PREDICTING NFL FOOTBALL GAME WINS

Jason Manuel, Put Dam, Nimish Garg, Cindy Li

Introduction

- Up until 2018, the Professional and Amateur Sports Protection Act (PASPA) made sports betting in the US illegal, with a few exceptions.[1]
- Despite its illegality, sports betting was still a common practice, with an estimated $380 billion being illegally gambled in 1999.[1]
- Since the overturn of PASPA, all but 12 states have introduced legislation related to sports betting.[2]

With sports betting being a very popular, we wanted to see whether we could predict the winners of NFL football games to aid in sports betting using machine learning methods.

- Current models include FiveThirtyEight's Elo Model and Microsoft's Cortana, which boasts accuracies of around 63% and 67% for the 2017 season, respectively.[3]

Methodology

- Game lineups were web-scraped from pro-football-reference.com.[4]
- Data was sourced from the SportRadar API[5], an official distributor of NFL statistics, using the lineups to get data per game.
  - 2012 – 2019 team game statistics were used
- Pandas and numpy libraries were used to clean data
  - Dropped columns containing null values
  - Dropped columns highly correlated with output
  - Game results were extracted from points columns
  - Sample: 1912 games with 232 features per game
- Train-validation-test split:
  - Train: 2012-2018 seasons game data
  - Validation: half of the 2019 season game data
  - Test: the remaining half of the 2019 season data
  - Additional Testing Set:
    - For each game, we took the average of the statistics of the away team and the home team, respectively, from their past 5 games to use as team game statistics
    - This is to see whether our model can indeed predict the outcome of future games
- Models: used scikit-learn and tensorflow
  - Logistic Regression
  - SVM
  - Feed forward neural network
- Feature Selection
  - used p-values as well as coefficients of logistic regression model to determine features to drop
  - used sklearn recursive feature elimination

Results and Analysis

Model	Accuracy	Accuracy (w/ feature selection)
Logistic Regression	63%	63%
SVM	52%	-
Feed-Forward	60%	63%
FiveThirtyEight's Elo	63%	-
Microsoft's Cortana	67%	-
Baseline: no training	~50%	-

Figure 1: Chart of the model and accuracy. Lighter blue indicates accuracy on our additional testing set. Darker blue indicates accuracy on the first 15 weeks of the 2017 NFL season.

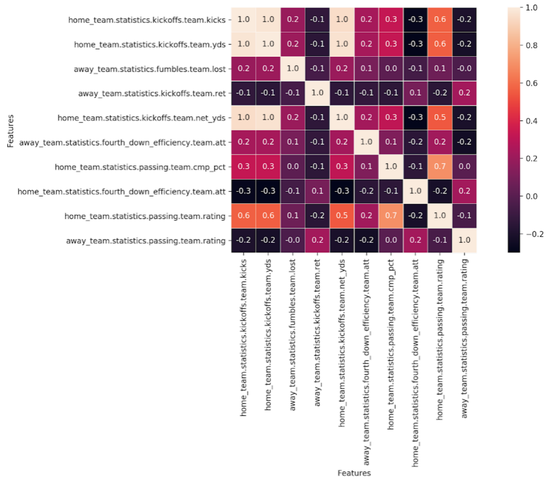
Logistic regression performed the best of our models with an accuracy of 63%, on par with the Elo model, but Cortana still performed the best with an accuracy of 67%.

Feature selection through recursive feature elimination improved the feed-forward network somewhat, but did not improve the logistic regression model.

Figure 2: The top 10 most important features based on the absolute value of their coefficients from the logistic model and their correlations.

We can see that some of these features are highly correlated with each other, in which case, we want to consider dropping them from our model.

In the end though, the model using only these features and the model only using features with p-values < 0.05 both performed worse than the model with all of the features.



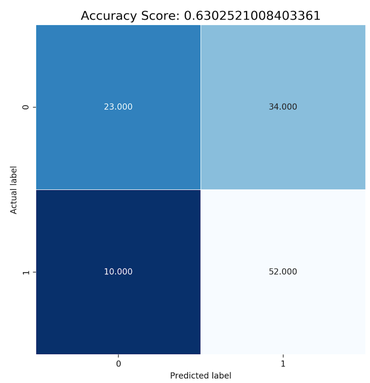


Figure 3: Confusion Matrix of logistic regression model predictions.

When true label is 1, our model has an accuracy of around 80%, but when true label is 0, the accuracy is around 40%.

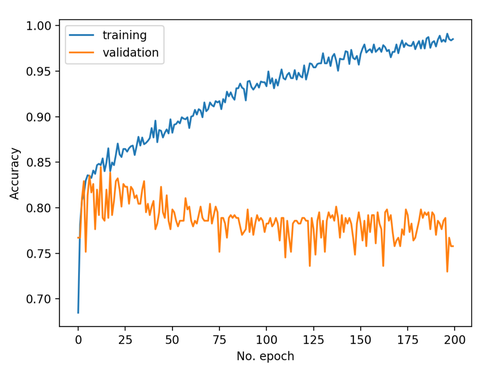


Figure 4: Feed-forward network training and validation accuracy over number of epochs.

As the number of epochs increases, training accuracy continues to increase while validation trends slightly downwards, suggesting overfitting.

Limitations & Future Works

Limitations

- Lack of data points we had, especially compared to our feature space
- Certain features not included (i.e. weather, stadium, etc.)
- Data source inconsistent in number of games per season and lack of access to other reputable sources

Future Work

- Use an autoencoder to reduce input dimensionality.
- Gather more data to see if this improves the models.
- Incorporate player data as features, as team performance can change based on who is playing.
- Learn how certain teams perform home vs away to see if this is why our model is biased towards home winning.

Conclusions

- Our models achieved a peak accuracy of 63%, with feature selection not improving the logistic regression model and slightly improving the feed-forward network.
- Our logistic regression and feed-forward models achieve an accuracy on par or relatively close to FiveThirtyEight's Elo and Microsoft's Cortana models.
- Currently, our model would not be very useful for sports betting. Some sports experts can achieve accuracies around the same if not slightly higher than our model.[6]
- A good model, however, can have some serious ethical implications.
  - If our model is very good, it may encourage more people to gamble and gamble more, which is not a pattern of behavior we wish to encourage

Bibliography

1. Meer, Eric. "The Professional and Amateur Sports Protection Act (PASPA): a Bad Bet for the States." *UNLV Gaming Law Journal,* vol. 2, 2011.
2. Moran, Patrick. "Anyone's Game: Sports-Betting Regulations after *Murphy v. NCAA*." *CATO Institute Legal Policy Bulletin*, 2019.
3. Gaines, Cork. "One of the Computer Models Picking NFL Games Has Taken a Big Lead - Here Is Who They like in Week 16." Business Insider, Business Insider, 21 Dec. 2017, www.businessinsider.com/nfl-picks-microsoft-cortana-elo-week-16-2017-12.
4. "Pro Football Statistics and History." Pro, Sports Reference LLC, www.pro-football-reference.com/.
5. "Sportradar: Sports Data and Content Solutions, Made Easy." Sportradar US, sportradar.us/.
6. "NFL Sports Betting Accuracy." *BettingPros*, 2020, www.bettingpros.com/nfl/accuracy/game/.