Machine Learning and Computational Statistics

DS-GA 1003

Spring 2018 Project Report

PREDICTING RESULT OF NCAA FOOTBALL GAMES

TEJAL LOTLIKAR
New York University
Courant Institute of Mathematical Science
tl2482@nyu.edu

DIWAKAR PALIWAL
New York University
Courant Institute of Mathematical Science
dp2757@nyu.edu

KUMAR MEHTA
New York University
Courant Institute of Mathematical Science
kjm627@nyu.edu

Advisor:

DAVID FROHARDT-LANE

Abstract

We present a sports prediction model that predicts the winning team and average margin of victory in NCAA football games. Our goal is to analyse the performance all games played by both the teams playing current match and predict the winning probability of each team in the current game using those statistics. We consider various data features (like points scored by a team against all teams, goal attempts, penalty etc) of each team that primarily influence the result of a game and use them in various data models to improve the accuracy of prediction. Along with the winning probability, we also predict the average margin victory in the current game by using different models of the data using appropriate features. We found which of the features from entire dataset are important as instruments using different feature selection algorithms.

We use cbstats.com dataset containing information on College Football Statistics from 2005 to 2013 for this analysis and prediction model.

1. Introduction

There are total 247 NCAA football teams in all seasons from 2005 to 2013 and each team has a unique team code. Many teams have been added over seasons. The games between these teams are organized according to unique game codes. Each game has different parameters like location of game (that decides which team is home team & which is visiting team), duration of game, attendance of each game along with actual game statistics(points scored, goals attempts made, actual goals scored, etc).

The basic factors that shape our model consist of each team's (home team and visiting team) performance in previous games and using that data we predict their individual winning probabilities. These factors vary from one season to another season and hence our model gives more weight to current data with gradually decreasing weights as we consider previous seasons' data. We use different algorithms like Logistic regression, Support Vector Machines & Neural Networks to make predictions of outcome of match and evaluate performance of each of these techniques. We predict margin victory by using Gradient Boosting regression technique and Neural Networks. Finally we analyze our models based on their individual losses and accuracy.

2. Datasets

We use complex NCAA football games data-set compiled by cbstats.com which has information on College Football Statistics from 2005-2013, recovered via wayback link. The data has statistics about approximately 800 games played between 240 teams, for every season from 2005-2013. Of these 240 teams, about half of them are FBS (Football Bowl Subdivision) teams and the other half are FCS (Football Championship Subdivision) teams. FBS teams have high number of matches (approximately 10-12 games) per season. FCS team are mentioned only when the match happened against a FBS team and therefore the data is very limited for these teams (0-2 games per team). Hence, our prediction is focused only for FBS teams and we do not consider games that includes FCS teams.

Each season's data has information on various fields and details as follows:

1. **Game statistics:**
   Home and visiting teams, points scored, detailed information on every rush, punt, reception, kickoff, pass, etc.
2. **Teams:**
   Information of team composition, combined statistics for the team in games like rush yards, number of passes completed, kickoff return yard, field goals made, tackles, fumbles, etc.
3. **Players:**
   Information about player positions, statistics of each player in every game like number of points scored, number of tackles, number of tackle assists, number of fumbles, etc.
4. **Environment:**
   Information on ground-surface, total attendance for the game, stadium statistics, total time duration of the game, etc.

Link to the dataset:
https://old.datahub.io/dataset/college-football-statistics-2005-2013

3. Data Preprocessing / Feature Extraction

3.1 Data Cleaning:

All the data was accumulated for each season from 2005 to 2013 from the datasets in the form of csv files. For data cleaning, we identified the records with outliers that are the extreme values falling way outside the range of other observations of each feature. If the number of outliers is very high, this can lead to overfitting of data and hence it is recommended to minimize the outliers. We did data normalization to bring the value of all features in the range of zero to one. This normalized data was then analyzed to identify and eliminate the records with outliers for individual features from our training data.

Box plot for features were plotted to check for the outliers. All the data points below and above from 25 percentile and 75 percentile respectively by 1.5 times Interquartile Range (75 Percentile - 25 Percentile) were removed.The representation of features before and after outlier removal is depicted below in Figure 1 and 2.

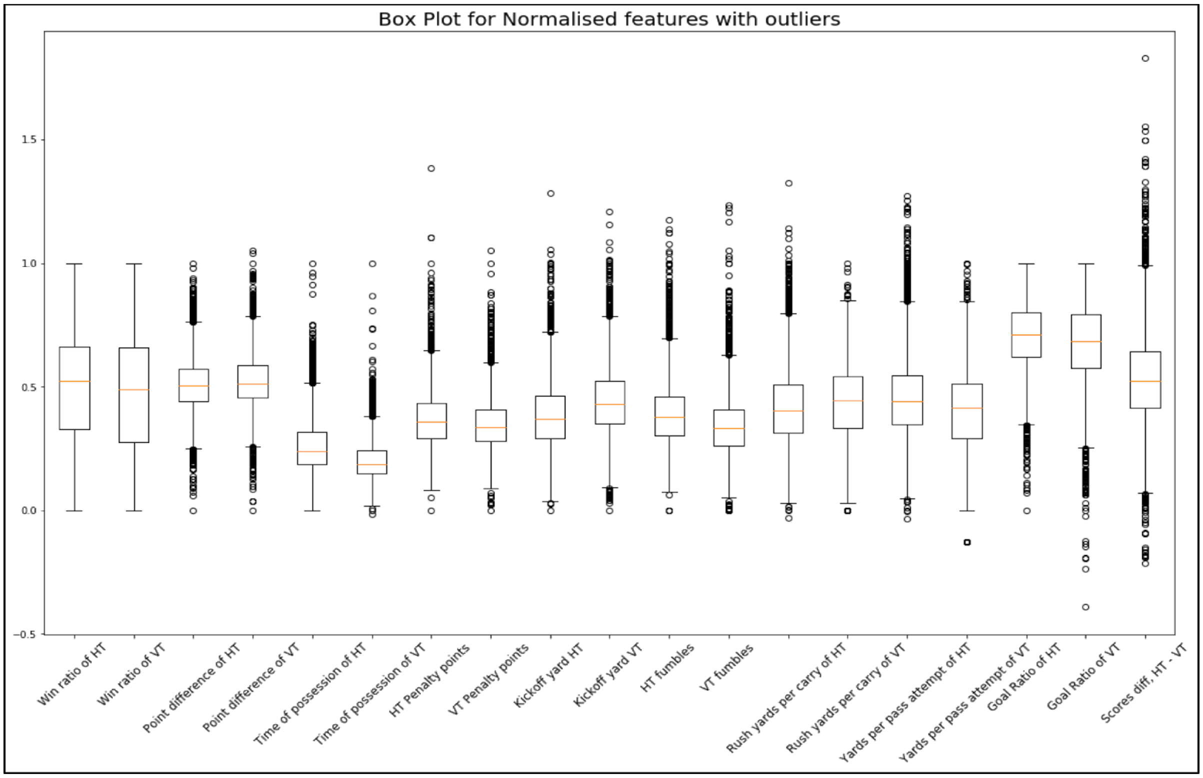


Fig 1. Box plot for Normalized data features with outliers

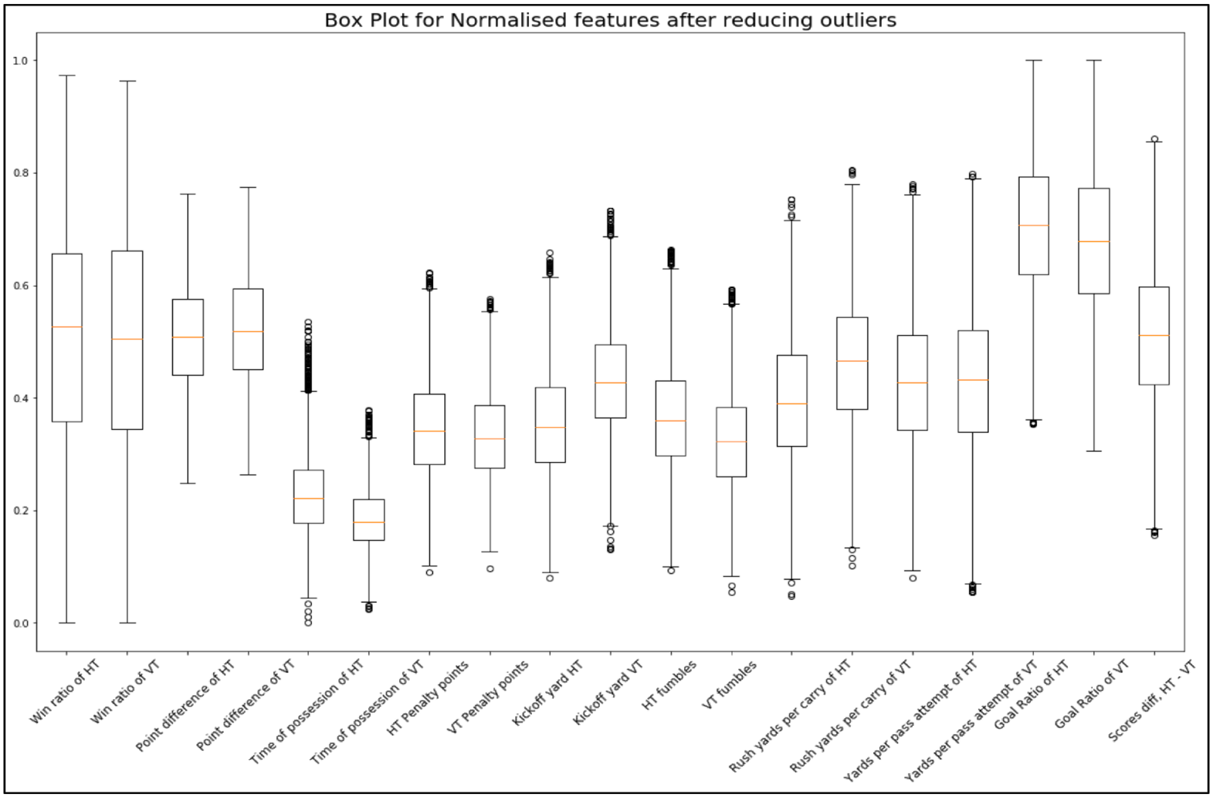


Fig 2. Box plot for Normalized data features after reducing outliers

3.2 Feature Selection

Feature Selection is a technique to identify and select the features that have high discrimination by understanding what aspects of complete dataset are significant in the prediction making and which aren't. We analyzed individual features based on their test accuracy as shown in below figure and extracted the relevant features to be used in final X_DATA (input examples of features) and Y_DATA (output labels).

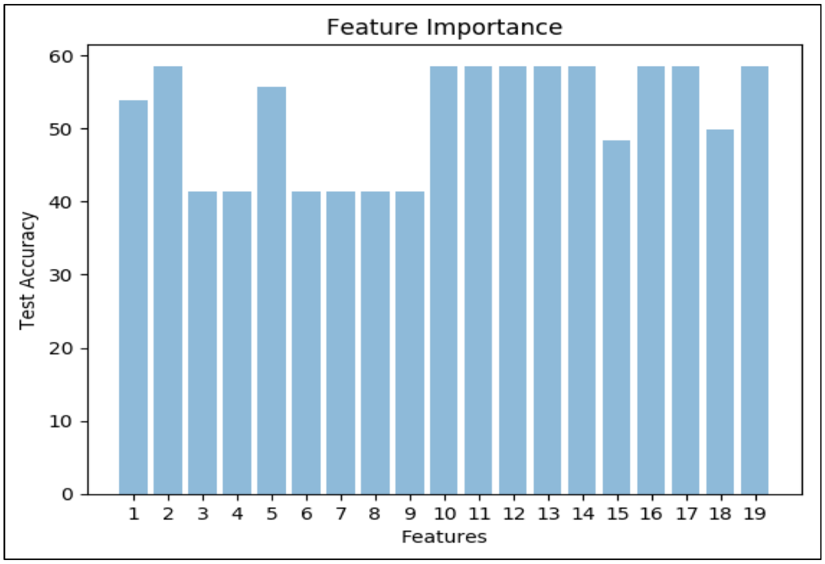


Fig. 3 Relative importance of selected features

Feature selection also provides following advantages:

- Dimensionality of data is reduced
- The training time is reduced
- As we remove outliers, variance decreases and hence overfitting does not take place.

3.3 Generating X_DATA and Y_DATA for training models:

By parsing the data, we extracted important features from data in the form of dictionary data structure for each new game and generated Game List and Team List for that entire year or season.

We have skipped the data for few initial games of each season and generated X_DATA and Y_DATA with relevant and appropriate features for that year. Match-related features like fumbles, goals made, rush attempts, etc are unknown until the match has been played. This means that only past average statistics for these features can be used to predict an current or upcoming match. Thus, we consider an average of these features for past matches for the individual teams.

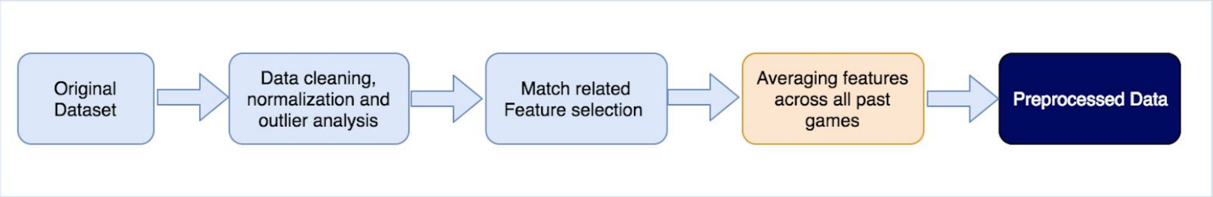

Original Dataset
Data cleaning, normalization and outlier analysis
Match related Feature selection
Averaging features across all past games
Preprocessed Data

Fig 4. Data Preprocessing workflow

Following features were computed/aggregated in X _DATA for each game in a season to be used for result computation.

X_DATA:

1. Win ratio of home team
2. Win ratio of visiting team
3. Point difference of home team
4. Point difference of visiting team
5. Time of possession of home team
6. Time of possession of visiting team
7. Home Team Penalty points
8. Visiting Team Penalty points
9. Kickoff yard Home Team
10. Kickoff yard Visiting Team
11. Home Team fumbles
12. Visiting Team fumbles
13. Rush yards per carry of Home Team
14. Rush yards per carry of Home Team
15. Yards per pass attempt of Home Team
16. Yards per pass attempt of Visiting Team
17. Goal Ratio of Home Team
18. Goal Ratio of Visiting Team
19. Scores difference so far Home Team - Visiting Team (Indicator variable)

1. **Win ratio:** Ratio of total number of matches won by a team and total matches played by that team.

2. **Point difference:**
   Difference in points for each game played by home team versus opponent team

3. **Time of possession:**
   Time of possession is computed from the first play initiated by one team from line of scrimmage till a score is made by or loss of possession takes place. This is an important feature in predicting intermediate values of outcome during a game.

4. **Penalty points:**
   Penalty is a sanction called against a team for a violation of the rules. Penalty points earned by each team in a match is one of the important factors to predict its outcome.

5. **Kick-off Yard:**
   Kickoff is a kick that puts the ball at start of each half in a game and player kicks the ball from center of the field. Kick-off yard distance measure is thus an important factor to be considered in estimating scoring probability of the team that kicks.

6. **Fumbles:**
   A fumble occurs when a player that has possession and control of ball loses it before being tackled or scoring. Number of fumbles in a match is thus a highly important feature as it explains number of time a team loses possession of a ball and hence affects goals made by that team.

7. **Rush yards per carry:**
   Rush yards per carry is ratio of Rush yards and Rush Attempts. This is value of yards per attempt rushing the ball is significant in estimating probability of goal and thus, overall match outcome

8. **Yards per pass attempt:**
   Deep passing plays during the course of a football game puts high emphasis on outcome of the match. This can be analysed by the ratio of pass yards per pass attempt for each team in a game.

9. **Goal Ratio:**
   Goal ratio feature is ratio of successful goals made by each team and goal attempts made by the team.

10. **Indicator Variable:**
    This field is used to taken into account strength of the teams. For any teams playing, the model is trained on their scores, the difference of which is later used as an input for learning as a completely new feature.

Y_DATA:

1. **0/1**
   1 if Home Team wins and 0 if Visiting Team wins

2. **Point difference (Victory Margin)**

Point Difference = Home Team points - Visiting Team points

This value positive if Home Team wins and negative if Visiting Team wins.

We generate the X_DATA with features mentioned above and divide each feature by number of matches played by the team +1. The plus one factor introduces random value to improve accuracy of prediction. Also, we assume that data for latest season is more importance than previous season data. To achieve this, we introduced a weight parameter alpha among features per season. For all features of X_DATA, we multiply each features' values till last season with alpha and add it for the current season. Alpha is a weight associated with last season's data.

We divided the complete data (X_DATA and Y_DATA) into 3 parts- training set (60%) , validation set(20%) and test set(20%). Hence 20% of our data is completely unknown for the model and can give effective Test accuracy values. As we have skipped data of first 100 games of each season, predictions are made on remaining matches in the season based on collective features of all previous games of that particular season.

Total Examples for 9 seasons data - 6320 (after featurization)
Examples for 7 seasons: 4834
Training examples - 3384
Validation examples - 1450

3.4 Data Visualization:

The distribution of the data when considering only win ratios of team (left in Fig) or point difference (right in Fig) looks like as shown. We can see two regions which can be separated linearly with some misclassification samples and correlation of these feature with the win of home and visiting team.

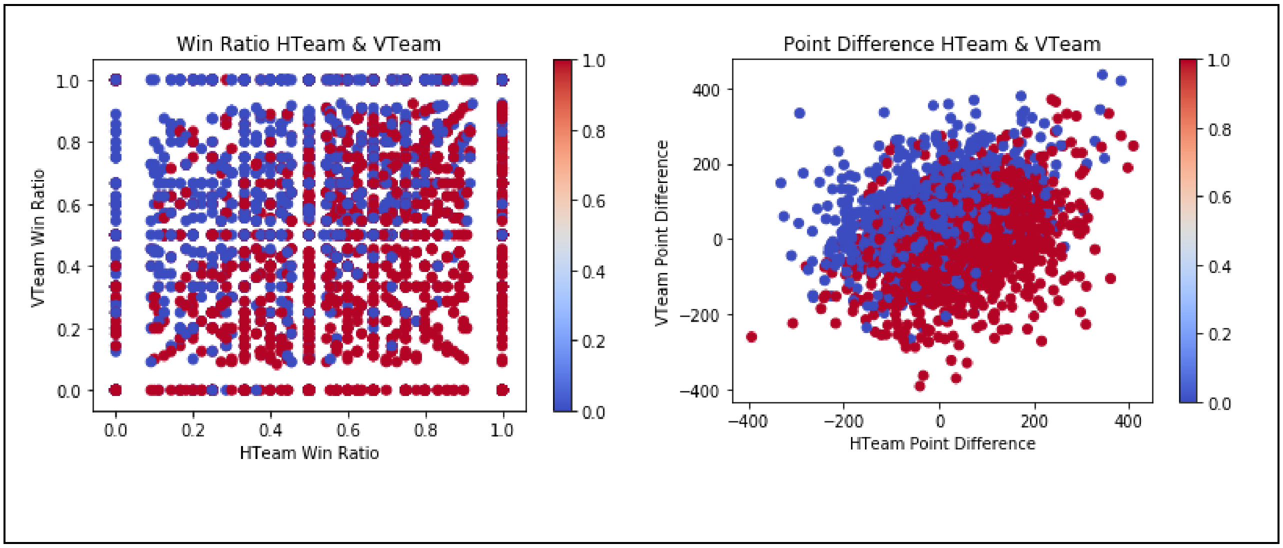


Fig 5. Data visualization of win ratio and point difference features for home team and visiting team

4. Modeling

We used following models in this project by tuning parameters like learning rates, number of epochs, etc.

4.1 Neural Network

Neural Networks are widely used to train non-linear models. They capture the essence of non-linearity in the data set by making use of hidden layers and different activation functions. We have tried with a small neural network which has two hidden layers with fifty nodes each. Activation function such as Relu and Sigmoid are used.

4.2 Support Vector Machine

SVM is one of the supervised training algorithms that builds a model from the training set and this model is then used to classify the test data to predict match outcome in this project. In SVM, we compared the accuracy obtained by using different types of kernels like RBF Kernel, Linear SVC and Polynomial kernel.

4.3 Logistic Regression

Logistic Regression tries to fit a regression model on the data by giving the output between [0,1]. The scores of prediction function from a linear regression are taken as input to a Sigmoid function. The final output can be used as probabilities of a win for the Home Team. Since this is a regression model, for the binary output labels, not normalising the features makes the model learn a constant function which always outputs either of wins or losses for the home team. Therefore, it was important to normalise the input features and bring all the features into a uniform range. We have used min-max normalisation for this purpose.

4.4 Gradient Boosting Regression

Gradient Boosting builds an additive model in a forward stagewise fashion which allows the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function. We used Gradient Boosting Regression to predict the victory margin i.e. point difference between home team and visiting team.

5. Design Diagram

The workflow of our project is as shown below:

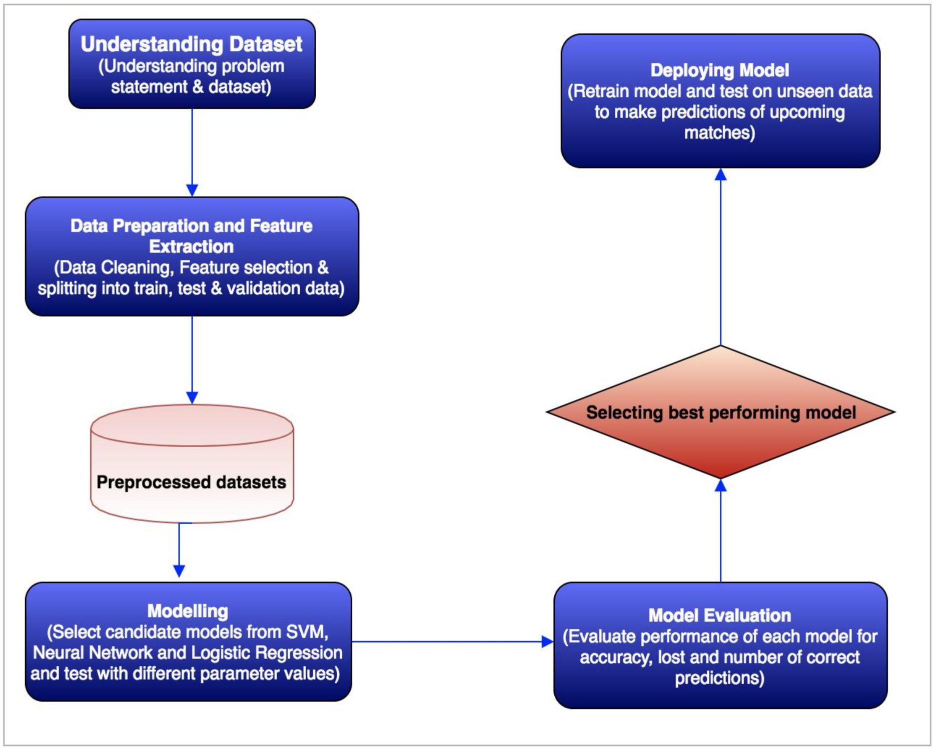

Understanding Dataset
(Understanding problem statement & dataset)
Data Preparation and Feature Extraction
(Data Cleaning, Feature selection & splitting into train, test & validation data)
Preprocessed datasets
Modelling
(Select candidate models from SVM, Neural Network and Logistic Regression and test with different parameter values)
Model Evaluation
(Evaluate performance of each model for accuracy, lost and number of correct predictions)
Selecting best performing model
Deploying Model
(Retrain model and test on unseen data to make predictions of upcoming matches)

Fig 6. Basic Design Diagram of project

6. Enhancing model with Feature addition

6.1 Baseline Algorithm - 1 (Number of Features = 2)

We used only one feature from the dataset for our first baseline algorithm and predicted the probability of the winning team with a better win ratio (defined as number of games won over total games played), among all games played before the current game in the same season. The probability is calculated as the ratio of win ratios for both the teams. In case of both teams with zero win ratio (a corner case), by definition both teams will have equal probability to win the current match since both of them have same win ratio.

6.2 Baseline Algorithm - 2 (Number of Features = 4)

We added few more features to our first baseline algorithm like the total point difference (difference of the number of points scored and the number of points conceded in all matches) for both teams, which is the home team and time of possession in the previous games, and finally test the model on these features to predict the winning probabilities for both teams in the current game.

6.3 Final Algorithm - 3 (Number of Features = 19)

In final algorithm, we have added many other features like yards per rush attempts, yards per pass attempts, ratio of goals made and goal attempts, kickoff yards, time of possession, number of fumbles of each of the teams playing the current game and indicator variable that has performance data of all teams. We run all three models on this data and predict outcome of match and evaluate the performance of all models. We use Gradient Boosting regression to predict victory margin or point difference of current game.

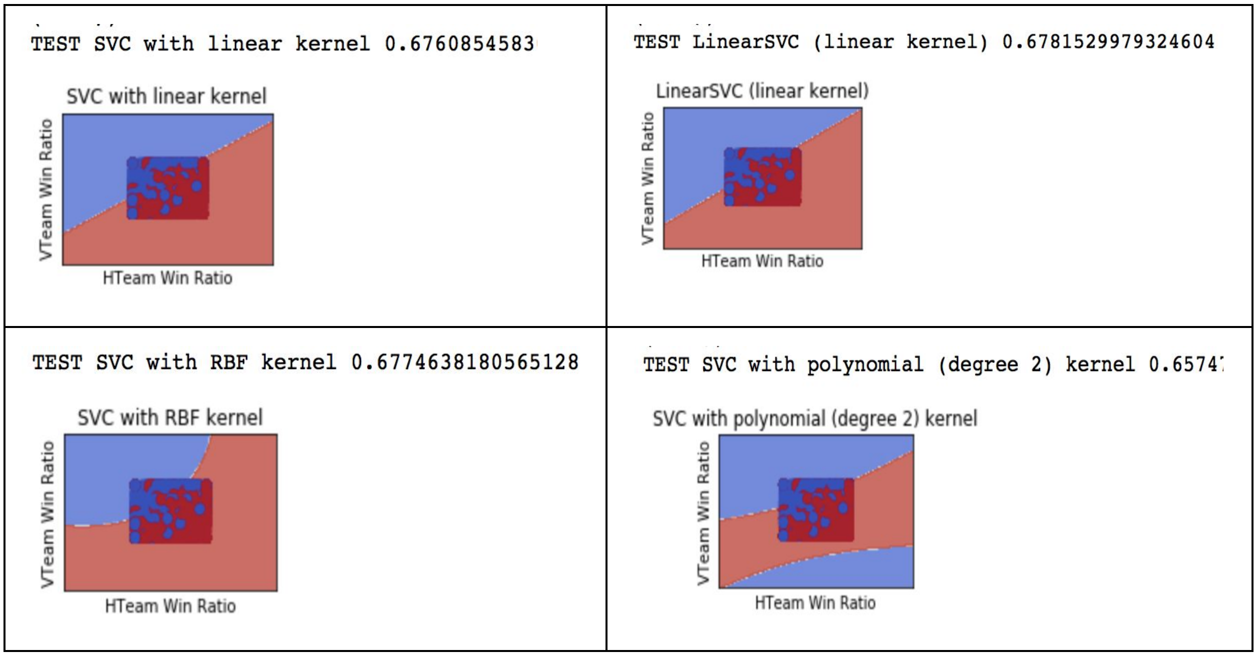
7. Results

7.1 Baseline Algorithm - 1

The results of this algorithm are as shown in below table:

Model	Loss on train	Loss on validation	Accuracy on train	Accuracy on validation
Logistic Regression	0.6768	0.6739	55.7788	56.3749
Neural Networks	0.5936	0.5852	68.40	69.56
SVM (Polynomial)	0.7307	0.7389	66.5009	65.7477

We also tried different kernels in SVM to predict accuracy on test data as shown in Figure 7:



```
TEST SVC with linear kernel 0.6760854583
```

```
TEST LinearSVC (linear kernel) 0.6781529979324604
```

```
TEST SVC with RBF kernel 0.6774638180565128
```

```
TEST SVC with polynomial (degree 2) kernel 0.6574
```

Fig 7. Result of Baseline Algorithm-1 using different kernels in SVM

7.2 Baseline Algorithm - 2

The results of this algorithm are as shown in below table:

Model	Loss on train	Loss on validation	Accuracy on train	Accuracy on validation
Logistic Regression	0.6560	0.6464	61.4838	61.1991
Neural Networks	0.5602	0.5528	70.93	71.09
SVM (Polynomial)	0.6916	0.7089	70.3517	69.7450

7.3 Final Algorithm - 3

We used the final model to make predictions on test data as well as validation data. The loss and accuracy observed are as shown in below table:

Model	Loss on train	Loss on validation	Loss on test (Unseen data)	Accuracy on train	Accuracy on validation	Accuracy on test (Unseen data)
Logistic Regression	0.5228	0.4932	0.4353	73.8064	74.3589	78.9367
Neural Networks	0.5212	0.4924	0.4343	73.8494	74.1935	79.2059
SVM (Polynomial)	0.4739	0.4908	0.4332	77.1586	74.9379	78.46567

It is observed that all three models give approximately same results and overall accuracy on validation as well as unseen data. This accuracy could be further improved by taking more features or creating new features by aggregating or manipulating the existing ones in order to reduce loss. We have used Logarithmic loss across all three models to compare their performance with each other.

7.4 Victory Margin Prediction using Gradient Boosting Regression:

In this model, we use Gradient Boosting Regression technique to analyze the statistical performance data of both teams playing current game and predict the average victory margin i.e. point difference between winning team and losing team of current match. This victory margin can be positive as well as negative. Since, victory margin is the difference of points between home team and visiting team, it will be positive if prediction says home team wins and negative if we predict that visiting team wins.

To calculate accuracy of predictions made, we select a threshold value (deviation from actual value) and compare the actual victory margin with the predicted one. If the absolute difference between actual and predicted value is less than or equal to the threshold, we consider it as a correct prediction and if this difference is larger than threshold, we consider it as incorrect prediction.

Based on our model, we get following results of accuracy on test and validation data for different values of threshold.

Deviation from actual value	Accuracy on Validation Data	Accuracy on Test Data
10	46.9331	45.1412
15	67.1950	62.9221
20	79.0489	76.9813

8. Performance Evaluation

We compare the performance of all three models (Neural Network, SVM and Logistic Regression) using Logarithmic loss. The performance is evaluated by majorly considering three factors:

1. Logarithmic loss
2. Accuracy
3. Number of Correct Predictions

We calculated these three metrics across train, validation and test data to observe that by proper data preprocessing and fine tuning the parameter values, an accuracy of approximately 74 % is achieved on validation data and around 78% is achieved on test data using all three models.

As seen from the Results (Section 7), increasing the number of features improved the accuracy especially for the regression model. Since, the regression model is linear and feature values were having a wide deviation, normalising them boosted the accuracy noticeably. The better performance on validation set and test set suggests that the model was not allowed to overfit the data during training. There was an increase in performance when a weight was given to previous seasons statistics for the teams.

9. Challenges

1. One of the significant challenge throughout the project has been data processing and cleaning. The data set used for the project is huge. We generated 19 features from the complete dataset which are considered to be relevant to our prediction model.
2. While parameter tuning, outliers posed a challenge which were not handled properly when we begin the project. Reducing the outliers showed significant improvement in performance.

10. Future Work

Future work comprises of extending the model to more enhanced and complex scenarios.

1. Since the training of models are not output label specific, they can be used to train on other outcomes. Other outcomes such as final scores of the game, various statistics like fumbles, yards, touchdowns, tackles etc. can be given as target inputs to the model to train on. They will use the same feature set to train, so feature removal will call for feature analysis again for training on a different outcome label.
2. Besides the features already used, the data set comprise of vast number of other features' information which can me included in the training after analysis to imitate more complex and real life scenarios.
3. Different teams keep changing their players each season which can affect its play. Player's positions such as Quarterback, Center and others can play an important role in deciding a team's winning chances.
4. Various other details for the plays in a game can be included into the our data. Not all features will be equally important. Hence, it will be required to carefully analyse and pick those that have significant impact on the performance of the model.

11. Acknowledgement

We would like to thank our supervisor, Prof. David Frohardt-lane, for the guidance, encouragement and advice he provided us during the course of the project. We are also grateful to Prof. David Rosenberg and Ben Jakubowski for their guidance throughout this course.

References

[1] Roary P. Bunker, Fadi Thabtah “A machine learning framework for sport result prediction.” (2017)

[2] Machine Learning for NFL Analysis: Prediction and Betting Evaluations https://becominghuman.ai/machine-learning-for-nfl-analysis-prediction-and-betting-evaluations-through-week-3-406027d2f7f

[3] College football explained (2012) https://www.theguardian.com/sport/blog/2012/oct/10/college-football-explained-ncaa

[4] Gradient Boosted Regression Trees https://blog.datarobot.com/gradient-boosted-regression-trees

[5] Christina Wadsworth, Francesca Vera, Michael Zhu “Predicting Point Spread in NFL Games”

[6] How I Used Machine Learning to Predict Soccer Games for 24 Months Straight https://doctorspin.me/digital-strategy/machine-learning/

[7] sklearn ensembleGradientBoostingRegressor http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

[8] Samuel Starkman, Emerson Boyd, Jeremy Eng, Kevin Kimelman, Xiangyu Li, David Kaeli “Machine Learning on Sports Prediction” http://staging-rise.s3.amazonaws.com/1362/3/1026/starkman_samuel_fxperw.pdf?AWSAccessKeyId=AKIAIZD5HUIXRXZ4FWDA&Expires=1775950506&Signature=xH4%2BIFjx9WRroezlDEhTHD2Ru5o%3D

[9] Support Vector Machines http://scikit-learn.org/stable/modules/svm.html

[10] Ben Ulmer, Matthew Fernandez “Predicting Soccer Match Results in the English Premier League”

[11] Josip Hucaljuk, Alen Rakipović“Predicting football scores using machine learning techniques“

[12] K. Sujatha, T.Godhavari, N.P.G. BhavaniI, B. Deepa Lakshmi “Football Match Statistics Predictions using Artificial Neural Networks”