Probabilidade de vitória na National Football League

Renan de Oliveira Ferretti

***Resumo*—A análise de esportes têm recebido mais atenção com o passar dos anos, sendo aplicada em qualquer aspecto que possa retornar uma vantagem competitiva. Seguindo essa tendência, foi usado um método de aprendizado de máquina a fim de criar modelos probabilísticos com maior acurácia para avaliar como cada jogada influencia no resultado do jogo.**

Palavras-Chave—Análise de esportes, Aprendizado de máquina, Probabilidade.

***Abstract*—Sports analysis has received more attention over the years, being applied to any aspect that can return a competitive advantage. Following this trend, a machine learning method was used to create probabilistic models with greater accuracy to assess how each move influences the outcome of the game.**

Keywords—Sports analytics, Machine learning, Probability.

I. Introdução

Em muitos esportes as famosas "viradas inesperadas" vêm ocorrendo ao longo dos anos, seja no futebol, basquete, vôlei, futebol americano ou qualquer outra modalidade esportiva. Tais resultados desafiam a acurácia de modelos de predição já consolidados que vinham sendo usados, visto que a probabilidade de um time ganhar varia de acordo com o andamento do jogo.

Os modelos de predição de vitória durante o jogo fornecem a probabilidade de que uma determinada equipe vencerá um jogo dado o estado atual da partida. Estes modelos estão se tornando cada vez mais populares durante o últimos anos, principalmente porque eles podem fornecer informações importantes para que sejam tomadas decisões, como táticas de jogo. Além disso, podem potencialmente melhorar a experiência dos fãs ao fornecer este tipo de informação.

O conceito de modelagem de probabilidades à medida que um evento progride não é novo ou exclusivo de eventos esportivos[1]. Além de serem usados para uma variedade de jogos do National Football League, College Basketball, Major League Soccer entre outros[2][3], as probabilidades de vitória também são usadas para previsões políticas, como em FiveThirtyEight[4].

Neste projeto, será elaborado um modelo de predição do vencedor de determinada partida da National Football League (NFL) levando em consideração os eventos que estão ocorrendo durante o jogo. Este modelo pode ajudar na avaliação dos jogadores, técnicos e dos times que estão jogando e nas tomadas de decisão dentro de campo.

A literatura deste tópico inclui experimentos de Dennis Lock e Dan Nettleton sobre a eficácia de vários algoritmos para previsão de probabilidade de vitória, no qual eles mostram a eficácia do algoritmo Random Forest[5], e o modelo de probabilidade de vitória do Pro Football Reference (PFR). O PFR é um dos modelos probabilísticos de vitória mais conhecidos atualmente, ele supõe que a diferença de pontos durante o jogo pode ser aproximada por uma distribuição normal.

II. Metodologia

Os dados utilizados no projeto contém todas as jogadas de todos os jogos da temporada regular de 2009 a 2016 obtidos diretamente do site NFL.com. A base de dados contém 102 variáveis e 362.447 observações. Cada observação nestes dados é uma peça ou evento que ocorreu durante um jogo.

Após a investigação inicial dos dados, o primeiro passo foi remover as variáveis que não são relevantes para a modelagem. As variáveis removidas neste estágio incluem informações do jogador e data do jogo. A abordagem de modelagem não se importa com quais jogadores estavam envolvidos em um jogo, apenas com o resultado do jogo em relação a pontuação, posição de campo, etc. Havia também várias variáveis que eram de alguma forma duplicadas ou transformações de outras variáveis, portanto também foram removidas.

Após as etapas iniciais de limpeza realizadas nos dados, o resultado do jogo para cada jogada foi anexado. Para cada jogo individual, identificou-se a equipe vencedora e a equipe perdedora. Em seguida, acrescentamos uma variável adicional, *result*, a cada jogada que identifica 'W' se o time que comanda a jogada ganhou o jogo, 'L' se a equipe que comanda a jogada perdeu o jogo, caso contrário acrescentou-se 'T' para o jogo que terminou empatado.

É sobre esses dados pré-processados que uma investigação estatística é realizada. Alguns resultados são mostrados na Tabela 1. Nossos resultados confirmam que os dados se comportam conforme o esperado. O maior volume de jogadas ocorre no início de cada quarto e no final do jogo (Figura 1). A maioria das jogadas são executadas no meio do campo, i.e. 40-60 jardas da *end zone* do ataque, e mais jogadas são executadas na primeira descida do que na segunda (Figura 2).

TABELA I. RESULTADOS DOS DADOS FILTRADOS

	Quantidade
Número total de jogadas ocorridas	267,788
Número total de descidas ocorridas	267,788
Número de touchdowns	10,268
Número de safeties	112
Número de interceptações	3,850
Número de penalidades	3,950
Número de field goals	7,767

O primeiro passo para a criação do modelo de aprendizagem de máquina é obter as estimativas de probabilidade para cada jogada, o qual indica como o resultado dessa jogada afeta a probabilidade da equipe com posse da bola ganhar o jogo. Para obter este resultado, a seleção de atributos precisa descobrir as variáveis que influenciam o resultado dos jogos. Algumas das mais importantes a serem incluídas são:

- Diferencial de pontuação: um valor positivo indica que o ataque está vencendo, um valor negativo valor significa que o ataque está perdendo, e as equipes estão empatadas se é zero.
- Descida e Distância a percorrer: identifica qual é a descida e o número de jardas para alcançar a próxima primeira descida.
- Posição de campo: distância da end zone do ataque.
- Tempo: medido em segundos para evitar ambiguidade entre os quatro períodos de jogo
- Resultado: para conduzir uma abordagem de aprendizado de máquina supervisionado, precisamos saber qual time ganha o jogo e qual time perde o jogo para cada jogada. Isso será codificado como 'W' e 'L'.

A etapa de seleção de atributos pode ser útil na captura de recursos não presentes nos dados fornecidos. Uma dessas transformações leva em consideração o tempo total de jogo, uma vez que a probabilidade de vitória aumenta conforme nos aproximamos do fim. Uma equipe perdendo por três pontos no primeiro quarto não é o mesmo que perder por três pontos faltando um minuto no último quarto.

Para o treinamento do modelo utilizamos todas as jogadas de 2009 a 2015 e os dados da temporada de 2016 foram usados como conjunto de testes. Empregou-se o algoritmo Random Forest para prever as probabilidades de vitória para cada jogada, visto que ele não cai em *overfitting* tão facilmente quanto muitos outros métodos e permite interações não lineares entre atributos[5]. Portanto, executamos o algoritmo Random Forest para criar probabilidades de vitória

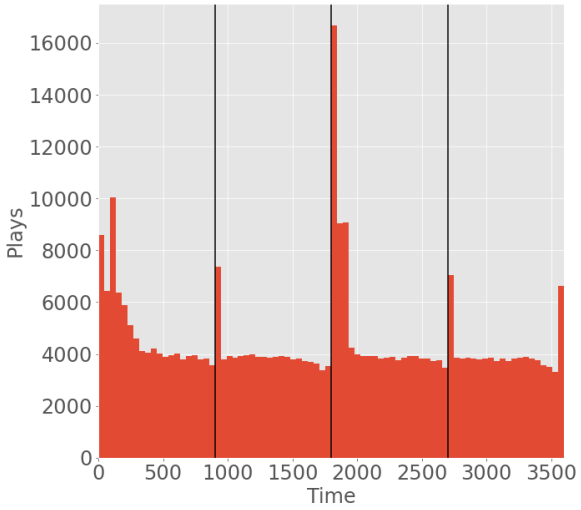

Fig. 1. Histograma do número de jogadas ao decorrer da partida

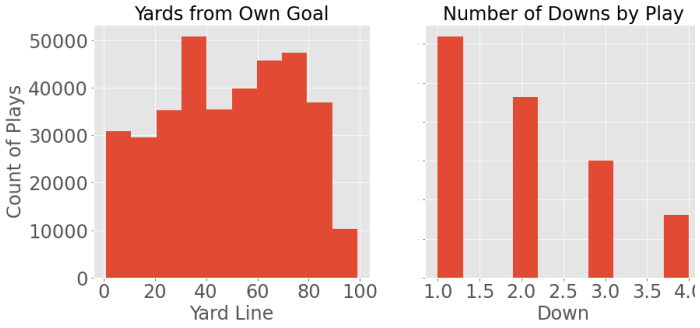

Fig. 2. O histograma à esquerda representa o número de jogadas em determinada parte do campo e o da direita a descida correspondente a cada jogada.

para cada jogada e realizar validação cruzada para identificar hiperparâmetros.

A Figura 3 mostra a mudança nas probabilidades de vitória para um determinado jogo. A partida foi escolhida aleatoriamente do conjunto de testes e resultou no confronto entre Seattle Seahawks contra San Francisco 49ers em 2016. O resultado do jogo foi 25-23 para os Seahawks. A curva azul indica a probabilidade de vitória do Pro Football Reference (PFR)[10], enquanto a linha vermelha é o modelo previsto. É possível observar que a previsão do modelo segue de perto o padrão da indústria, com um ligeiro aumento na volatilidade dos pontos. Essas diferenças são difíceis de explorar sem mais informações sobre o modelo PFR.

Sabemos que o tempo possui influência direta sobre a probabilidade de vitória de um time, então queremos aumentar o peso dele no nosso modelo. Para isso seria vantajoso usar outro método de aprendizagem de máquina. A regressão logística é uma classificação eficaz que foi usada nesta etapa da metodologia. Ela fornece um conjunto de pesos mais otimizado para o problema atual. Neste caso, a formulação da regressão logística calcula a probabilidade do time da casa vencer dado um vetor de probabilidades de vitória amostrado durante o jogo, conforme descrito na equação 1.

$$Pr(R = 1|x) = \frac{exp(w^T * x)}{1 + exp(w^T * x)} \quad (1)$$

R é a variável de resultado dependente do nosso modelo representando se o time da casa ganha ou perde, x é um vetor de probabilidades amostradas de vários pontos em um jogo, enquanto o vetor coeficiente w inclui os pesos para cada

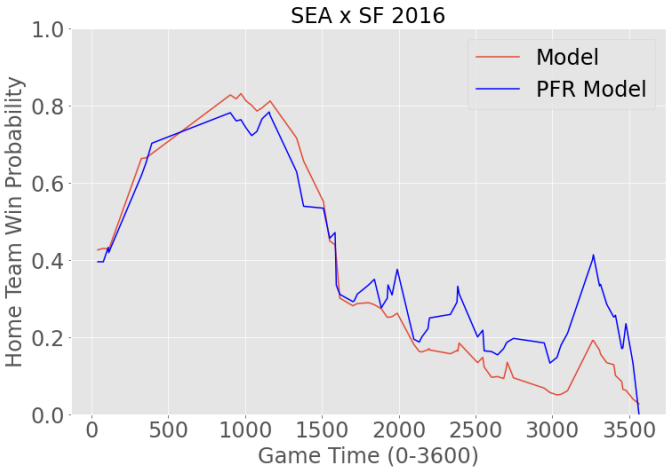

Fig. 3. Probabilidade de vitória para um modelo que utiliza Random Forest versus o modelo do Pro Football Reference.

probabilidade [6]. Dessa maneira, novamente criamos dois hiperparâmetros, um tempo de deslocamento e um número de probabilidades, para usar como entrada.

Para cada combinação de hiperparâmetros, iteramos através de cada jogo e amostramos o número de probabilidades, excluindo o tempo especificado pelo tempo de deslocamento do final e início do jogo. Depois de percorrer todos os jogos para um conjunto de hiperparâmetros, teremos essencialmente criado um novo conjunto de dados com uma linha para cada jogo. Usando o resultado do jogo como rótulo, criamos um modelo de Regressão Logística para avaliar o quão bem as predições se comportam em relação aos valores reais.

III. Resultados

É necessário reconhecer que os valores obtidos são meras estimativas, sendo que a probabilidade real de vitória de uma equipe é quase impossível de determinar com certeza. Tendo isso em mente, serão apresentados os resultados do projeto nesta seção. A Figura 4 ilustra uma curva ROC mostrando como a predição das probabilidades são mapeadas para os rótulos reais (0 é uma derrota, 1 é uma vitória). Pode-se notar que o modelo resulta em um valor de AUC de 0,87. Tendo em vista que a acurácia para o modelo foi de 77%, a combinação entre essas duas métricas demonstram que o modelo está funcionando adequadamente e o conjunto de dados não é desbalanceado.

A Figura 5 mostra um mapa de calor de acurácia em uma grade de hiperparâmetros, utilizando a metodologia de

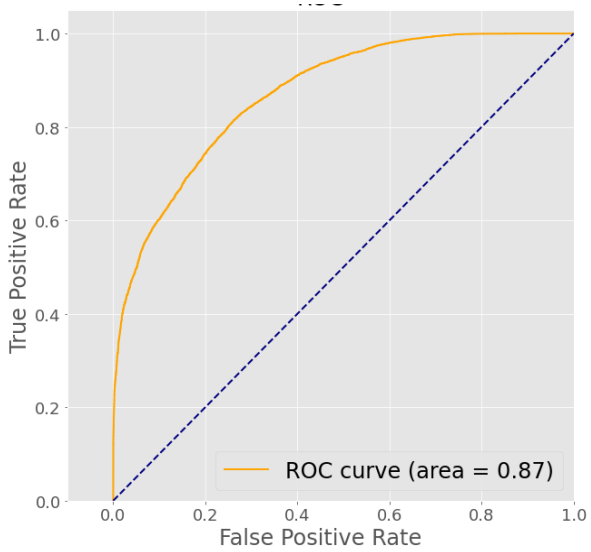

Fig. 4. ROC e AUC para o modelo Random Forest desenvolvido.

regressão logística especificada acima. Foram utilizados 1.486 jogos no conjunto de treinamento e 208 jogos no conjunto de testes. À medida que definimos o tempo de deslocamento para estar mais longe do final do jogo (decrescente no eixo y), as predições se tornam cada vez menos precisas. O interessante da imagem é que o número de probabilidades amostradas (eixo x) não mostra nenhum padrão discernível quanto à precisão de resultados.

IV. Conclusões

Como foi dito na introdução, “viradas fantásticas" acontecem nos esportes. Na verdade, esses acontecimentos são muitas vezes os momentos mais famosos e gratificantes do esporte. Neste projeto, foram estimadas probabilidades para testar o quão cedo em um jogo pode-se prever o resultado com precisão e quantas amostras de probabilidades no jogo precisaríamos. Um fato interessante dos resultados é que o modelo consegue classificar corretamente quase 90% dos jogos a 12 minutos do final. Isso indica que a maioria dos jogos são decididos logo após o começo do último quarto.

Os resultados indicam que as probabilidades fornecidas pelo nosso modelo são consistentes e bem calibradas. Vale ressaltar que o estudo não visa desacreditar os modelos de probabilidade de vitória existentes, mas sim explorar a capacidade de modelos simples e interpretáveis alcançarem a ser pesquisado futuramente é a reavaliação de tipos semelhantes de modelos, dado que o jogo muda rapidamente desempenho similar aos mais complicados. Um ponto crucial tanto devido a mudanças nas regras, mas também devido a mudanças nas habilidades dos jogadores ou mesmo a análises. Esses tipos de mudanças podem ter uma implicação em quão segura uma pontuação diferencial de x pontos é, pois as equipes podem cobrir a diferença mais rápido. Assim, os detalhes do modelo também podem mudar.

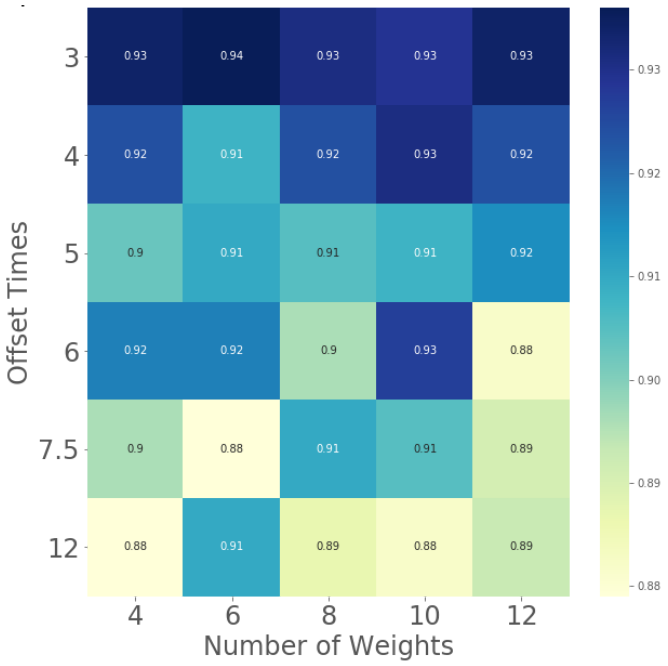

Fig. 5. Acurácia da predição de vitória para diferentes tempos de deslocamento e probabilidades de vitória realizadas durante o jogo.

Agradecimentos

Ao Prof. Dr. Leonardo Tomazeli Duarte pela explicação do conteúdo ao longo do semestre, o que possibilitou a compreensão e o desenvolvimento da pesquisa realizada.

Referências

[1] MADSEN, Richard. On the Probability of a Perfect Progression. **The American Statistician.** New York, p. 214-216. maio 1987.

[2] KVAM, Paul; SOKOL, Joel S. A logistic regression/Markov chain model for NCAA basketball. **Naval Research Logistics (NrL)**, v. 53, n. 8, p. 788-803, 2006.

[3] RYDER, Alan. Win Probabilities: A tour through win probability models for hockey. **Hockey Analytics, www. hockeyanalytics. com**, 2004.

[4] Nate Silver. 2016. 2016 Election - FiveThirtyEight. (2016). fivethirtyeight.com/politics/elections/.

[5] LOCK, Dennis. **Statistical methods in sports with a focus on win probability and performance evaluation**. 2016. Tese de Doutorado. Iowa State University.

[6] ROBBERECHTS, Pieter; VAN HAAREN, Jan; DAVIS, Jesse. Who will win it? An in-game win probability model for football. **arXiv preprint arXiv:1906.05029**, 2019.