game in the dataset featuring exactly one rushing yard for a team, when the mean of the entire dataset was 118 rushing yards. SVM model can be the easiest to deploy into additional data without a need to take into account significant outliers in the data.

Comparing the AUC-ROC curves of the RFC and SVM models in Figures 13 and 14, the performance of the two models is close. The SVM model has a better Area Under the Curve, indicating better predictive ability. While a small difference, if the models would be used for example for betting purposes, the SVM model may be notably better at generating returns when combined with the expected value of betting odds.

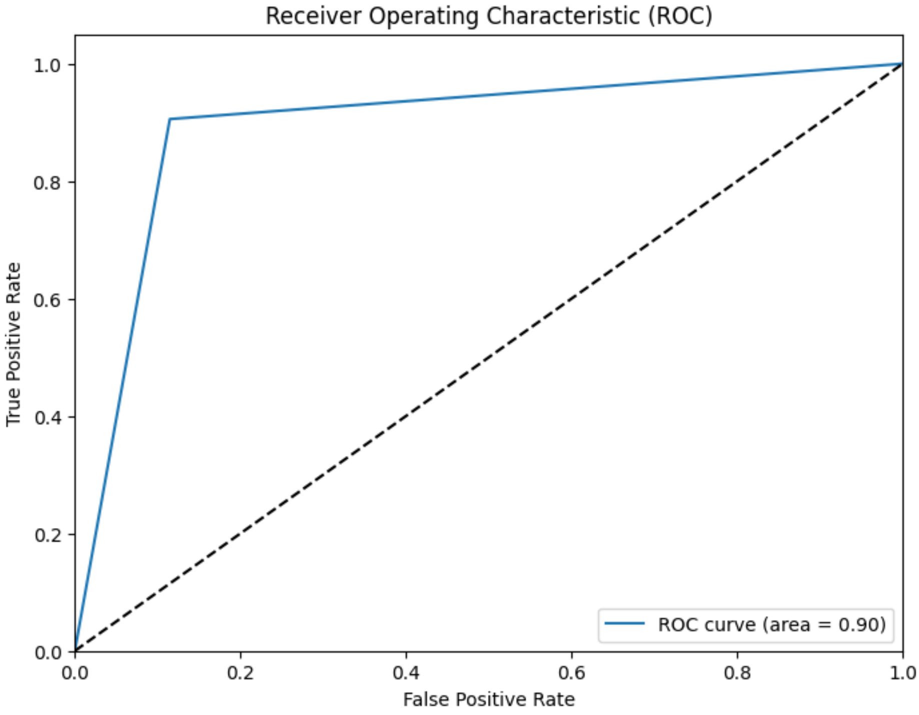


Figure 13: ROC curve of the RFC model

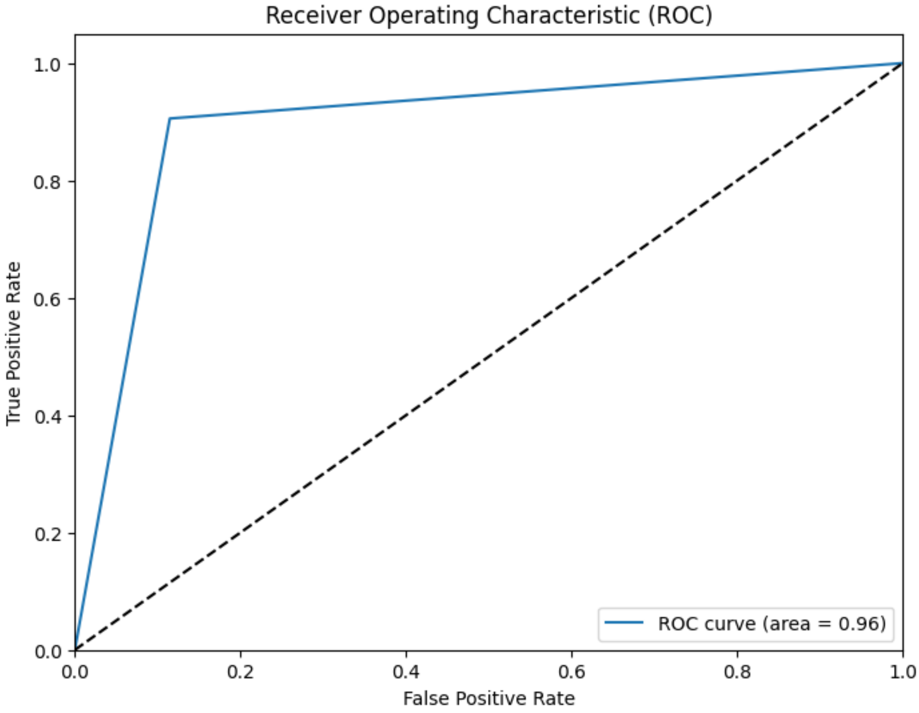


Figure 14: ROC curve of the SVM Classifier model

5.2 Performance of models compared to prediction in other sports

It is notable that all three of the models used during this thesis performed well on the dataset. The accuracy of the models were all above 85%, with RFC and SVM models above 90%. This is somewhat higher than the accuracy of models that were part of the literature review: models that analyzed American or Canadian Football includes Hill's (2022) research, which had an accuracy of 73%, while Gifford and Bayrak (2020) were able to predict the result of the game 83% of the time.

In other sports, Zhang and Abdelhamid (2019) achieved accuracies ranging from 73% to 80% when trying different machine learning models on NBA game data. Egidi and Ntzoufras (2020) analysed volleyball games with a Bayesian model, achieving an accuracy of 78% in predictions. In cricket, Vistro et al. (2019) achieved an 80% accuracy with a Random Forest Classifier model.

An explanation for possibly why the models were so accurate for the NFL dataset in this thesis can at least partially be explained by the fact that many statistics in American Football that are available and measured for each game are directly impactful to performance of the teams. More yards gained, less turnovers, more completed passes all point out to the team having possession of the ball for longer periods, often leading to more opportunities to score, and also less opportunities for the opposing team to score. By nature American Football games have many statistic points recorded that can directly be used to assess which team is performing better, which may not always directly guarantee a win, but can strongly point towards that direction.

6 Conclusions

Using machine learning to predict different aspects of sports matches, including the results has been a popular topic academically, but also for commercial purposes outside of betting. NFL has partnered with Amazon Web Services for many machine learning applications in the game of American football, and other sports have also adopted machine learning for purposes such as infotainment: a live win percentage to indicate the likelihood of a team winning, possibly boosting the impressiveness of an unlikely event happening during a match, with an example in Figure 15 of one such graphic, provided by a sports analytics called Opta (theanalyst.com, 2021).

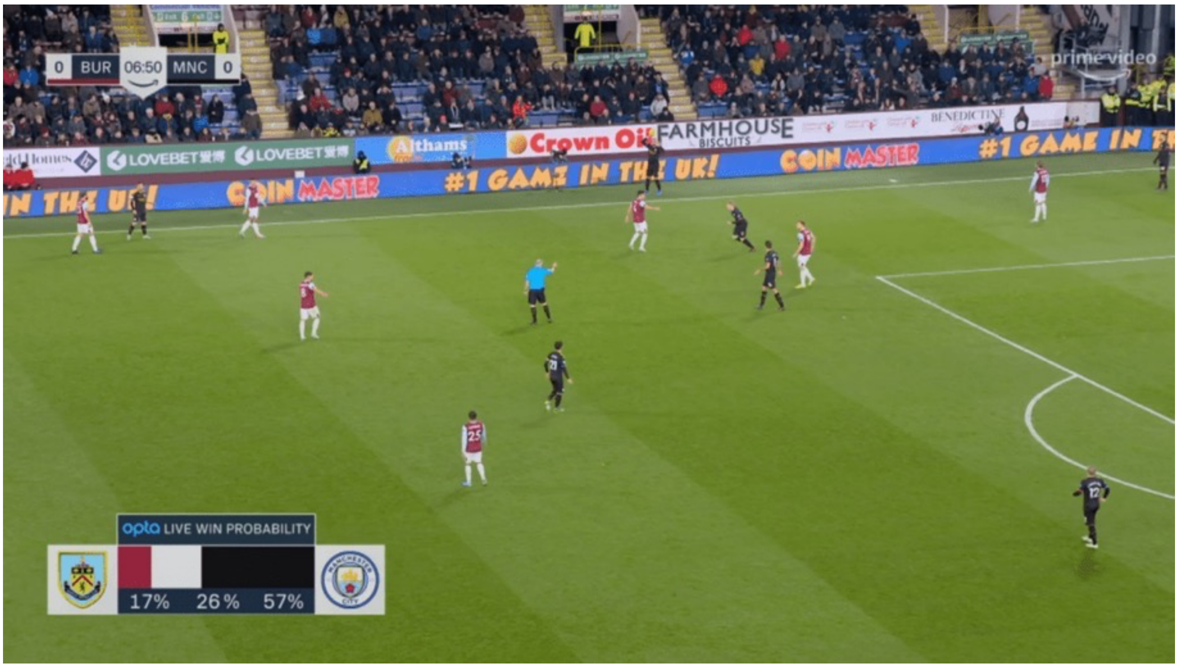


Figure 15: Live win probability between Burnley and Manchester City in an Association football match

The research objectives of this thesis were to determine:

- Can machine learning models predict the outcome of NFL matches reliably
- Can the most influential statistical variables that affect the outcome of games be determined

Based on the results, it can be stated that machine learning models are reliably able to predict the outcome of games based on game statistics. While many statistics gathered are very indicative of performance (yards gained often contribute to more scoring opportunities), they cannot always explain the outcome of games reliably. More interestingly, the models are able to indicate which statistics are most important for the individual models. While the research conducted by Gifford and Bayrak (2020) for example indicated turnovers as the most significant factor, it was not the most important feature for two of the models in this thesis, though turnovers were among the top features for all models. Turnovers for the opposing team was the most important feature in the Logistic regression model. This provides an interesting reflection, especially as Gifford and Bayrak's research was conducted before the extension of the regular season. The results of the thesis provide an additional viewpoint to the latest changes in rules and season length.

There is strong evidence in the results of this thesis that the models used can be applied to predictive analytics and for making predictions on future NFL games. Compared to past literature, the results provide a recent look into what factors influence the outcome of games in the current state of the NFL. Rule changes and the evolution of playbooks is likely to change the dynamics of the games, along with the inclusion of an additional regular season game. These models are a good fit to use for the current NFL games according to the results.

Applications for use of the models can include optimizing the point predictions for the fantasy leagues, or providing infotainment material such as in Figure 15. Betting services could include the application of these models to adjust the odds of different bets to provide an efficient betting market.

Other more niche applications could be for individual teams to analyze their performance: possibly extracting their own games and running analysis on those games, perhaps extending the dataset by a few more recent seasons. This could provide insight to coaching staff and team management if particular factors arise in their games to be most significant.

Another niche field is video games which could use the results and data to adjust for example the level of play a computer-controlled opponent has against a player, or adjusting the internal game mechanics for example how likely certain events like turnovers happen by random when a contact event occurs within the game.

For predicting the results of future games, the models could be applied perhaps to halftime statistics of upcoming NFL games to test out the predictive capabilities.

7 Limitations

This part of the thesis will consider the most significant limitations related to the research, methodology and available data used. The two main limitations in this thesis are:

- Multicollinearity issues in the dataset
- Availability of variables and data points

When implementing a logistic regression model for this dataset, the model ran into serious multicollinearity issues as many statistics are closely related (total yards consisting of rushing and passing yards, for example) which are natural for a sports statistic dataset, but create complications for analysis. One method that is possible to combat this is PCA feature clustering to group similar variables into a single variable in the dataset, which was not done in this thesis. A related technique of grouping variables was considered, but ultimately decided that it has some drawbacks, such as losing nuance if for example passing and rushing yards were combined to total yards. While this could help with the issue, it was decided that eliminating factors with the VIF method was the more suitable option in this thesis. There are also other forms of tackling multicollinearity issues not utilized for this thesis, such as Ridge regression which is particularly used for linear regression problems, shrinking some coefficients to reduce their impact. Another similar solution is Lasso regression which works in a similar manner, but also eliminates some coefficients completely to achieve the effect. Ultimately, collecting more data is also a simple method that can reduce multicollinearity as it can help diversify the ranges of the variables. Particularly in this thesis' situation as we are taking the data of only two NFL seasons, further analysis can tackle the problem possibly by simply having more data.

For replicating the results of this thesis, this creates issues with narrowing down the available datasets for further analysis to other possible NFL seasons or other season data. For this thesis however, using the Variance Inflation Factor method alone was deemed sufficient as the analysis was also done with two other machine learning models that do not suffer from multicollinearity. If more models were prone to multicollinearity issues, or the performance of the Logistic Regression model was low after solving for multicollinearity, applying additional methods would have been more necessary.

Availability of variables is another remarkable limitation. This thesis relies on a dataset of data from Sports Reference, which features box score statistics and other easily available data. This rules out other interesting factors that could have an impact on games, such as weather conditions, playing surface (grass or artificial turf), among many other interesting nuances. Additionally, key player injuries can have an impact on game outcomes that is invisible to this dataset. On a more advanced level, having data of starting lineups, possible injuries players are playing through, are just some of the variables that could provide a deeper level of understanding outcomes in games.

8 Further Research

The use of machine learning models to predict sports events and outcomes is likely to continue trending in academic areas as well as commercial purposes. The main research question of the thesis was to see if machine learning models can predict the outcome of matches and it seems well supported that this is possible.

Some of the possibilities for further research include is incorporating additional seasons into the models. Now only the last two NFL seasons were included, but the model could be tested back on previous years. It could be interesting to see how far back the model can remain accurate, and when the game dynamics have been so different that the accuracy starts to decrease notably. Another opportunity is to take data from either the NCAA Division 1 College Football games, that is the closest equivalent to NFL games, or even the Canadian Football League (CFL), and run the predictive models there. Alternatively, the same machine learning models could be trained on recent season data from the NCAA and CFL games and then compare if there are differences in feature importances for example.

More advanced models, such as XGBoost or combination of different machine learning algorithms is another avenue of research that could dive down further into the subject area.

References

Books and reports

Murty, & Raghava, R. (2016). *Support Vector Machines and Perceptrons Learning, Optimization, Classification, and Application to Social Networks* (1st ed. 2016.). Springer International Publishing. https://doi.org/10.1007/978-3-319-41063-0

Articles

Bai, Gedik, R., & Egilmez, G. (2022). What does it take to win or lose a soccer game? A machine learning approach to understand the impact of game and team statistics. *The Journal of the Operational Research Society*, *ahead-of-print*(ahead-of-print), 1–22. https://doi.org/10.1080/01605682.2022.2110001

Korpimies, S. (2020) Predicting players' success on the PGA-Tour.

Hsu, Liu, K.-S., & Chang, S.-C. (2019). Choking under the pressure of competition: A complete statistical investigation of pressure kicks in the NFL, 2000–2017. *PLoS ONE*, *14*(4), e0214096–e0214096. https://doi.org/10.1371/journal.pone.0214096

Bock. (2016). Empirical Prediction of Turnovers in NFL Football. *Sports (Basel)*, *5*(1), 1–. https://doi.org/10.3390/sports5010001

Wilson. (2020). College Football Overtime Outcomes: Implications for In-Game Decision-Making. *Frontiers in Artificial Intelligence*, *3*, 61–61. https://doi.org/10.3389/frai.2020.00061

Thabtah, Zhang, L., & Abdelhamid, N. (2019). NBA Game Result Prediction Using Feature Analysis and Machine Learning. Annals of Data Science, 6(1), 103–116. https://doi.org/10.1007/s40745-018-00189-x

Mohsin, & Gebhardt, A. (2022). A stochastic model for NFL games and point spread assessment. *Journal of Applied Statistics*, *ahead-of-print*(ahead-of-print), 1–14. https://doi.org/10.1080/02664763.2022.2120973

Shank. (2019). NFL Betting Biases, Profitable Strategies, and the Wisdom of the Crowd. *International Journal of Sport Finance*, *14*(1), 3–12. https://doi.org/10.32731/IJSF/141.022019.01

Gifford, M., & Bayrak, T. (2020). What Makes a Winner? Analyzing Team Statistics to Predict Wins in the NFL. *AMCIS 2020 Proceedings*. 35.

https://aisel.aisnet.org/amcis2020/data_science_analytics_for_decision_support/data_science_analytics_for_decision_support/35

Majumdar, Bakirov, R., Hodges, D., Scott, S., & Rees, T. (2022). Machine Learning for Understanding and Predicting Injuries in Football. *Sports Medicine - Open*, *8*(1). https://doi.org/10.1186/s40798-022-00465-4

Borghesi. (2007). The home team weather advantage and biases in the NFL betting market. *Journal of Economics and Business*, *59*(4), 340–354. https://doi.org/10.1016/j.jeconbus.2006.09.001

Borghesi. (2007). The late-season bias: explaining the NFL's home-underdog effect. *Applied Economics*, *39*(15), 1889–1903. https://doi.org/10.1080/00036840600690314

Paul. (2017). The impact of atmospheric conditions on actual and expected scoring in the NFL. *International Journal of Sport Finance*, *12*(1), 14–14.

Lock, D. (2016). *Statistical methods in sports with a focus on win probability and performance evaluation* (Doctoral dissertation, Iowa State University).

Hill. (2022). In-game win probability models for Canadian football. *Journal of Business Analytics*, *5*(2), 164–178. https://doi.org/10.1080/2573234X.2021.2015252

Egidi, & Ntzoufras, I. (2020). A Bayesian quest for finding a unified model for predicting volleyball games. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *69*(5), 1307–1336. https://doi.org/10.1111/rssc.12436

Silverman, & Suchard, M. (2013). PREDICTING HORSE RACE WINNERS THROUGH A REGULARIZED CONDITIONAL LOGISTIC REGRESSION WITH FRAILTY. *Journal of Prediction Markets*, *7*(1), 43–52. https://doi.org/10.5750/jpm.v7i1.595

Breiman, L. (2001). Random Forests. *Machine Learning* **45**, 5–32. https://doi.org/10.1023/A:1010933404324

Jia, Zhao, Y., Chang, F., Zhang, B., & Yoshigoe, K. (2020). A Random Forest Regression Model Predicting the Winners of Summer Olympic Events. *Proceedings of the 2020 2nd International Conference on Big Data Engineering*, 62–69. https://doi.org/10.1145/3404512.3404513

Vistro, D.M., Rasheed, F., & David, L.G. (2019). The Cricket Winner Prediction With Application Of Machine Learning And Data Analytics. *International Journal of Scientific & Technology Research, 8*, 985-990.

Internet-references

Posti (1999). Online. Available at: www.posti.fi, [15.2.2000].

IBM (2021). Online. How SVM Works. Available at: https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works

Sports Reference (2023). Online. About Sports Reference. Available at: https://www.sports-reference.com/about.html

Whitmore, J. (2021). Explaining Live Win Probability (LWP). *Opta Analyst* Available at: https://theanalyst.com/eu/2021/11/live-win-probability/