PREDICTIVE MODELING & GAME SIMULATION

SUPERBOWLLIII

...

KAITLYN DRAKE
JAVIER ORRACA

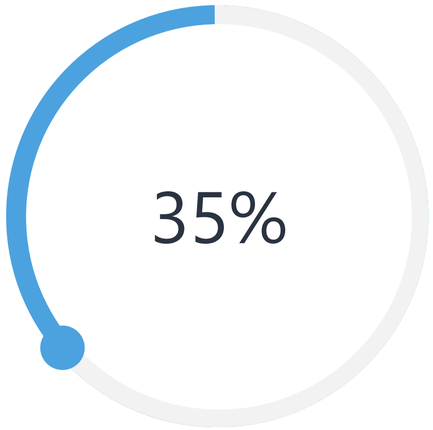
ABOUT US

Kaitlyn Drake and **Javier Orraca** are graduate students in UCI's Master of Science in Business Analytics ("**MSBA**") program.

David Savlowitz and **Michael Ponton** teach an MSBA course, *Applied Predictive Modeling*, for graduate students in the MSBA program. Predictive modeling techniques are taught through advanced software. David Savlowitz is the Founder & CEO of **Competitive Analytics** and Michael Ponton is the firm's Director of Analytics.

TOOLS USED IN THIS ANALYSIS:
R, WEKA, JupyterLab, dplyr(R), SQLDF(R), ggplot2(R), Plotly(R), gganimate(R)

PREDICTIVE MODELING PROCESS



1: DATA COLLECTION

Data sets were sourced from ESPN, Fortune, ProFootball, and other online sources.

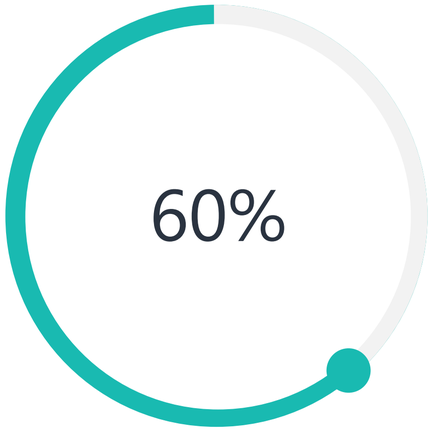
2: DATA MANIPULATION

Data was sanitized, manipulated, and reviewed for completeness in R.
Data frames were created as needed throughout R program.

3: VISUALIZATIONS

Data exploration through visualizations supports the analysis and modeling process.

PREDICTIVE MODELING PROCESS



4: POISSON REGRESSION

$$\Pr(Y_i = y_i \mid \mu_i, t_i) = \frac{e^{-\mu_i t_i}(\mu_i t_i)^{y_i}}{y_i!}$$

where

$$\mu_i = t_i \mu(\mathbf{x}_i'\boldsymbol{\beta})$$
$$= t_i \exp(\beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})$$

5: SIMULATION MODELING

Relied on Poisson regression and J48 decision tree models, and further created a simulation function in R to predict scores & probabilities of those scores.

6: CONCLUSIONS

The Patriots are better scorers than the average NFL team, but *not as good of scorers* as the Los Angeles Rams.

DATA COLLECTION & MANIPULATION

PRIMARY DATA SET

The main data set consisted of the 2018 season NFL data including Team, Opponent, 15 variables of interest (score, rushing vs passing yards, turnovers, etc.).

MANIPULATION

Data points were imputed from the primary data set, growing the number of variables from 15 to 30, including net metrics, home vs away metrics, etc.

VISUALIZATIONS

The data was initially viewed in table form, then transformed to scatter plots and charts, and interactive visualizations were developed to better understand the NFL.

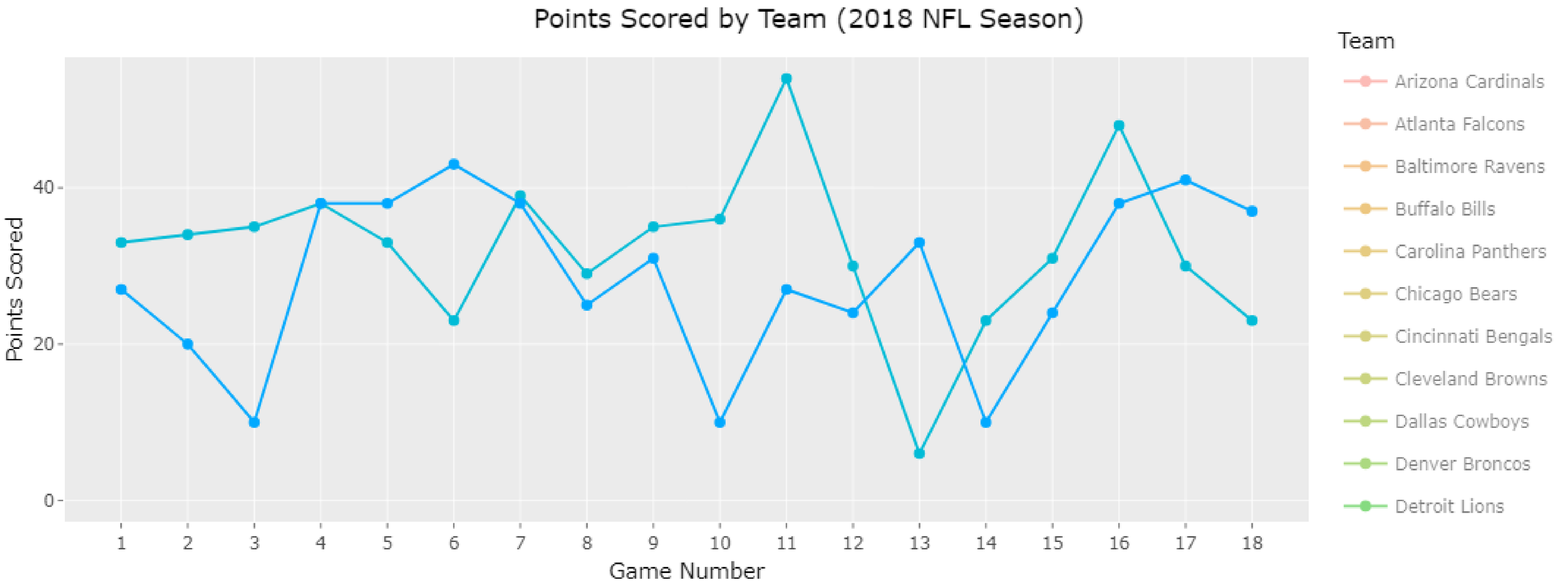
SAMPLE R CODE

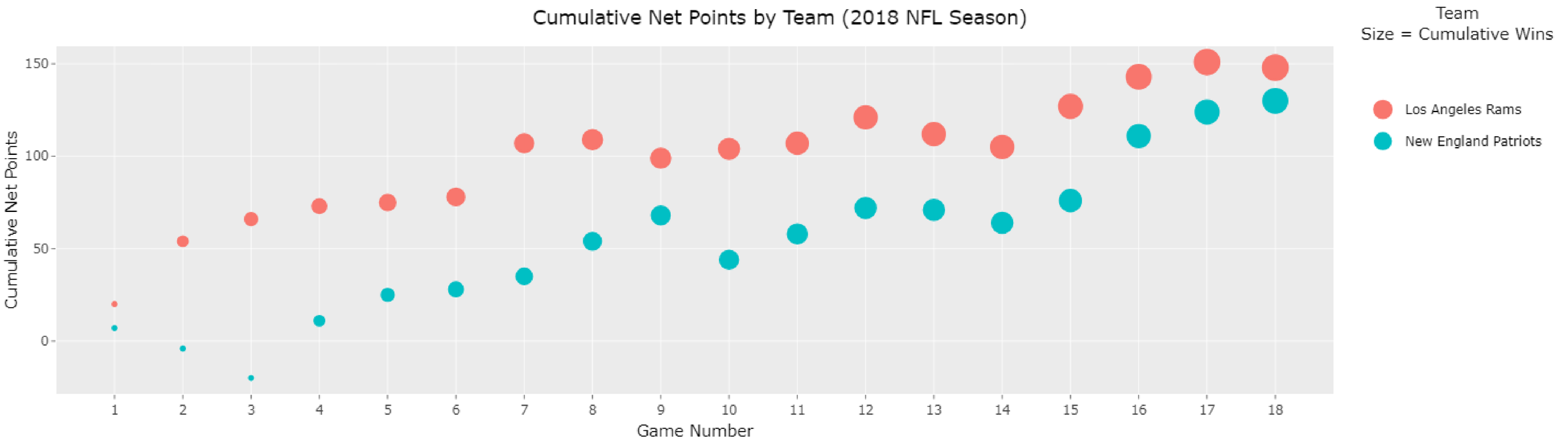
```
# Plot 1: Time-series interactive plot, by Team
NFL_TimeSeriesLine <- ggplot(NFL_Trim, aes(GameNumber, Team_Score, group=Team, colour=Team)) +
                      geom_line() + geom_point() + ylab("Points Scored") + xlab("Game Number") +
                      scale_x_continuous(breaks=seq(1,18,1)) +
                           ggtitle("Points Scored by Team (2018 NFL Season)")
ggplotly(NFL_TimeSeriesLine)


# 4.2: Create new data frame and run Poisson regression
NFL_Poisson <- rbind(
    data.frame(Points=NFL_Trim$HomeGoals,
               Team=NFL_Trim$Team,
               Opponent=NFL_Trim$Opponent,
               Home=1),
    data.frame(Points=NFL_Trim$AwayGoals,
               Team=NFL_Trim$Opponent,
               Opponent=NFL_Trim$Team,
               Home=0)) %>%
glm(Points ~ Home + Team + Opponent, family=poisson(link=log), data=.)
```

VISUALIZATIONS

Plots created in **R** with **ggplot2**.
Made interactive with **Plotly** and **gganimate**.

- **R** is a programming language and **free, open-source software** for statistical computing and visualizations.
- **ggplot2**, part of the *Tidyverse*, is an open-source graphical system and R-package for creating data visualizations.
- **Plotly** and **gganimate** are animation packages that wrap around the R visuals to create interactive, web-based maps and plots.


Points Scored by Team (2018 NFL Season)
Points Scored
0
20
40
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Game Number
Team
Arizona Cardinals
Atlanta Falcons
Baltimore Ravens
Buffalo Bills
Carolina Panthers
Chicago Bears
Cincinnati Bengals
Cleveland Browns
Dallas Cowboys
Denver Broncos
Detroit Lions


Cumulative Net Points by Team (2018 NFL Season)
Cumulative Net Points
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Game Number
Team
Size = Cumulative Wins
Los Angeles Rams
New England Patriots

MODELS SELECTED

A **Poisson regression** is a form of generalized linear model used for analyzing multivariate problems, deriving data-driven insights, and building predictive models. The high-mean NFL scoring appears normally distributed, but variable significance and model performance was stronger with Poisson vs Linear regression.

The **C4.5 algorithm** is used for statistical classification problems to generate decision trees. We used **J48**, an open-source Java implementation of C4.5, to maximize information game at every tree node split.

VARIABLES

All combinations of NFL home-team advantages, teams, and opponents were considered in our Poisson regression.

POISSON MODEL

Our formula (Points ~ Home + Team + Opponent) Iterated six times to predict game scores and game score probabilities.

DECISION TREES

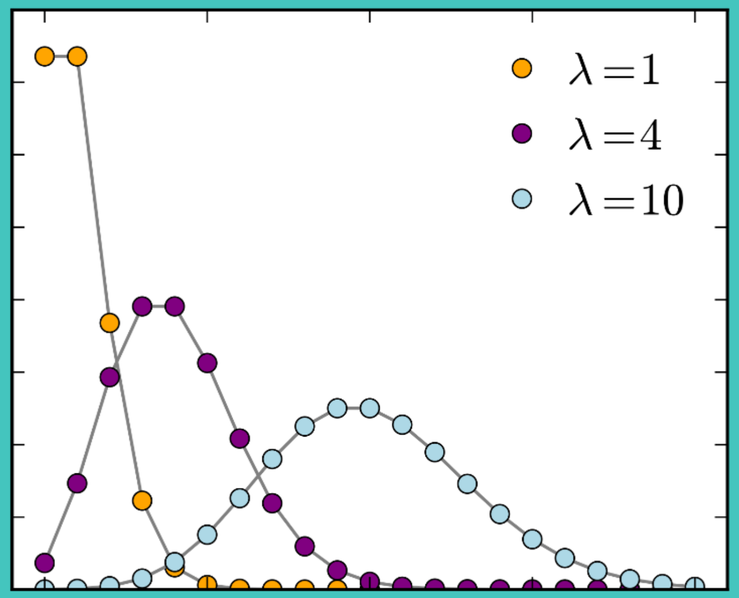
The J48 decision tree algorithm was utilized, via WEKA, to predict the Super Bowl champion, and the probability of that event.

SIMULATION

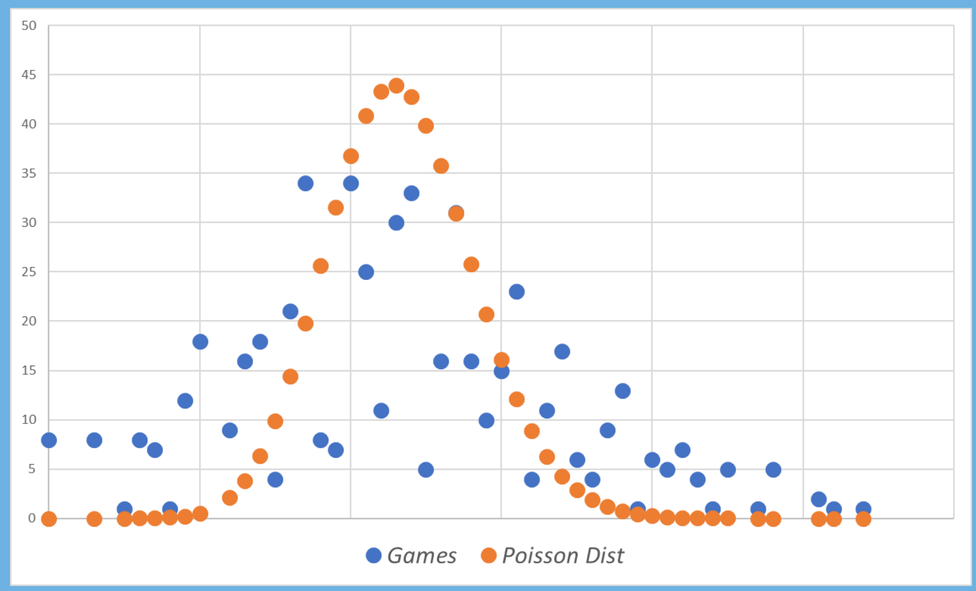
The simulation function allowed for team score predictions & probabilities of all possible score combinations.

POISSON DISTRIBUTION

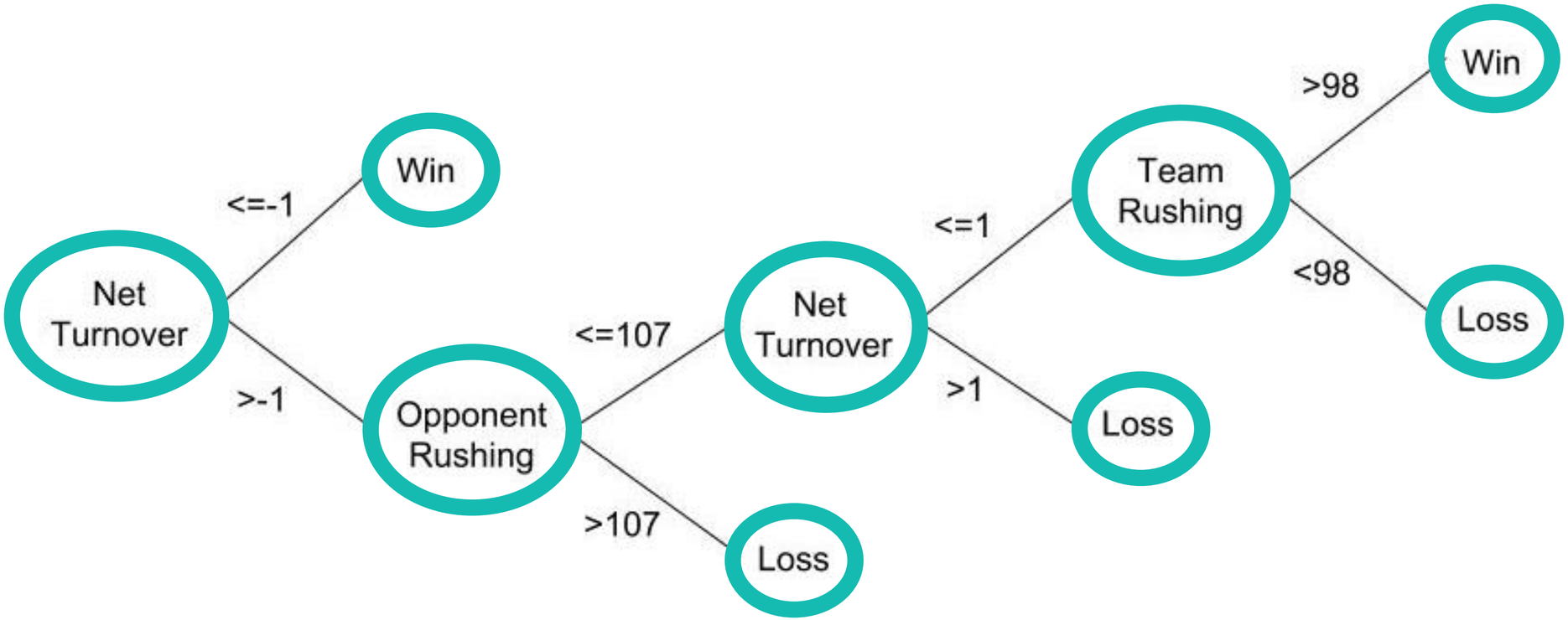
POISSON PROBABILITY MASS FUNCTION

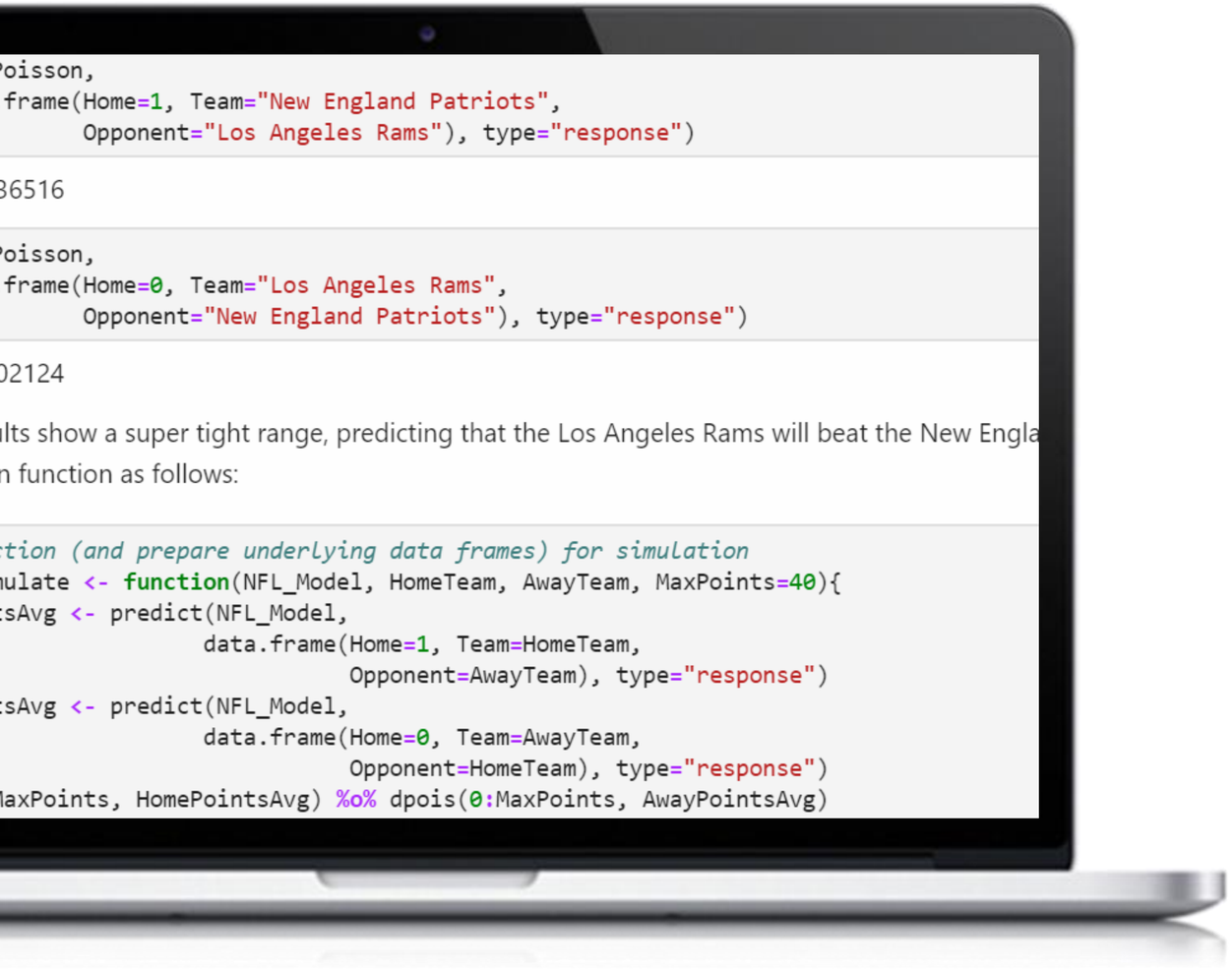


2018 NFL SCORE DISTRIBUTION



J48 DECISION TREE VIA WEKA





GAME SIMULATION

HOME TEAM ADVANTAGE

While not significant to the overall conclusions, the Patriots were assigned home-team advantage given expected crowd size at Super Bowl 53.

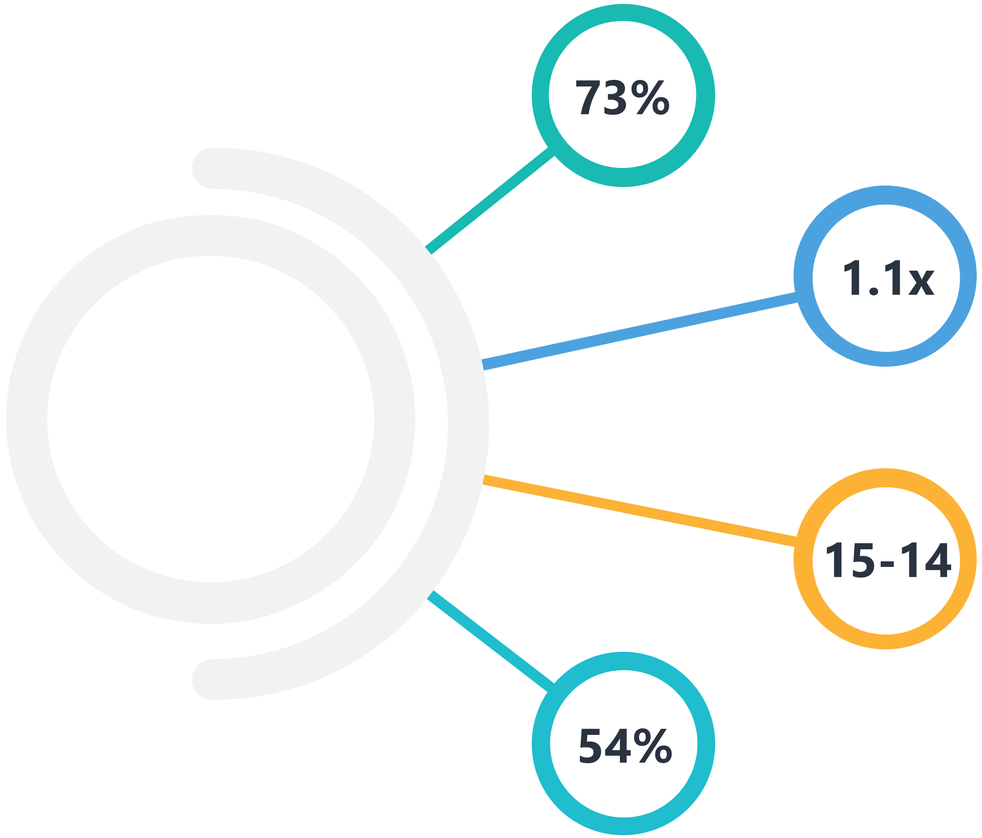
GAME SIMULATION

The Rams vs Patriots were passed through the simulation function and Poisson regression model to develop a matrix of final scores between the teams.

LIKELIHOOD OF PREDICTED SCORES

Matrix calculations allowed us to better understand the odds of the Rams winning vs losing.

MODELING INSIGHTS



RAMS WIN!

The J48 decision tree model reinforced individual score predictions.

CLOSE GAME

Our model predicts a very close game, with Rams winning 15-14.

HOME ADVANTAGE

Exponentiating the Home coefficient from the regression model, the Patriots are expected to have a 1.1x scoring advantage

SCORE PROBABILITY

Matrix calculations of all potential Super Bowl scores indicates a 54% chance of the 15-14 predicted score.

THANK*YOU*

...

CONTACT US

JAVIER ORRACA

Email: jorraca@uci.edu

LinkedIn: https://www.linkedin.com/in/Orraca/

GitHub: https://javorraca.github.io/Home/

KAITLYN DRAKE

Email: kdrake1@uci.edu

LinkedIn: https://www.linkedin.com/in/kaitdrake/

University of California, Irvine

The Paul Merage School of Business

4293 Pereira Dr, Irvine, CA 92697