Universidade Federal do Rio Grande do Norte
Centro de Ciências Exatas e da Terra
Departamento de Estatística

Web Scraping, Análise e Predição de Resultados de times mandantes na NFL utilizando um modelo de Random Forest

Lucas Medeiros dos Santos

Relatório sobre Web Scraping, Análise e predição de resultados de times mandantes na NFL, da disciplina Big Data ministrada pela professora Marcus Nunes.

Natal

6 de agosto de 2024

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Na primeira parte, explicamos como usamos o Web Scraping para obter informações das partidas de futebol americano de várias temporadas, de 2013 a 2023. Isso envolve extrair dados de um mesmo site online porém de diferentes URLS, uma para cada ano, para ter uma base sólida para nossa análise.

Na segunda parte, analisamos os dados coletados, organizando-os e removendo qualquer confusão. Essa etapa é essencial para entender melhor as partidas e preparar os dados para a próxima fase, onde construímos nosso modelo de previsão.

Na terceira parte, entramos na construção do modelo de previsão, utilizando Random Forest. Ajustamos o modelo para obter os melhores resultados, mexendo em alguns ajustes para que ele funcione da melhor maneira possível.

Na quarta parte, avaliamos o desempenho do nosso modelo. Usamos termos como AUC-ROC, acurácia e outras métricas para entender quão bem nosso modelo pode prever os resultados dos mandantes dos jogos.

Finalizamos o relatório com algumas conclusões sobre o que aprendemos e sugerimos caminhos para pesquisas futuras. Este relatório é como um guia, usando a tecnologia para entender e prever o que pode acontecer nos jogos da NFL.

2 Metodologia

A metodologia adotada para explorar e analisar os dados de partidas da NFL será dividida em duas etapas principais: Web Scraping e Modelagem com Random Forest.

2.1 Web Scraping

Na primeira etapa, empregamos a técnica de Web Scraping para coletar dados cruciais de partidas da NFL. Isso envolveu a seleção de fontes confiáveis, extração e limpeza de dados brutos, e transformação para um formato mais adequado para nossa análise. As etapas específicas incluíram:

2.1.1 Seleção e Extração de dados

Utilizamos um conjunto de URLs específicas retiradas do site: https://www.pro-football-reference.com para coletar informações detalhadas sobre partidas da NFL.

Removemos linhas duplicadas, incluindo aquelas que correspondiam aos títulos, para evitar repetições ao combinar os dados.

2.1.2 Limpeza e Transformação de Dados

Excluímos colunas desnecessárias, como Dia, Data, e Hora, simplificando assim o conjunto de dados.

Transformamos as variáveis relevantes em formato numérico para facilitar análises futuras.

Adicionamos a coluna Year para identificar o ano de cada partida.

2.1.3 Agregação e Filtragem de Dados

Consolidamos os dados de todas as temporadas em um único dataframe.

Excluímos as linhas correspondentes às semanas 13 até a última semana de jogos, incluindo playoffs, da temporada atual, para focar nas partidas regulares.

	Week	Winner/tie	...3	Loser/tie	...5	PtsW	PtsL	YdsW	TOW	YdsL	TOL	Year
1	1	Detroit Lions	@	Kansas City Chiefs	boxscore	21	20	368	1	316	1	2023
2	1	Atlanta Falcons		Carolina Panthers	boxscore	24	10	221	0	281	3	2023
3	1	Cleveland Browns		Cincinnati Bengals	boxscore	24	3	350	2	142	0	2023
4	1	Jacksonville Jaguars	@	Indianapolis Colts	boxscore	31	21	342	2	280	3	2023
5	1	Washington Commanders		Arizona Cardinals	boxscore	20	16	248	3	210	2	2023
6	1	Baltimore Ravens		Houston Texans	boxscore	25	9	265	2	268	1	2023
7	1	Tampa Bay Buccaneers	@	Minnesota Vikings	boxscore	20	17	242	0	369	3	2023
8	1	New Orleans Saints		Tennessee Titans	boxscore	16	15	351	2	285	3	2023
9	1	San Francisco 49ers	@	Pittsburgh Steelers	boxscore	30	7	391	1	239	2	2023
10	1	Green Bay Packers	@	Chicago Bears	boxscore	38	20	329	0	311	2	2023
11	1	Las Vegas Raiders	@	Denver Broncos	boxscore	17	16	261	1	260	0	2023
12	1	Miami Dolphins	@	Los Angeles Chargers	boxscore	36	34	536	2	433	0	2023
13	1	Philadelphia Eagles	@	New England Patriots	boxscore	25	20	251	1	382	2	2023
14	1	Los Angeles Rams	@	Seattle Seahawks	boxscore	30	13	426	0	180	0	2023
15	1	Dallas Cowboys	@	New York Giants	boxscore	40	0	265	0	171	3	2023
16	1	New York Jets		Buffalo Bills	boxscore	22	16	289	1	314	4	2023
17	2	Atlanta Falcons		Green Bay Packers	boxscore	25	24	446	1	224	0	2023
18	2	Buffalo Bills		Las Vegas Raiders	boxscore	38	10	450	0	240	3	2023

Figura 1: Dados de partidas da NFL

2.1.4 Criação de Variáveis Adicionais

Introduzimos quatro novas variáveis: Jardas do Time da Casa, Jardas do Time Visitante, Turnovers do Time da Casa e Turnovers do Time Visitante.

Utilizamos lógica condicional com a função ifelse para atribuir valores com base na presença do símbolo @ em determinadas colunas, indicando se o time da casa ganhou ou perdeu.

2.1.5 Classificação e Identificação

Criamos uma nova coluna indicando se o time da casa ganhou (1) ou perdeu (0), com base na lógica condicional aplicada anteriormente.

Introduzimos a variável Result at Home, categorizando os resultados em Win, Lost/Tie com base nas condições estabelecidas.

2.1.6 Validação para Playoffs

Adicionamos uma coluna identificando se a partida é válida para os playoffs (1) ou não (0), com base nas semanas correspondentes a essa fase.

2.2 Modelagem com Random Forest

2.2.1 Seleção de Variáveis de Interesse

Escolhemos as variáveis relevantes para a análise subsequente, incluindo Resultado do mandante, Jardas do Time da Casa, Jardas do Time Visitante, Turnovers do Time da Casa, Turnovers do Time Visitante e Playoffs, com isso teremos o seguinte conjunto de dados:

result_at_home	yds_at_home	yds_away	to_at_home	to_away	playoffs
Lost/Tie	316	368	1	1	0
Win	221	281	0	3	0
Win	350	142	2	0	0
Lost/Tie	280	342	3	2	0
Win	248	210	3	2	0
Win	265	268	2	1	0
Lost/Tie	369	242	3	0	0
Win	351	285	2	3	0
Lost/Tie	239	391	2	1	0
Lost/Tie	311	329	2	1	0
Lost/Tie	260	261	0	1	0
Lost/Tie	433	536	0	2	0
Lost/Tie	382	251	2	1	0
Lost/Tie	180	426	0	0	0
Lost/Tie	171	265	3	0	0
Win	289	314	1	4	0
Win	446	224	1	0	0
Win	450	240	0	3	0

Figura 2: Dados do resultados de partidas dos mandantes na NFL

2.2.2 Pré-processamento de Dados

Dividimos o conjunto de dados em conjuntos de treino e teste.

Normalizamos e escalonamos os dados.

2.2.3 Exploração e Tuning do Modelo

Exploramos diferentes parâmetros do modelo Random Forest para encontrar a combinação ideal.

Utilizamos técnicas de validação cruzada para garantir a generalização do modelo.

2.2.4 Avaliação do Modelo

Utilizamos métricas como AUC-ROC, acurácia, matriz de confusão, sensibilidade e especificidade para avaliar o desempenho do modelo na previsão de resultados.

3 Resultados

3.1 Análise

Nesta seção iremos mostrar o quantitativo de vitórias que o time mandante conquistou nesse intervalo de tempo e quantos derrotas obteve:

Tabela 1: Resultados dos mandantes

Resultado do mandante	Partidas
Ganhou:	1.585
Perdeu ou Empatou:	1.274

3.1.1 Componente Principal

Vamos mostrar o gráfico de componentes principais dos nossos dados para visualizarmos como estão distribuidos meus resultados:

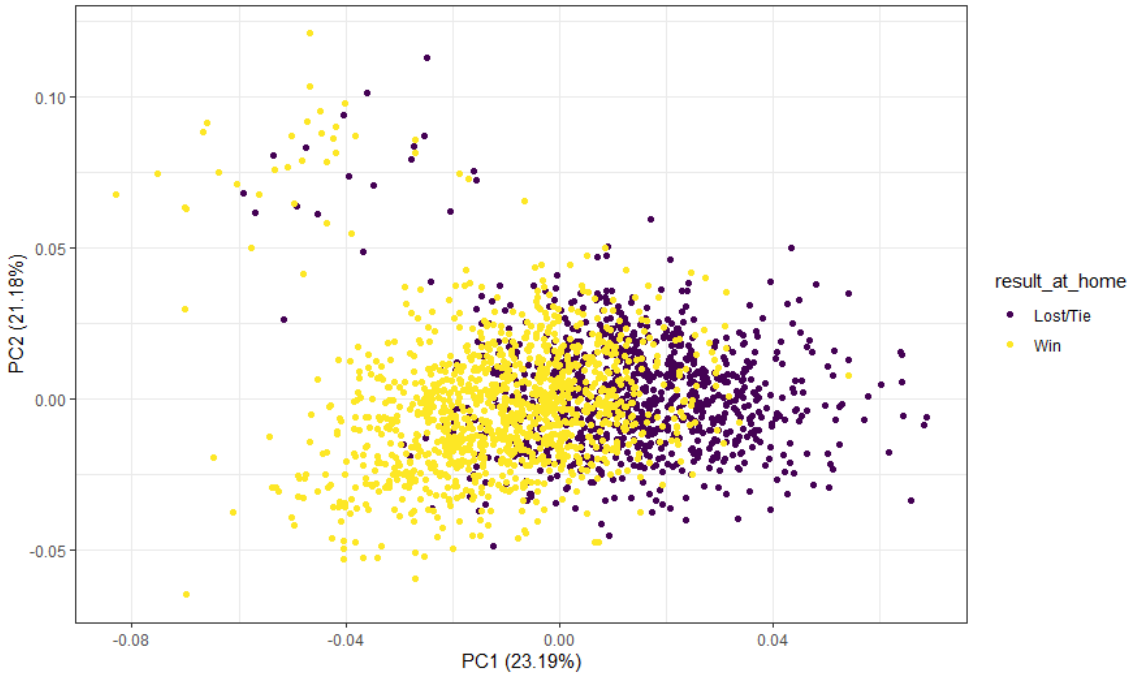

Figura 3: Gráfico da Componente principal 1 e 2

3.1.2 Matriz de dispersão

A matriz de dispersão é útil para visualizar padrões de associação ou correlação entre variáveis. Cada célula da matriz mostra um gráfico de dispersão entre duas variáveis, enquanto a diagonal principal pode exibir histogramas ou densidades univariadas para cada variável.

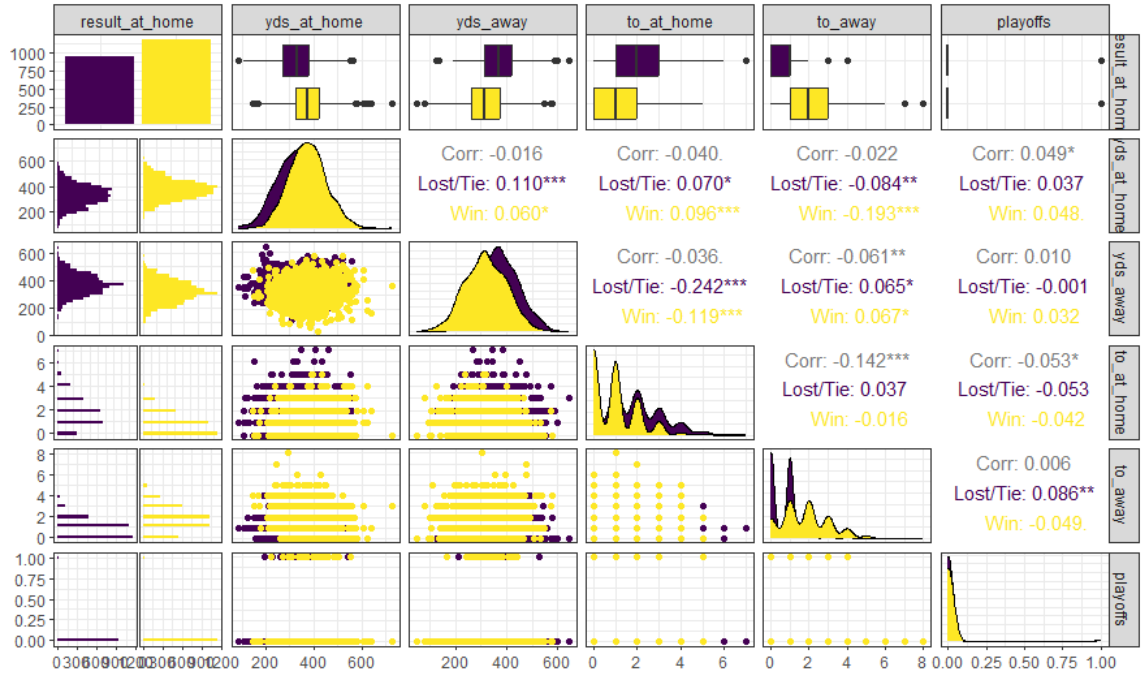

Figura 4: Matriz de dispersão

3.2 Melhor modelo

Para encontramos o melhor modelo iremos utilzar um validação com 10 iterações, juntamente com um 1.000 árvores, com um mtry de 1 a 5, o min_n de 10 a 50 e utilizaremos o tunnig para o meu grid de procura para realizar uma busca sistemática para encontrar a combinação de valores dos hiperparâmetros (mtry e min_n) que otimizem métrica de desempenho, que no nosso caso será a acurácia do modelo e a área sob a curva.

Com isso definido teremos então o gráfico com as métricas de acurácia e área sob a curva: Portanto teremos o melhor modelo, em que se teve uma acurácia de 0.796 no conjunto de treino, foi treinado com 1000 árvores, usando uma amostra de 1910 observações e 5 variáveis independentes. Com uma variável considerada para divisão em cada nó, 20 tamanho mínimo permitido para um nó folha e configurado para prever probabilidades, com uma abordagem de importância de variáveis baseada na impureza de Gini. O erro de previsão fora da amostra (Brier Score) é indicado como aproximadamente 0.1558439.

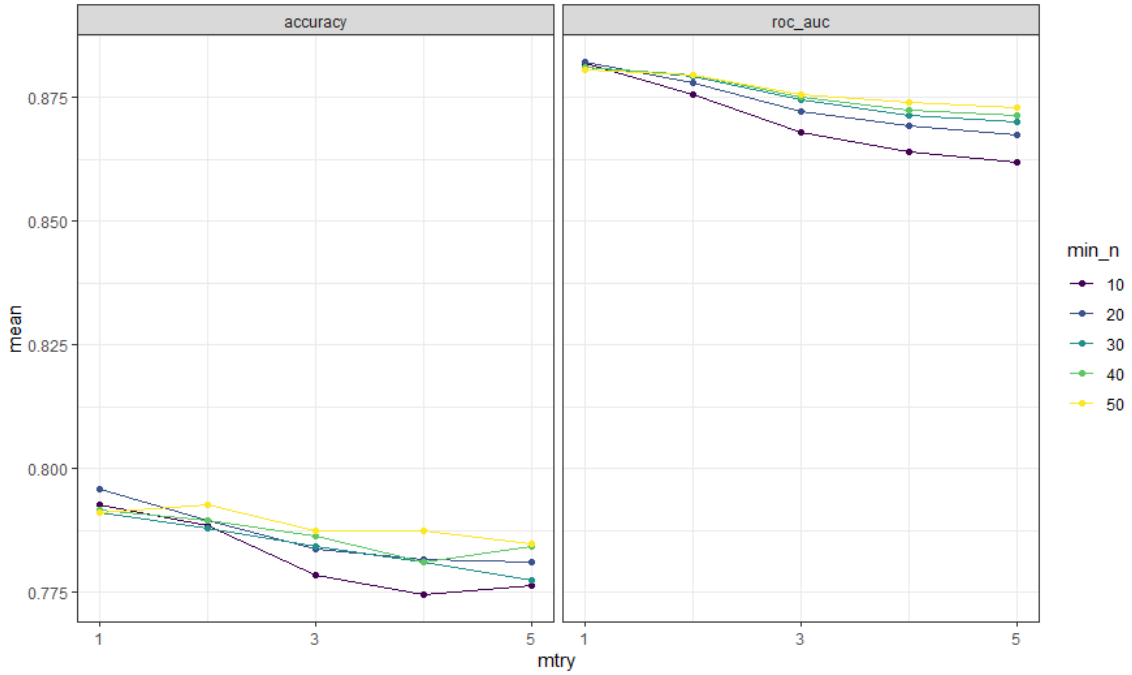

Figura 5: Acurácia e Área sob a curva

3.3 Variáveis mais importantes

Teremos então que a importância das variáveis são as seguintes:

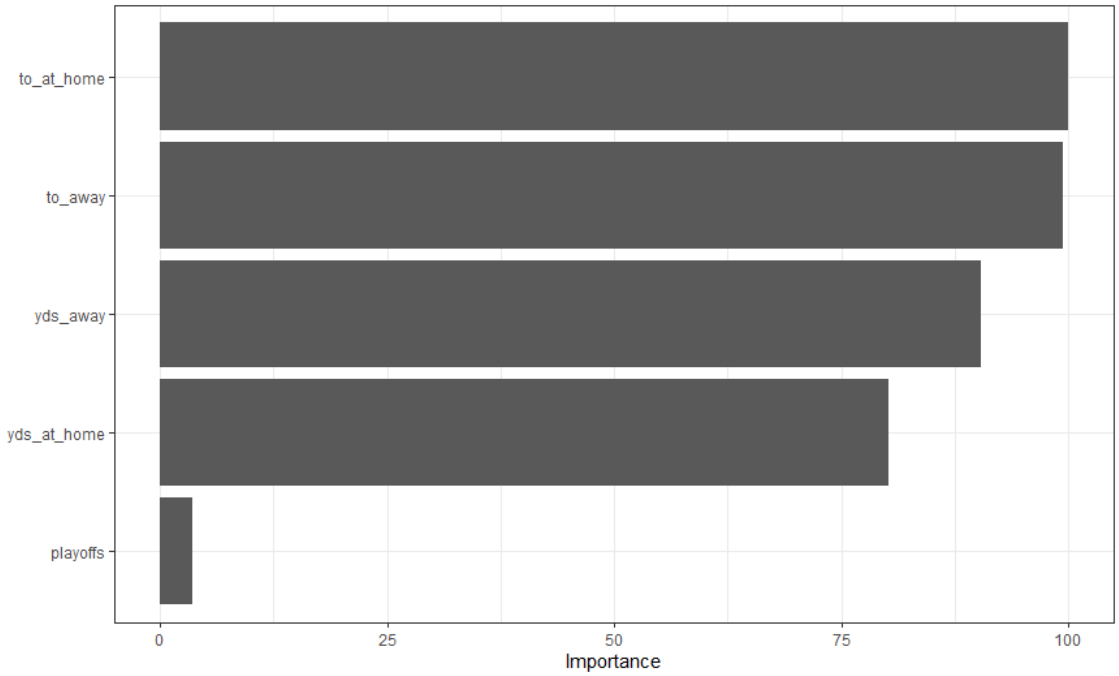

Figura 6: Importância das variáveis

Portanto as variáveis que mais importam para sabermos o resultado do time mandante são respectivamente os Turnovers que o time da casa cometeu, Turnovers que o time visitante cometeu, jardas conquistadas do time vistante, jardas conquistadas do time mandante e

o se é jogo de playoffs.

3.4 Resultados no conjunto de teste

Ao aplicarmos o modelo de treino no conjunto de teste temos então que a acurácia aumentou comparada ao conjunto de treino, como demostrado na tabela abaixo:

Tabela 2: Acurácia nos conjuntos de treino e teste

Métrica	Estimada
Acurácia (treino)	0.796
Acurácia (teste)	0.810

Obtemos também um kappa de 0.619 sugere que o está com uma concordância considerável em relação com as observações reais.

3.5 Matriz de confusão

Para visualizarmos os resultados que meu modelo previu no conjunto de teste utilizaremos a matriz de confusão para verificarmos, como segue a figura abaixo:

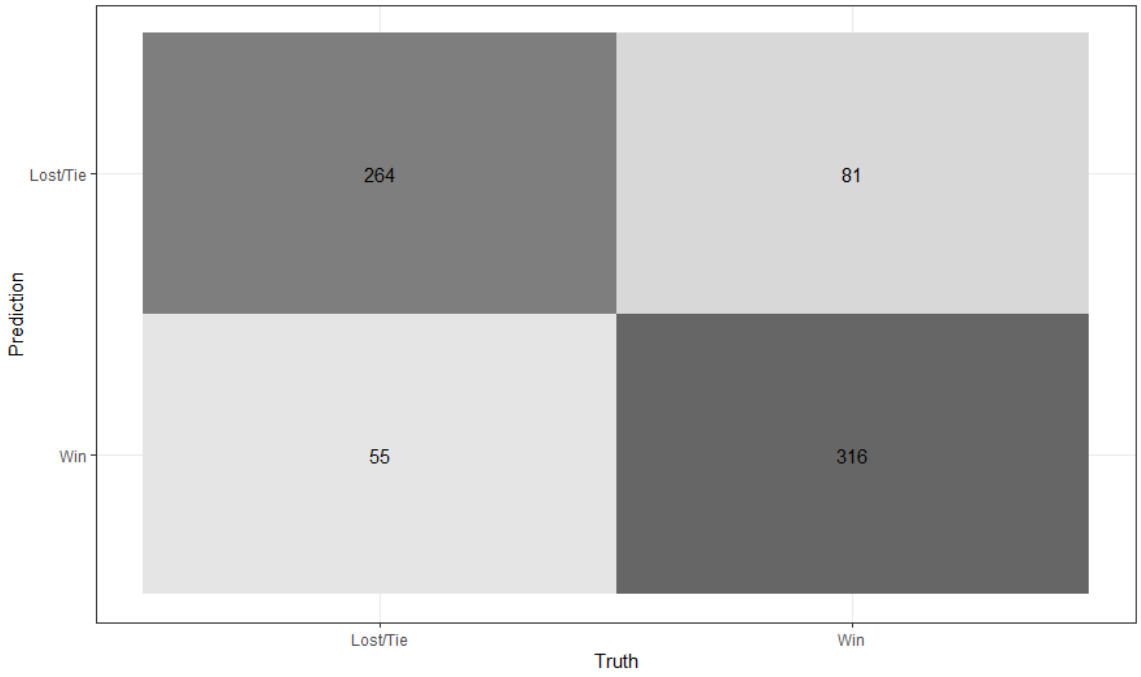

Figura 7: Matriz de Confusão

Com isso temos que meu modelo previu corretamente como Derrota ou Empate do time mandante em 264 de 319 e preveu corretamente 316 de 397 vitórias do time mandante. Obtendo assim uma sensibilidade de 0.828 e uma especifidade de 0.796.

4 Conclusão

Este trabalho proporcionou uma abordagem abrangente, integrando técnicas de Web Scraping, análise exploratória e modelagem preditiva. Os resultados obtidos não apenas adicionam conhecimento ao cenário do futebol americano, mas também abrem caminhos para futuras pesquisas e aplicações práticas. Este estudo representa um passo significativo na aplicação da análise de dados ao esporte, oferecendo insights valiosos para entusiastas, estrategistas e pesquisadores.

A modelagem preditiva utilizando o algoritmo Random Forest representou um passo crucial para compreender as dinâmicas subjacentes aos resultados dos mandantes nas partidas, em que o mesmo demonstrou uma acurácia significativa na previsão dos resultados dos mandantes nas partidas, indicando uma capacidade robusta de generalização para dados não vistos. A seleção cuidadosa de parâmetros, a validação cruzada e a avaliação meticulosa do desempenho do modelo nos conjuntos de treino e teste proporcionaram uma visão integral de sua eficácia.

Recomenda-se a adição de novas variáveis que aumentem a acurácia do modelo, como exemplo Número de Jogadores All-Pro do time mandante ou ao conjunto de dados. Isso pode fornecer insights valiosos sobre a influência de jogadores excepcionais no desempenho da equipe. A presença de talentos individuais de alto nível pode ser um fator crucial em determinar os resultados das partidas.

5 Referência

Referências

[1] Notas de aula;

[2] https://www.pro-football-reference.com

[3] https://introbigdata.org/