Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística

Felipe Cordeiro Melo

Modelagem Estatística para Previsão de Pontuações no Fantasy Football: Uma Abordagem Baseada em Dados Históricos da NFL

Curitiba
2025

Felipe Cordeiro Melo

Modelagem Estatística para Previsão de Pontuações no Fantasy Football: Uma Abordagem Baseada em Dados Históricos da NFL

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Graduação em Estatística da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Orientador(a): Profa. Dra. Amanda Merian Freitas Mendes

Curitiba
2025

Dedico este trabalho ao meu falecido pai, por ser minha fonte de inspiração.

Agradecimentos

À minha mãe, Luciane, agradeço por todo amor e carinho que recebo, por sempre me incentivar e me cobrar. Agradeço pela força nos últimos anos e por sempre proteger eu e meus irmãos.

Ao meu pai, Sidnei, agradeço por ter sido um exemplo pra mim, por tudo o que me ensinou desde a infância e por ter me motivado a seguir na carreira de exatas. Sei que a cada novo passo da minha vida você sempre estará do meu lado.

À minha companheira de vida, Gabriele, por todos os dias estar do meu lado, me apoiando e sendo meu porto seguro.

Ao meu irmão e à minha irmã, Arthur e Júlia, por me desafiarem a ser uma pessoa melhor, e todas as habilidades que conquistei pelo simples motivo de tentar ser melhor que meus irmãos.

À minha sogra e ao meu sogro, Rosiele e Tiago, por terem me acolhido desde o primeiro dia, e estarem do meu lado durante toda essa caminhada.

Às minhas amadas filhas, Cláudia e Isabella, pelo apoio emocional.

Aos meus avós, por sempre serem um ponto de fuga da rotina, e por trazer paz e tranquilidade nos momentos juntos.

Aos meus amigos, Bruno e Henrique, por estarem ao meu lado desde a infância, crescendo e compartilhando histórias juntos.

Aos meus amigos de faculdade, Nilton, Paulo, Pedro e Willian, por todos os momentos, seja fazendo trabalho ou estudando pra provas, seja em momentos de descontração.

E, por fim, à minha orientadora, Amanda, por ter aceitado esse desafio, por todos os ensinamentos e por ter confiado em mim.

A perfeição não é atingível. Mas se perseguir a perfeição,
podemos alcançar a excelência.
Vince Lombardi.

Resumo

O objetivo deste estudo é desenvolver e avaliar modelos estatísticos para a previsão das pontuações de jogadores no Fantasy Football, utilizando dados históricos da carreira dos atletas. Para isso, foram empregadas estatísticas de desempenho de jogadores da *National Football League* e do *College Football*. Com base nesses dados, aplicaram-se técnicas avançadas de modelagem preditiva, incluindo um modelo em duas etapas para lidar com a inflação de zeros, algoritmos de *ensemble* baseados em árvores de decisão (*Random Forest* e *XGBoost*) para tarefas de classificação e regressão, além de Modelos Lineares Generalizados com distribuição Binomial para classificação e Gamma para regressão. Os modelos foram avaliados para todas as posições ofensivas contempladas no *Fantasy Football* (*Quarterbacks*, *Running Backs*, *Wide Receivers* e *Tight Ends*), com distinção entre jogadores calouros e veteranos, a fim de investigar variações de desempenho conforme posição e nível de experiência. Os métodos utilizados foram capazes de alcançar bons resultados tanto em modelos de classificação, lidando adequadamente com a inflação de zeros, quanto em modelos de regressão. Além disso, a segmentação por posição e experiência contribuiu para a melhoria da precisão preditiva, revelando padrões distintos de desempenho. As principais contribuições deste estudo incluem a demonstração da aplicabilidade de técnicas estatísticas e de aprendizado de máquina na previsão de pontuações no *Fantasy Football*, fornecendo uma base metodológica robusta para o tratamento e análise desses dados.

Palavras-chave: Fantasy Football. NFL. Modelagem em Duas Etapas. Modelos Baseados em Árvores. Modelos Lineares Generalizados.

Sumário

1 Introdução

A *National Football League* (NFL) é a principal liga de futebol americano do mundo. Ela foi criada em 1920 e é composta por 32 times, divididos entre as conferências *American Football Conference* (AFC) e *National Football Conference* (NFC). Sua temporada regular dura 18 semanas, com cada time disputando 17 jogos e tendo uma semana de descanso, os 7 melhores times de cada conferência se classificam para *playoffs*, em que eles se enfrentam em confrontos eliminatórios, essa fase tem duração de 5 semanas. Assim, cada conferência tem o seu campeão, e os dois representantes das conferências se enfrentam na final, o *Super Bowl* (NFL, 2025a).

Cada partida é composta por duas equipes com 11 jogadores em campo, divididos entre unidades de ataque e defesa, cujas funções variam conforme suas posições. O objetivo principal é avançar com a bola até a extremidade do campo para marcar um *Touchdown*, que representa a maior pontuação possível no jogo. Para isso, a equipe ofensiva dispõe de quatro tentativas (ou descidas) para percorrer ao menos 10 jardas. Caso tenha êxito, conquista uma nova série de quatro jogadas, podendo prosseguir até a zona de pontuação. Se não alcançar esse avanço, a posse de bola é transferida para a equipe adversária. O avanço da bola pode ocorrer de duas formas principais: por meio de passes aéreos, em que um jogador lança a bola para um companheiro, ou por corridas, em que o jogador recebe a bola e tenta ganhar jardas correndo com ela (NFL, 2025c).

Cada time, possuí substituições ilimitadas, fazendo com que haja rotação nas formações, estilo de jogo e dos próprios jogadores dentro de um mesmo jogo. No time ofensivo, as posições existentes são de *Quarterback, Wide Receiver, Tight End, Running Back* e jogadores da linha ofensiva. Já na defesa, atuam os jogadores da linha defensiva, os *Linebackers* e os *Defensive Backs*. Cada posição carrega estereótipos físicos característicos. Por exemplo, jogadores das linhas ofensiva e defensiva tendem a ser mais altos e pesados, sendo responsáveis, respectivamente, por proteger os demais atletas no ataque e por compor a primeira linha de contenção na defesa (NFL, 2025b). Existem ainda jogadores chamados "especialistas" que são os responsáveis por jogadas de chute, geralmente há 3 por time, o *Kicker*, o *Punter* e o *Long Snapper*.

A temporada da NFL é curta, por conta da fisicalidade do esporte, ela dura menos de 5 meses, e cada time terá de 17 a no máximo 21 jogos, se chegar até o *Super Bowl*. Se compararmos com ligas de outros esportes, como a *National Basketball Association* (NBA), que têm duração de 6 meses, e que os times jogam de 82 a 110 jogos (NBA, 2025). Outro exemplo nos Estados Unidos é o da *Major League Baseball* (MLB) (MLB, 2025), cuja temporada também dura 6 meses, e o número de jogos pode chegar até 185. Em outros países, nas principais ligas de futebol, a temporada dura mais de 11 meses, como é o caso no Brasil (GE, 2024a), e na Europa em que times chegam a disputar 80 partidas na mesma temporada (GE, 2024b).

Essa escassez de jogos fez com que os fãs buscassem maneiras alternativas de se manterem conectados à liga e seus jogadores, impulsionando o surgimento e a popularização do *Fantasy Football*. Essa modalidade permitiu que torcedores acompanhassem o desempenho de múltiplos jogadores e times, não apenas de suas equipes favoritas, aumentando o engajamento e o consumo de conteúdo da NFL. Como resultado, o *Fantasy Football* se tornou um fenômeno bilionário, com impacto direto na receita da liga, no mercado de mídia esportiva e na indústria do entretenimento digital (ALMEIDA; ALMEIDA; LIMA, 2015).

O *Fantasy Football* é uma competição virtual em que seus participantes se reúnem em uma liga e ali cada um monta seu time no início da temporada, e rodada a rodada escala jogadores titulares e reservas, que vão pontuar de acordo com seu desempenho dentro de campo. Cada liga pode ter regras personalizadas, para o presente estudo iremos considerar as normas padrões do *Fantasy* da NFL (NFL, 2024).

Uma liga geralmente é composta de 8 a 16 times, e cada time é formado por 15 jogadores reais da NFL. As equipes são montadas por meio de um *Draft* com 15 rodadas, no formato *Snake*. Nesse formato, é sorteada uma ordem aleatória para a escolha dos jogadores, e cada time realiza uma seleção por rodada, com a ordem sendo invertida a cada nova rodada. Ou seja, o time que tiver a primeira escolha na primeira rodada terá a última escolha na segunda, a primeira na terceira, e assim por diante. O time com a primeira escolha pode selecionar qualquer jogador disponível, o segundo, qualquer jogador exceto aquele já escolhido, e assim sucessivamente. Dessa forma, os jogadores não se repetem entre os times.

Dentre os 15 jogadores selecionados, cada time pode escalar apenas 09 titulares por rodada e, os 6 restantes permanecem no banco de reservas, e apenas a pontuação dos titulares será considerada. A escalação dos jogadores titulares deve respeitar os seguintes limites por posição, a saber:

- 1 *Quarterback* (QB)
- 2 *Running Backs* (RB)
- 2 *Wide Receivers* (WR)
- 1 *Tight End* (TE)
- 1 *Flex* (RB/WR)
- 1 *Kicker* (K)
- 1 Defesa completa

Os jogadores na reserva podem ser de qualquer posição. A pontuação dos jogadores depende do desempenho deles em campo, cada jarda conquistada e *touchdown* anotado soma na pontuação, cada erro também será contado, porém, de forma negativa.

A cada rodada, há confrontos diretos entre os times, e aquele que somar mais pontos no duelo é considerado o vencedor. Esse formato se mantém durante as 14 primeiras

rodadas da temporada. A partir da 15ª rodada, iniciam-se os *playoffs*, nos quais os times classificados continuam se enfrentando, porém agora em confrontos eliminatórios, quartas de final, semifinal e final, até a definição do campeão da liga.

Na última década, observa-se um crescimento significativo no uso de estatísticas avançadas por parte dos times da NFL e pela liga como um todo (GROTHAUS, 2024). Um dos principais exemplos dessa tendência é o processo de *Draft* da NFL, que é a seleção de jogadores provenientes do futebol americano universitário para a liga profissional (Schneider Downs, 2023).

Contudo, o *Fantasy Football* é um tema que não acompanhou esse desenvolvimento. Abadzic, Cheun e Patel (2024), Morgan et al. (2019) e Lutz (2015), por exemplo, exploram a estimativa da pontuação dos atletas e, ao analisar as bases de dados da Web of Science (2025) e Scopus (2025), não foram encontrados estudos que modelassem a pontuação anual de todas as posições ofensivas, levando em consideração, de forma simultânea, também os jogadores calouros. A NFL e sites especializados frequentemente produzem rankings e projeções de desempenho para os atletas, mas essas estimativas costumam se basear em fatores como o desempenho coletivo das equipes e expectativas subjetivas sobre o papel que cada jogador exercerá no ataque (SMOLA, 2025). Essas projeções, portanto, dependem fortemente da interpretação e opinião de especialistas. Diante disso, a presente discussão se justifica pela proposta de desenvolver um modelo preditivo baseado em dados objetivos, que considere o desempenho histórico individual dos atletas. Ao integrar jogadores com diferentes níveis de experiência em uma mesma estrutura analítica, busca-se oferecer uma alternativa sistemática e transparente às previsões tradicionais, contribuindo para a ampliação da literatura acadêmica sobre o tema e fornecendo ferramentas tanto para entusiastas do *Fantasy Football* quanto para analistas esportivos.

Para tanto, o objetivo desta pesquisa é estimar a pontuação anual de atletas da NFL em todas as posições ofensivas contempladas pelo Fantasy Football, adotando abordagens distintas para jogadores calouros, oriundos do College Football, e para atletas já atuantes na liga profissional.

2 Revisão de Literaura

A revisão de literatura foi guiada por estudos que buscavam compreender o desempenho de jogadores no *Fantasy Football*, além de pesquisas voltadas à transição de atletas do futebol universitário para a NFL. Também foram considerados trabalhos que utilizaram variáveis resposta contínuas e assimétricas, semelhantes às observadas neste estudo.

No que diz respeito à predição no *Fantasy Football*, destaca-se o trabalho de Abadzic, Cheun e Patel (2024), no qual os autores elaboram, para a temporada de 2024, um *ranking* com as 12 maiores projeções para cada posição ofensiva e também para os *Kickers*. Para isso, foram utilizadas variáveis relacionadas às estatísticas dos atletas - como jardas e touchdowns -, bem como ao contexto em que o jogador está inserido, incluindo idade e força dos adversários. Foram testadas diferentes técnicas de modelagem, como regressão *Ridge* e *Ridge* Bayesiana, regularização *Elastic Net*, *Random Forests* e *Gradient Boosting*. Os modelos foram ajustados separadamente por posição, com diferentes variáveis explicativas em cada caso. Os melhores resultados foram obtidos com *Random Forests*, no entanto, os autores não consideraram atletas calouros, além disso, só foram apresentados os resultados das 12 maiores pontuações de cada posição, o que limita a avaliação do modelo para jogadores fora do topo do ranking.

Morgan et al. (2019), nesse aspecto, procuram indentificar atletas subvalorizados no *Fantasy Draft*, definidos como aqueles selecionados após a quarta rodada e que finalizaram a temporada entre os 20 melhores de sua posição, contrastando com Abadzic, Cheun e Patel (2024), pois incluem jogadores oriundos do *College Football*. Jogadores calouros frequentemente se enquadram nesse perfil, já que, por não possuírem histórico na NFL, representam escolhas arriscadas nas rodadas iniciais do *Draft*. Para identificar esses atletas, os autores utilizaram variáveis que vão além do desempenho em campo, como métricas físicas que indicam o potencial atlético do jogador e informações contextuais sobre sua provável participação no ataque da equipe. Os resultados foram satisfatórios com o uso de regressão Lasso, *Random Forests* e *XGBoost*, embora o estudo tenha se limitado às posições de RB e WR.

Em Mulholland e Jensen (2016), o objetivo foi modelar o desempenho futuro de jogadores em transição do futebol universitário para a NFL, com foco na projeção de sucesso desses atletas na liga profissional, e não em sua pontuação no *Fantasy*. Para isso, os autores criaram a variável `NFL_Career_Score`, calculada como a soma das jardas recebidas com 19,3 vezes o número de *touchdowns*, refletindo o valor esperado desses eventos em pontos. Foram utilizadas variáveis relacionadas ao desempenho e estatísticas na carreira universitária, características físicas (como altura, peso e IMC), além da conferência universitária em que o atleta atuou. O estudo considerou apenas as posições de *Wide Receiver* e *Tight End*, sendo construídos modelos separados para cada uma. Como

estratégias de modelagem, os autores utilizaram Regressão Linear Múltipla e Árvores de Decisão. Ambas as variáveis resposta associadas ao desempenho apresentaram distribuição assimétrica à direita.

Assim como no trabalho de Mulholland e Jensen (2016), as variáveis resposta analisadas neste estudo também apresentam distribuições assimétricas à direita. Dessa forma, foram consideradas referências metodológicas que tratam diretamente desse tipo de problema, mesmo fora do contexto esportivo. Destaca-se, nesse sentido, o estudo de Ng e Cribbie (2017), que utiliza um Modelo Linear Generalizado (MLG) com distribuição Gamma para modelar variáveis contínuas assimétricas e com heterocedasticidade, em aplicações na área da psicologia, utilizando tanto variáveis explicativas contínuas quanto categóricas.

Outro desafio comum em variáveis assimétricas é a inflação de zeros, especialmente em situações em que parte significativa das observações assume valor zero. Esse problema foi tratado em Rožanec et al. (2025) por meio de um modelo em duas etapas (*two-part model*), que consiste, inicialmente, em uma classificação binária para identificar se a variável resposta é igual a zero ou não, seguida por um modelo de regressão aplicado apenas aos casos com valor diferente de zero.

A partir da revisão, observou-se que os estudos voltados à modelagem do desempenho de atletas da NFL frequentemente adotam métodos baseados em Árvores de Decisão, como *Random Forests* e modelos de *Boosting*. Dentre as variáveis utilizadas, além do desempenho estatístico em campo, destacam-se características físicas dos atletas. Outro ponto relevante abordado foi a distribuição da variável resposta, com destaque para o uso de modelos MLG com distribuição Gamma e modelos em duas etapas. Assim, o presente estudo se fundamenta nas abordagens discutidas na literatura, adaptando-as ao contexto do *Fantasy Football* com foco específico na previsão da pontuação anual dos atletas, incluindo jogadores calouros.

3 Material e Métodos

3.1 Material

3.1.1 Base de Dados

Os dados analisados neste trabalho foram extraídos de três fontes principais: os pacotes `nflfastR` (CARL; BALDWIN, 2024), `nflreadR` (HO; CARL, 2024) e `cfbfastR` (GILANI et al., 2021), todos disponíveis no *software* R (R Core Team, 2025). No total, foram coletadas cinco bases de dados distintas, descritas a seguir:

3.1.1.1 College Football Rosters

Esta base foi extraída por meio da função `cfbfastR::load_cfb_rosters()` e reúne informações sobre os elencos das equipes universitárias de futebol americano por temporada. As colunas disponíveis contemplam dados pessoais dos jogadores, bem como informações sobre os times em que atuaram em cada ano. Cada registro possui um identificador único (`athlete_id`), além do nome completo do jogador (`full_name`), ano da temporada (`season`), nome da equipe universitária (`team`), conferência da equipe (`conference`) e a posição em que o jogador atua (`position`).

Embora não inclua os elencos de todas as equipes do *College Football*, com a ausência de algumas conferências menores e times menos expressivos, a base contempla a maioria dos principais times e jogadores, sendo amplamente representativa. No total, abrange dados das temporadas de 2004 a 2024, com 245.410 registros, dos quais 13.262 correspondem a QBs, 24.336 a RBs, 38.693 a WRs e 14.721 a TEs.

3.1.1.2 College Stats

Esta base foi obtida utlizando a função `cfbfastR::cfbd_stats_season_player()` e apresenta as estatísticas de desempenho dos jogadores do futebol americano universitário por temporada.

Cada registro contém, além do identificador (`athlete_id`) e do ano da temporada (`season`), estatísticas relacionadas a passe, corrida, recepção e defesa. Neste estudo, apenas as estatísticas ofensivas foram consideradas. A base contém um total de 164.458 registros, sendo 9.395 de QBs, 18.944 de RBs, 27.694 de WRs e 9.078 de TEs.

As variáveis consideradas nesta base incluem:

- Estatísticas de passe:
  - `passing_att`: tentativas de passe,
  - `passing_completions`: passes completos,
  - `passing_yds`: jardas passadas,

  - `passing_td`: passes para touchdown,
  - `passing_int`: passes interceptados;

- Estatísticas de corrida:

  - `rushing_car`: corridas realizadas,
  - `rushing_yds`: jardas corridas,
  - `rushing_td`: corridas para touchdown;

- Estatísticas de recepção:

  - `receiving_rec`: recepções realizadas,
  - `receiving_yds`: jardas recebidas,
  - `receiving_td`: recepções para touchdown.

3.1.1.3 NFL Roster

Esta base foi extraída por meio da função `nflfastR::fast_scraper_roster()` e apresenta estrutura semelhante à da tabela *College Football Roster*, com a diferença de que contempla os elencos dos 32 times da NFL. Abrange as temporadas de 1999 a 2024, sendo cada jogador identificado de forma única pela coluna `gsis_id`.

Estão presentes algumas colunas comuns às do *College*, como `full_name`, `season`, `team` (neste caso, referente ao time da NFL) e `position`. Além dessas, a base inclui informações adicionais como a data de nascimento (`birth_date`), altura em polegadas (`height`), peso em libras (`weight`), número da escolha no *draft* da NFL (`draft_number`), ano de entrada na liga (`entry_year`), anos de experiência na NFL (`years_exp`) e o identificador `espn_id`, que permite a integração com as bases de dados do *College Football*.

Essa base é composta de 63.343 linhas, sendo 2.917 QBs, 4.821 RBs, 7.476 WRs e 4.053 TEs.

3.1.1.4 NFL Stats

Esta tabela é extraída por meio da função `nflfastR::calculate_stats()`. Assim como nas tabelas de *rosters*, sua estrutura é semelhante à da base do *College*, ela traz uma coluna adicional de partidas disputadas pelo jogador (`games`), e também as estatísticas de desempenho em campo dos jogadores da NFL, incluindo métricas de passe, corrida, recepção e também estatísticas defensivas. Todas as colunas presentes na base *College Stats* também estão contidas aqui, com a adição de algumas variáveis específicas:

- Estatísticas de passe:

  - `sacks_suffered`: número de vezes em que o *Quarterback* foi derrubado pela defesa adversária.

- Estatísticas de corrida:

  - `rushing_first_downs`: número de corridas que resultaram em uma primeira descida.

- Estatísticas de recepção:

  - `targets`: número de vezes em que o jogador foi alvo de um passe,
  - `receiving_yards_after_catch`: jardas recebidas após a recepção,
  - `receiving_first_downs`: número de recepções que resultaram em uma primeira descida.

Esta é também a base que contém a variável resposta `fantasy_points`, calculada com base nas Equações (3.1), (3.2) e (3.3).

$$\begin{aligned} \textbf{Pontuação de Passe} = {} & 0{,}04 \times \text{jardas passadas} \\ & + 4 \times \text{passes para touchdown} \\ & - 2 \times \text{passes interceptados} \end{aligned} \tag{3.1}$$

$$\begin{aligned} \textbf{Pontuação de Corrida} = {} & 0{,}1 \times \text{jardas corridas} \\ & + 6 \times \text{corridas para touchdown} \\ & - 2 \times \text{fumbles sofridos} \end{aligned} \tag{3.2}$$

$$\begin{aligned} \textbf{Pontuação de Recepção} = {} & 0{,}1 \times \text{jardas recebidas} \\ & + 6 \times \text{recepções para touchdown} \\ & - 2 \times \text{fumbles sofridos} \end{aligned} \tag{3.3}$$

Desse modo, os pontos no *fantasy* são dados pela soma dessa pontuações, conforme apresentado na Equação (3.4).

$$\textbf{fantasy\_points} = \text{Pontuação de Passe} + \text{Pontuação de Corrida} + \text{Pontuação de Recepção} \tag{3.4}$$

3.1.1.5 Draft Picks

Essa tabela contém as informações do histórico do *Draft* da NFL, e foi utilizada para complementar a informação contida na coluna `draft_number` da tabela *NFL Roster*, com ela, conseguimos também trazer a rodada em que o jogador foi selecionado. Essa tabela é extraída com a função `nflreadr::load_draft_picks()`, e contém 3 colunas:

- `season`: ano da temporada;
- `round`: rodada em que o jogador foi selecionado;
- `pick`: posição em que o jogador foi selecionado.

3.1.2 Recursos Computacionais

Para o desenvolvimento do trabalho, foi utilizado o *software* `R`, versão 4.4.3 (R Core Team, 2025). Desde a extração dos dados, com os pacotes `nflfastR` (CARL; BALDWIN, 2024), `cfbreadR` (HO; CARL, 2024) e `cfbfastR` (GILANI et al., 2021), passando por etapas de tratamento, análise exploratória e visualização dos dados, realizadas com os pacotes pertencentes à família `tidyverse` (WICKHAM et al., 2019), até as etapas de modelagem, em que foram utilizados os pacotes `ranger` (WRIGHT; ZIEGLER, 2017), para o treinamento dos modelos *Random Forest*, e `mlr` (BISCHL et al., 2016), para os modelos de *XGBoost*.

Para documentação e reprodutibilidade, os dados foram armazenados e gerenciados em banco de dados no ambiente `DBeaver`, versão 25.0.5 (DBeaver Community, 2025). As transformações necessárias até a obtenção da base final foram realizadas por meio de consultas escritas na linguagem `SQL` (Oracle Corporation, 2025).

3.2 Métodos

O presente trabalho exigiu a aplicação de diferentes técnicas de classificação e regressão para a modelagem da variável resposta. Nesta seção, são descritos os métodos estatísticos utilizados para prever a pontuação anual de atletas da NFL no *Fantasy Football*. Modelos de classificação foram empregados nos casos em que a variável resposta apresentava inflação de zeros, com o objetivo de predizer se o jogador obteve ou não pontuação positiva. Para os atletas que pontuaram, foram aplicadas técnicas de regressão voltadas à estimativa da pontuação obtida.

3.2.1 Processamento dos Dados

A etapa de pré-processamento dos dados foi fundamental para consolidar, transformar e preparar as informações brutas antes da aplicação dos modelos. Inicialmente, foi realizada a junção das bases contendo os elencos dos jogadores (*rosters*) com as respectivas estatísticas por temporada, tanto para a NFL quanto para o *College Football*. Em seguida, os dados foram filtrados para manter apenas as posições ofensivas de interesse neste estudo: *Quarterbacks*, *Running Backs*, *Wide Receivers* e *Tight Ends*.

Para cada jogador, foi criada a variável de idade com base na data de nascimento (`birth_date`) e no ano da temporada (`season`). Valores negativos eventualmente presentes na variável resposta foram ajustados para zero, evitando distorções nos modelos preditivos. Também foram incorporadas informações do *draft*, utilizando a base de *Draft Picks*, a partir das colunas `draft_number` e `year`, permitindo a identificação da rodada em que o atleta foi selecionado. Com base nessa informação, foi criada a variável categórica `draft_round_group`, agrupando os jogadores em cinco categorias: `1st round` (rodada 1),

`Early Rounds` (rodadas 2 e 3), `Mid Rounds` (rodadas 4 e 5), `Late Rounds` (rodadas 6 e 7) e `Undrafted` (não selecionados).

No caso dos dados do *College Football*, foi realizada a agregação das estatísticas individuais ao longo da carreira universitária de cada jogador. Nesse processo, foi criada a variável `seasons_played`, que representa o número de temporadas disputadas, a partir da contagem de registros por jogador na base original. Adicionalmente, elaborou-se a variável `conference_group`, com o intuito de distinguir o nível competitivo das universidades. Essa variável categórica classifica as conferências em dois grupos: `Power5`, que engloba as cinco principais conferências do *College Football* (*ACC*, *Big Ten*, *Big 12*, *Pac-12* e *SEC*), e `Others`, representando as demais conferências. Essa distinção permite considerar o contexto competitivo no qual os jogadores atuaram durante sua formação universitária, fator potencialmente relevante para a modelagem de desempenho futuro.

A variável resposta da base de *College* foi obtida por meio da junção com a base da NFL, utilizando os identificadores `athlete_id` e `espn_id` para localizar, na temporada de estreia, a pontuação do atleta no *Fantasy Football*. Além da pontuação, também foram incorporadas variáveis de perfil referentes ao ano de calouro, como idade, altura, peso e informações do *draft*.

Na base da NFL, por sua vez, a variável resposta foi ajustada para refletir a pontuação do jogador na temporada seguinte àquela em que as estatísticas explicativas foram registradas, possibilitando a modelagem prospectiva do desempenho. Após essas transformações, os dados foram segmentados por posição, originando bases específicas para QB, RB, WR e TE. Por fim, foram mantidos apenas os registros com valores não nulos na variável `fantasy_points`, garantindo a consistência dos conjuntos de dados utilizados na modelagem.

3.2.2 Estratégia de Modelagem

Após o pré-processamento, foram obtidas oito bases de dados distintas para a realização das modelagens propostas. Para cada uma das quatro posições ofensivas analisadas, foram construídos dois modelos: um voltado para atletas calouros, em sua primeira temporada na NFL, utilizando informações relativas à carreira universitária, e outro para atletas veteranos, aqueles a partir da segunda temporada na NFL, com base em dados da temporada anterior na liga profissional.

Em cada modelo houve separação dos dados em dois conjuntos: treinamento e teste, com o objetivo de ajustar os modelos preditivos e avaliar seu desempenho em dados independentes. A separação foi realizada de forma aleatória, respeitando a proporção de aproximadamente (70% para treino e 30% para teste). Todas as etapas de ajuste de hiperparâmetros e seleção de variáveis foram conduzidas exclusivamente no conjunto de treinamento, enquanto o conjunto de teste foi reservado para a avaliação final dos modelos.

3.2.3 Modelos em duas etapas

Devido à natureza de algumas posições na NFL, há um número significativo de jogadores reservas que não pontuam no *Fantasy Football*. Um exemplo são os *Quarterbacks*, posição na qual os times podem ter até três atletas no elenco, mas geralmente apenas um atua durante toda a temporada. Como consequência, observa-se uma inflação de zeros na variável resposta. Para contornar esse problema, adotou-se a modelagem em duas etapas, que consiste, primeiramente, em classificar se um jogador pontuará ou não e, em seguida, estimar a pontuação apenas daqueles que efetivamente pontuaram.

Essa técnica foi inicialmente proposta por Cragg (1971) no contexto de variáveis dependentes limitadas, sendo posteriormente aplicada em diferentes áreas, como saúde, sendo posteriormente difundida em diversas áreas com dados excessivamente zerados. Embora os autores não estabeleçam um limite fixo para a proporção mínima de zeros que justifique o uso desse tipo de modelagem, estudos posteriores aplicaram a abordagem em cenários com proporções de zeros superiores a 25% (NG; CRIBBIE, 2017; ROŽANEC et al., 2025). Esse foi o valor de referência considerado neste trabalho para definir os casos em que seria necessário o ajuste por meio de modelos em duas etapas.

3.2.4 Modelos de Classificação

Em casos em que se detectou a presença de inflação de zeros, foi ajustado um modelo de classificação preliminar para predizer se um jogador pontuaria ou não. Para isso, a variável resposta `fantasy_points` foi dicotomizada: valores iguais a zero foram codificados como 0, e valores superiores a zero como 1. O propósito dos modelos classificatórios é estimar a probabilidade de um indivíduo pertencer a uma das duas classes (0 ou 1), com base em um conjunto de variáveis explicativas. A partir dessas probabilidades, define-se um limiar de decisão (*threshold*) que determina a classe predita. Neste estudo, esse limiar foi ajustado com o objetivo de maximizar a acurácia no conjunto de teste.

Foram considerados, nesta etapa, modelos de Regressão Logística Binomial e algoritmos baseados em árvores de decisão, como o *Random Forest* e o *Extreme Gradient Boosting* (*XGBoost*). A comparação entre os modelos ajustados para cada posição foi realizada com base na métrica de Acurácia Balanceada (3.5), além da simplicidade do modelo, priorizando-se aqueles com menor número de variáveis explicativas. Também, foi utilizada a métrica de Acurácia (3.6) para avaliar a capacidade preditiva do modelo final.

$$\text{Acurácia Balanceada} = \frac{1}{2}\Big(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\Big) \tag{3.5}$$

Onde:

- **TP** = Verdadeiros Positivos
- **TN** = Verdadeiros Negativos

- **FP** = Falsos Positivos
- **FN** = Falsos Negativos

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.6}$$

3.2.5 Modelos de Regressão

Na segunda etapa dos casos em que houve inflação de zeros, e também para os casos sem essa característica, procedeu-se à modelagem por regressão, visando predizer uma variável contínua. Nos casos com inflação de zeros, a regressão foi aplicada apenas às observações com valores positivos da variável resposta, enquanto nos demais casos, utilizou-se o conjunto completo.

Os métodos explorados incluíram a Regressão Gamma, além dos modelos de *Random Forest* e *XGBoost*. A avaliação e comparação dos modelos de regressão foi realizada com base na métrica de Erro Médio Absoluto (MAE) (3.7). Assim como nos modelos classificatórios, foi considerada também a parcimônia, priorizando-se modelos com bom desempenho preditivo e menor número de variáveis.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.7}$$

Onde:

- y_i = valor real da observação
- $\hat{y}_i$ = valor predito pelo modelo para a observação
- n = número total de observações

Tanto nos modelos de classificação quanto nos de regressão, foram empregadas técnicas de seleção de covariáveis, as quais serão detalhadas nas subseções dedicadas à explicação de cada modelo.

3.2.6 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (MLG), propostos por Nelder e Wedderburn (1972), são modelos em que se busca uma relação entre o preditor linear, que é uma combinação linear das covariáveis, representado por η, e o parâmetro da distribuição da variável resposta, μ, por meio de uma função de ligação $g(\cdot)$. A forma geral do modelo é expressa pela Equação (3.8).

$$g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}, \qquad i = 1, \ldots, n \tag{3.8}$$

Esses modelos têm como característica principal a flexibilidade e adaptabilidade a diferentes tipos de distribuições da variável resposta, sendo a função de ligação o

componente responsável por essas adaptações. No presente estudo, tais modelos foram empregados tanto para classificação, utilizando a família binomial, quanto para regressão, utilizando a distribuição Gamma.

Para seleção das covariáveis foi utilizado o algoritmo de *stepwise*, com o objetivo de garantir simplicidade e parcimônia no modelo final. O procedimento é descrito a seguir:

1. O algoritmo é iniciado com um modelo contendo algumas variáveis;

2. Calcula-se o AIC do modelo atual;

3. Adiciona-se individualmente cada variável ainda não incluída no modelo e calcula-se o AIC de cada novo modelo;

4. Remove-se individualmente cada variável atualmente no modelo e calcula-se o AIC de cada novo modelo;

5. Compara-se o AIC de todos os modelos gerados nos passos **3** e **4**;

6. O modelo com menor AIC entre os testados é definido como novo modelo;

7. Os passos entre **2** e **6** se repetem até que nenhuma adição ou remoção de variável melhore o AIC do modelo atual.

O Critério de Informação de Akaike (AIC) penaliza modelos mais complexos (com mais variáveis), o que contribui para a obtenção de modelos mais simples, sem comprometer o poder preditivo. A equação do AIC está apresentada em (3.9), em que L representa a máxima verossimilhança do modelo e k o número de parâmetros estimados:

$$\text{AIC} = 2k - 2\ln(L) \tag{3.9}$$

3.2.6.1 Modelo de Regressão Binomial

O modelo de regressão Binomial, também conhecido como modelo de regressão logística, é um MLG utilizado em problemas de classificação, e temos que a resposta Y_i condicionada as variáveis explicativas, $\mathbf{x}_i$, segue a distribuição expressa em (3.10).

$$Y_i|\mathbf{x}_i \sim \text{Binomial}(m, \mu_i), \qquad i = 1, \ldots, n \tag{3.10}$$

Para realizar o ajuste foram consideradas duas formas para a função de ligação, a função logito (3.11) e a função probito (3.12).

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right), \tag{3.11}$$

$$g(\mu_i) = \Phi^{-1}(\mu_i) \tag{3.12}$$

3.2.6.2 Modelo de Regressão Gamma

O modelo de regressão Gamma é um MLG utilizado para modelar dados contínuos, positivos e assimétricos. Assim, ele foi empregado nos casos em que houve inflação de

zeros na variável resposta, sendo ajustado apenas aos casos em que o valor da resposta foi estritamente positivo. A distribuição condicional da variável resposta é expressa a seguir:

$$Y_i|\mathbf{x}_i \sim \text{Gamma}(\mu_i, \phi) \tag{3.13}$$

As funções de ligação utilizadas aqui foram a função logarítmica (3.14) e a função inversa (3.15).

$$g(\mu_i) = \log(\mu_i) \tag{3.14}$$

$$g(\mu_i) = \frac{1}{\mu_i} \tag{3.15}$$

3.2.7 Árvore de Decisão

Inicialmente, é necessário apresentar o método de árvore de decisão, o qual serve de base para modelos robustos como *Random Forest* e *XGBoost*. Árvores de decisão são algoritmos de aprendizado supervisionado, ou seja, utilizados quando se conhece o valor da variável resposta para cada observação. Esses modelos podem ser aplicados tanto a problemas de regressão quanto de classificação, sendo originalmente propostos por Morgan e Sonquist (1963) e formalizados por Breiman et al. (1984).

No contexto da classificação, o objetivo das árvores de decisão é particionar o espaço das covariáveis em regiões que sejam o mais homogêneas possível em relação às classes da variável resposta. Conforme descrito por Hastie, Tibshirani e Friedman (2009), uma árvore de decisão define uma função preditiva da forma:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m), \tag{3.16}$$

em que c_m representa a classe majoritária (ou a probabilidade estimada de uma classe) dentro da região R_m, e $R_1, R_2, \ldots, R_M$ são regiões disjuntas do espaço das covariáveis. Cada observação $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ pertence a uma dessas regiões, definidas recursivamente por divisões binárias nas covariáveis.

Para definir essas regiões, dado um atributo j e um ponto de divisão s, utiliza-se a seguinte regra de particionamento:

$$R_1(j, s) = \{X \mid X_j \leq s\} \quad \text{e} \quad R_2(j, s) = \{X \mid X_j > s\}. \tag{3.17}$$

Cada divisão da árvore é chamada de nó, e as regiões finais, nas quais se define a predição da classe, são chamadas de folhas. O critério de divisão mais comum em problemas de classificação é a redução da impureza da partição, medida por índices como a entropia ou o índice de Gini. Este último é definido pela seguinte equação:

$$G = \sum_{k=1}^{K} p_k(1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2, \tag{3.18}$$

em que p_k representa a proporção de observações da classe k em um determinado nó. Quanto menor o valor de G, maior a pureza do nó.

No contexto de regressão, o funcionamento do modelo é semelhante, com a principal alteração sendo a natureza contínua da variável resposta y. Nesse caso, não se utiliza o índice de Gini para medir impureza, e sim a Soma dos Quadrados dos Resíduos (RSS), definida em (3.19):

$$RSS = \sum_{i \in R_m} (y_i - \bar{y}_{R_m})^2, \tag{3.19}$$

em que y_i representa os valores observados da variável resposta e $\bar{y}_{R_m}$ é a média desses valores na região R_m.

Com essa medida, observa-se que o melhor valor preditivo $\hat{c}_m$ para minimizar o RSS em cada região é justamente a média dos valores observados da variável resposta dentro da região R_m:

$$\hat{c}_m = \bar{y}_{R_m}. \tag{3.20}$$

Embora modelos de árvore sejam intuitivos e úteis, apresentam limitações importantes, como tendência ao *overfitting*, alta variabilidade e desempenho insatisfatório em contextos mais complexos (BREIMAN et al., 1984; QUINLAN, 1996; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Para contornar essas limitações, surgiram métodos baseados em conjuntos de árvores, como o *Random Forest* e o *XGBoost*, que aprimoram tanto o desempenho preditivo quanto a estabilidade dos modelos.

3.2.7.1 Random Forest

O modelo de *Random Forest*, ou Floresta Aleatória, foi introduzido por Breiman (2001). Trata-se de um algoritmo baseado em múltiplas árvores de decisão que utiliza o princípio do *ensemble learning*, uma técnica que combina diversos modelos com o objetivo de alcançar um desempenho superior ao obtido por modelos individuais.

A ideia central do *Random Forest* é construir diversas árvores de decisão a partir de subconjuntos aleatórios do conjunto de dados, utilizando o método de amostragem com reposição conhecido como *bagging* (abreviação de Bootstrap Aggregation), proposto por Breiman (1996). Em cada iteração, uma nova árvore é treinada sobre um subconjunto de observações, e, adicionalmente, em cada nó da árvore, apenas um subconjunto aleatório das covariáveis é considerado para a divisão. Esse processo reduz a correlação entre as árvores e aumenta a diversidade do conjunto.

No contexto de classificação, ao final do processo, cada árvore gera uma predição categórica, e o *Random Forest* pode operar de duas formas. A primeira consiste na predição por voto majoritário, em que a classe final é definida como aquela que foi mais frequentemente atribuída entre todas as árvores. A segunda consiste na estimativa da probabilidade de cada classe, obtida a partir da proporção de árvores que atribuíram determinada classe à observação. Esta abordagem probabilística permite aplicar limiares de decisão (*thresholds*) customizados, como foi feito no presente estudo, em que o threshold foi ajustado para maximizar a acurácia.

A probabilidade estimada de que uma observação pertença à classe 1 é dada por:

$$\hat{P}(y = 1 \mid x) = \frac{1}{B} \sum_{b=1}^{B} I(\hat{y}^b = 1) \tag{3.21}$$

em que B representa o número total de árvores do modelo, e $I(\hat{y}^b = 1)$ é uma função indicadora que vale 1 quando a árvore b classifica a observação na classe 1, e 0 caso contrário.

Para o contexto de regressão, a previsão final corresponde à média das previsões de todas as árvores do modelo. Representada por:

$$\hat{y} = \sum_{b=1}^{B} \frac{\hat{y}^b}{B} \tag{3.22}$$

Esse método corrige as principais limitações das árvores de decisão isoladas. Graças à aleatorização, cada árvore possui alta variância individual, mas o conjunto apresenta baixo viés e maior robustez, resultando em modelos menos suscetíveis ao *overfitting* e com melhor desempenho preditivo em contextos mais complexos (JAMES et al., 2013).

O algoritmo de *Random Forest* depende de alguns hiperparâmetros que podem influenciar significativamente seu desempenho. Para obter os melhores resultados, foram aplicadas técnicas de otimização conhecidas como *tuning* de hiperparâmetros. Dentre as diversas abordagens existentes, a técnica adotada neste trabalho foi a de *grid search*. Nessa estratégia, são definidos previamente alguns valores para cada hiperparâmetro, e todas as combinações possíveis entre esses valores são testadas, gerando um modelo para cada configuração. O modelo com os melhores resultados de avaliação é então selecionado (LIASHCHYNSKYI; LIASHCHYNSKYI, 2019).

A descrição dos hiperparâmetros utilizados, bem como os valores estipulados para cada um deles, está apresentada na Tabela 1 (PROBST; WRIGHT; BOULESTEIX, 2019).

Para a seleção de variáveis, foi calculada a importância relativa de cada variável explicativa com base na redução média da impureza (*Mean Decrease in Impurity* - MDI) (BREIMAN, 2001). Esse algoritmo contabiliza quantas vezes cada variável é utilizada para dividir os nós das árvores no modelo. Para cada uma dessas divisões, é computada a redução na impureza, no caso da classificação, utilizando o Índice de Gini, conforme apresentado em (3.18). Essas reduções são somadas ao longo de todas as árvores da floresta.

Tabela 1 – Hiperparâmetros do *Random Forest*, suas respectivas descrições e valores testados.

Hiperparâmetro	Descrição	Valores
mtry	Proporção de variáveis candidatas sorteadas aleatoriamente em cada divisão.	5%, 15%, 25%, 33% e 40%
sample.fraction	Proporção de observações sorteadas para cada árvore.	50%, 63%, 80% e 90%
replace	Define se as observações são sorteadas com ou sem reposição.	TRUE (com reposição) e FALSE (sem reposição)
nodesize	Número mínimo de observações em um nó terminal.	1, 3, 5 e 10
num.trees	Número de árvores na floresta.	750 e 1000

O valor final de importância relativa, atribuído a cada covariável, é uma medida agregada de sua contribuição para tornar os nós mais puros, ou seja, para aumentar o desempenho preditivo local do modelo.

Após esse cálculo inicial, foi aplicado o algoritmo de *Recursive Feature Elimination* (RFE), originalmente apresentado por Guyon et al. (2002) em um modelo de *Support Vector Machines*, e aqui adaptado para o *Random Forest*. O procedimento é composto pelas seguintes etapas:

1. Treinamento do modelo com o conjunto completo de variáveis;

2. Cálculo das importâncias relativas das covariáveis;

3. Eliminação da variável com menor importância;

4. Reajuste do modelo e cálculo das métricas de avaliação;

5. Repetição do processo até atingir o número mínimo de variáveis.

Ao final, com todas as métricas computadas para os diferentes subconjuntos de variáveis, é selecionado o modelo que apresenta o melhor equilíbrio entre parcimônia e capacidade preditiva.

3.2.7.2 Grandient Boosting

O algoritmo de *Gradient Boosting* foi introduzido por Friedman (2001). A ideia central é construir modelos aditivos compostos por diversas árvores de decisão que, individualmente, são fracas, mas que, combinadas iterativamente, aprendem com os erros dos modelos anteriores. A atualização do modelo ocorre na direção do gradiente da função de perda, otimizando a predição a cada nova iteração.

Seja $L(y, \gamma)$ a função de perda do modelo, o procedimento segue os seguintes passos:

1. Inicializa-se o modelo com a constante $f_0(x)$, que minimiza a função de perda:

$$f_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma). \tag{3.23}$$

2. Para m variando de 1 até M (número total de iterações):

   a) Calcular os pseudo-resíduos com base no gradiente da perda:

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}. \tag{3.24}$$

b) Ajustar uma nova árvore de decisão usando r_{im} como variável resposta. Essa árvore define as regiões R_{jm}, onde $j = 1, 2, \ldots, J_m$.

c) Para cada folha R_{jm}, calcular o valor ótimo a ser atribuído:

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma). \tag{3.25}$$

d) Atualizar o modelo:

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}). \tag{3.26}$$

A predição final é dada por $\hat{f}(x) = f_M(x)$. No caso de classificação, essa predição é transformada em probabilidade por meio da função logística:

$$\hat{P}(y = 1 \mid x) = \frac{1}{1 + e^{-\hat{f}(x)}}. \tag{3.27}$$

Modelos de boosting podem proporcionar ganhos significativos de desempenho em relação a árvores simples, embora apresentem maior custo computacional. Para mitigar esse custo, surgiram métodos mais eficientes como o *XGBoost*, que será abordado a seguir (CHEN; GUESTRIN, 2016; KE et al., 2017).

3.2.7.3 XGBoost

O *XGBoost* (*Extreme Gradient Boosting*) foi introduzido por Chen e Guestrin (2016) e é uma variação otimizada do algoritmo de *Gradient Boosting*, incorporando técnicas de regularização, paralelização e poda eficiente de árvores. Ele foi projetado para alcançar alto desempenho em tarefas supervisionadas, incluindo problemas de classificação binária.

A cada iteração t, o *XGBoost* busca minimizar a seguinte função objetivo:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), \tag{3.28}$$

em que $L(y, \hat{y})$ representa a função de perda, $f_t(x)$ é a nova árvore ajustada na iteração t, e $\Omega(f_t)$ é o termo de regularização que penaliza a complexidade da árvore.

Como essa função não pode ser minimizada diretamente de forma eficiente, utiliza-se uma expansão de Taylor de segunda ordem, resultando na aproximação:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t), \tag{3.29}$$

em que:

Tabela 2 – Hiperparâmetros do *XGBoost*, suas respectivas descrições e intervalos utilizados no espaço de busca.

Hiperparâmetro	Descrição	Intervalo
max_depth	Profundidade máxima da árvore.	2 a 10
eta	Taxa de aprendizado (learning rate).	0,01 a 0,3
nrounds	Número de iterações (boosting rounds).	50 a 200
subsample	Proporção de amostras usadas por árvore.	0,5 a 1
colsample_bytree	Proporção de variáveis usadas por árvore.	0,5 a 1

- $g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}$ é o gradiente da função de perda;
- $h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$ é a hessiana (segunda derivada).

Essa estrutura permite atualizações mais informadas e estáveis, contribuindo para a robustez e eficiência do modelo. Ao final, a saída do modelo também pode ser convertida em uma probabilidade estimada de classe por meio da função logística, como no *Gradient Boosting* padrão.

Assim como no *Random Forest*, a predição com o *XGBoost* também depende da definição de hiperparâmetros que influenciam diretamente o desempenho do modelo. O *tuning* desses parâmetros seguiu a técnica de *random search* (LIASHCHYNSKYI; LIASHCHYNSKYI, 2019), que se assemelha ao *grid search*, com a diferença de que não são geradas todas as combinações possíveis dos hiperparâmetros. Em vez disso, a cada modelo treinado, realiza-se um sorteio aleatório de valores para os hiperparâmetros, dentro de intervalos previamente definidos. A descrição de cada hiperparâmetro e os respectivos valores testados são apresentados na Tabela 2, seguindo a descrição de Chen e Guestrin (2016).

O *Gain*, métrica utilizada para avaliação da importância das variáveis no *XGBoost*, é definido como a redução na função de perda regularizada provocada por uma divisão (*split*). Formalmente, o ganho ao realizar um *split* é dado por:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda}\right]$$

onde:

- G_L e G_R são os gradientes acumulados nos nós esquerdo e direito, respectivamente;
- H_L e H_R são os hessianos acumulados nos mesmos nós;
- $G = G_L + G_R$, $H = H_L + H_R$;
- λ é o parâmetro de regularização L2.

Essa métrica quantifica a contribuição de cada variável na melhoria da função objetivo. Assim como na métrica de importância por impureza utilizada no *Random Forest*, o *Gain* permite a ordenação das variáveis explicativas com base em sua relevância para o

modelo. Dessa forma, é possível aplicar o algoritmo de eliminação recursiva de variáveis (RFE) no *XGBoost* da mesma maneira, promovendo a seleção progressiva das variáveis mais informativas.

4 Resultados e Discussão

Nesta seção são apresentados os resultados da análise do conjunto de dados. É feito uma análise descritiva e o ajuste dos modelos para cada posição separadamente, considerando em cada subseção o fato das observações serem de calouros ou jogadores veteranos da NFL.

4.1 Quarterbacks

Os *Quarterbacks* são os responsáveis por conduzir os ataques no futebol americano, sendo encarregados de passar a bola para seus companheiros. Dessa forma, a maioria das jogadas ofensivas passam por suas mãos, o que faz com que essa posição seja amplamente reconhecida como a mais importante do esporte (BROOKS, 2015). No *Fantasy Football*, cada time titular conta com apenas um *Quarterback*, o que torna essencial contar com um atleta confiável e com bom desempenho em termos de pontuação.

No entanto, a modelagem estatística para essa posição apresenta desafios específicos. Há uma quantidade limitada de dados, tanto pelo número reduzido de *Quarterbacks* calouros que ingressam na NFL anualmente, quanto pelo fato de existirem apenas 32 vagas de titulares na NFL. Além disso, os *Quarterbacks* reservas raramente entram em campo, o que resulta em um conjunto de dados com grande proporção de pontuações muito baixas ou até mesmo zeradas.

4.1.1 Calouros

Para os calouros, foram como covariáveis utilizadas as estatísticas de passe e corrida em suas carreiras universitárias, características fisícas em seu primeiro ano na NFL, e o contexto em que ele estava inserido no *College*, além das variáveis a respeito do seu processo no *Draft* da NFL. Todas essas variáveis serão melhor exploradas na análise descritiva, antes de seguir para a modelagem.

Por ser a posição mais importante do futebol americano, *Quarterbacks* vindos do universitário frequentemente enfrentam dificuldades de adaptação em sua primeira temporada. Em razão disso, é comum que os times optem por mantê-los na reserva, permitindo seu desenvolvimento até estarem prontos para assumir a titularidade. Esse fenômeno pode ser observado na Figura 1, que apresenta a distribuição da variável resposta para *Quarterbacks* calouros: cerca de 57,6% dos jogadores não pontuam em seu ano de estreia, o que representa um desafio adicional para a modelagem.

Diante desse cenário, optou-se por utilizar um modelo em duas etapas. Primeiramente, foi aplicado um método de classificação para prever se o jogador pontuará ou não na temporada. Em seguida, para os jogadores que pontuaram, foi ajustado um modelo de regressão para estimar a pontuação esperada.

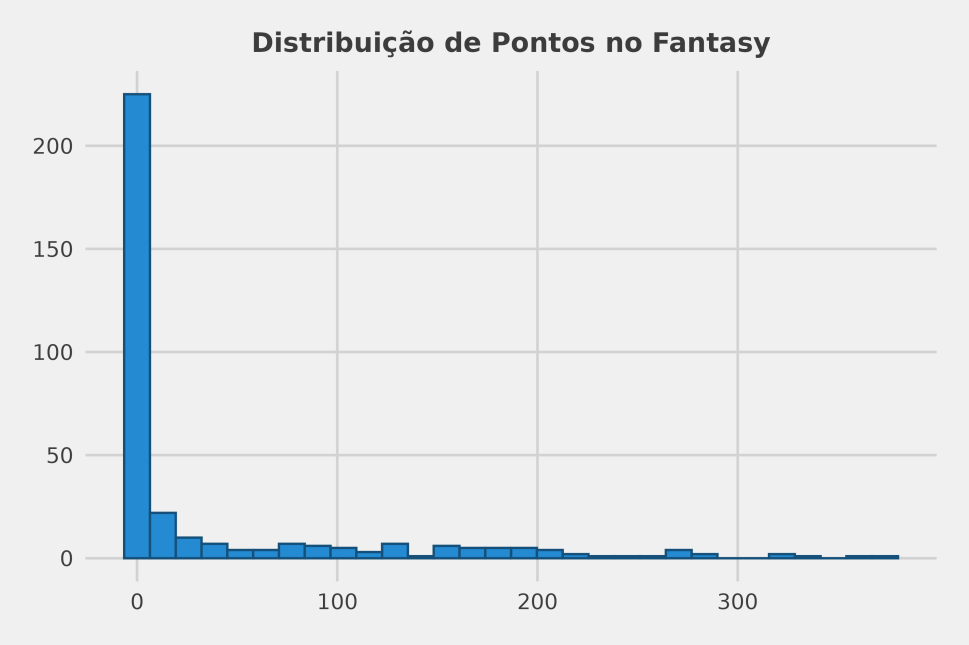

Figura 1 – Distribuição da pontuação no *Fantasy* para *Quarterbacks* em seu primeiro ano na NFL.

O desempenho de um calouro parece estar fortemente relacionado às expectativas do time em que ele está inserido. Na Figura 2, observa-se que jogadores selecionados nas rodadas iniciais do *Draft* (gráfico à esquerda) recebem significativamente mais oportunidades de jogo em suas temporadas de estreia, enquanto aqueles escolhidos em rodadas intermediárias, finais ou que não foram "draftados" tendem a permanecer como reservas no início da carreira, obtendo assim pontuações mais baixas que os demais. Em relação à conferência em que o atleta atuou no universitário (gráfico à direita), nota-se uma diferença considerável, de modo que jogadores oriundos de conferências menores apresentam menor tempo de jogo e pontuação em comparação àqueles que competiram nas conferências *Power 5*.

Características físicas também podem influenciar o desempenho de um jogador. Observou-se que atletas mais jovens tendem a apresentar pontuações superiores em comparação àqueles que ingressam na NFL em idade mais avançada. No entanto, essa relação pode estar associada à rodada em que o jogador é selecionado no *Draft* da NFL, uma vez que atletas de destaque no futebol universitário costumam se declarar mais cedo para o *Draft* e, devido a esse destaque, são escolhidos nas primeiras rodadas.

Jogadores mais baixos e leves podem ter dificuldades para lidar com a intensidade física da liga profissional, enquanto atletas muito altos e pesados podem enfrentar limitações de mobilidade e velocidade, o que também pode comprometer seu desempenho. Essas relações estão ilustradas na Figura 3: na linha superior, observa-se a relação entre idade e pontuação, enquanto na linha inferior, à esquerda, é apresentada a distribuição da pontuação por altura (em polegadas), e à direita, por peso (em libras).

Por fim, outro fator relevante é o desempenho do atleta durante sua trajetória no

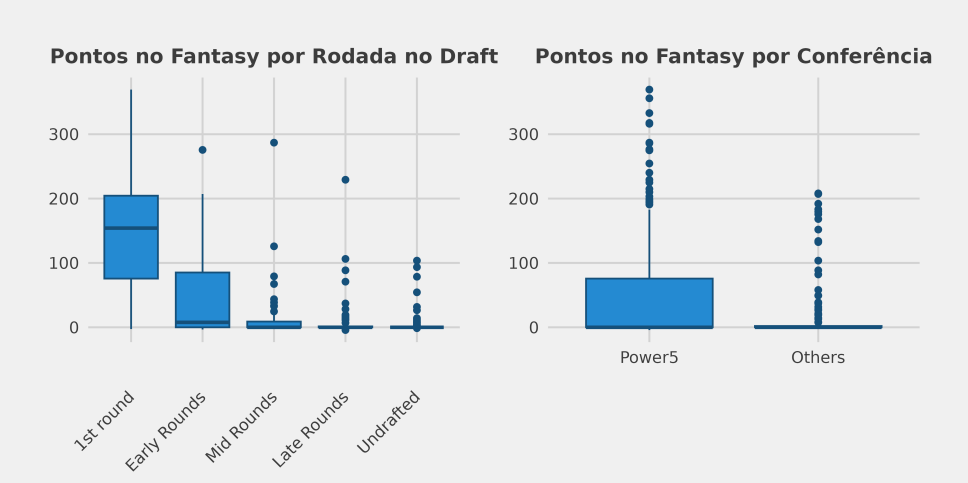

Figura 2 – Distribuição da pontuação no *Fantasy* por rodada em que o *Quarterback* foi selecionado no *Draft* da NFL, e de acordo com a conferência do *College Football* ele atuou.

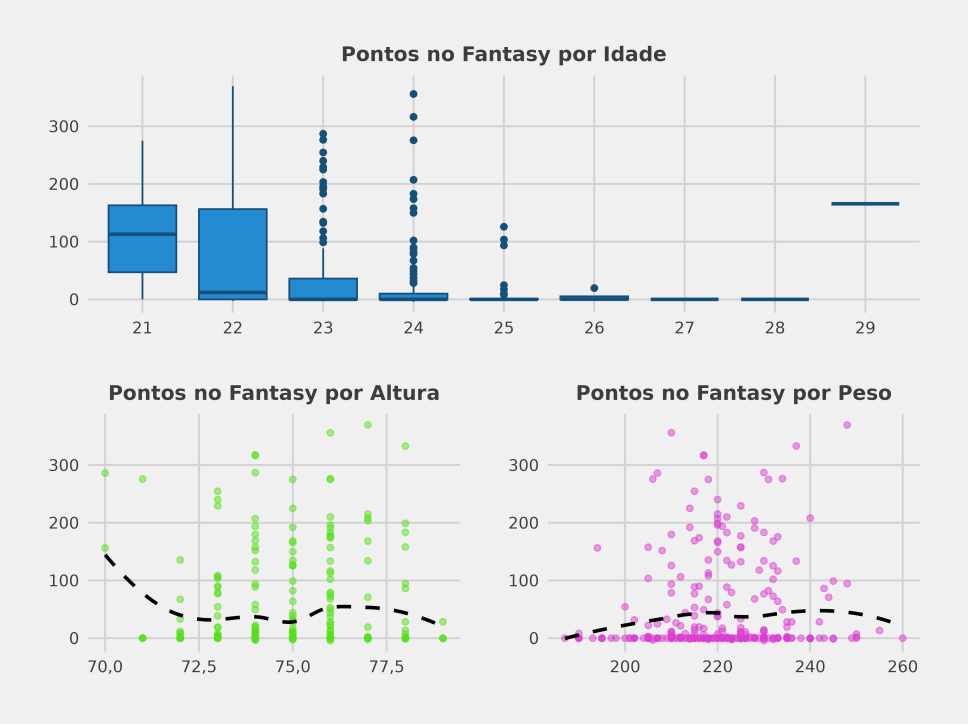

Figura 3 – Distribuição da pontuação no *Fantasy* por características pessoais dos QBs calouros.

Tabela 3 – Análise de correlação para as variáveis de passe para os QBs calouros.

	passing_completions	passing_att	passing_yds	passing_td	passing_int
passing_completions	1,00	0,99	0,98	0,90	0,80
passing_att	0,99	1,00	0,97	0,87	0,85
passing_yds	0,98	0,97	1,00	0,94	0,77
passing_td	0,90	0,87	0,94	1,00	0,63
passing_int	0,80	0,85	0,77	0,63	1,00

Tabela 4 – Análise de correlação para as variáveis de corrida para os QBs calouros.

	rushing_car	rushing_yds	rushing_td
rushing_car	1,00	0,87	0,88
rushing_yds	0,87	1,00	0,90
rushing_td	0,88	0,90	1,00

futebol universitário. No entanto, muitas das variáveis disponíveis apresentam características semelhantes entre si. Por exemplo, se um *Quarterback* realiza um grande número de tentativas de passe, é esperado que ele também acumule mais jardas aéreas, o que indica uma alta correlação entre essas variáveis.

Dessa forma, as Tabelas 3 e 4 apresentam as correlações entre as estatísticas universitárias relacionadas ao jogo aéreo e ao jogo terrestre, respectivamente, permitindo uma análise mais aprofundada desse impacto. Variáveis com correlação superior a 0,90 foram removidas para evitar problemas de multicolinearidade, sendo selecionadas para a modelagem as seguintes: `passing_td`, `passing_int`, `rushing_car`, `rushing_yds` e `rushing_td`.

4.1.1.1 Classificação

O modelo de classificação teve como objetivo identificar se um *Quarterback* pontuaria (resposta = 1) ou não (resposta = 0) em sua temporada de estreia. Para isso, os dados foram divididos em um conjunto de treino (70%) e outro de teste (30%). Foram utilizadas as variáveis de passe e corrida selecionadas na análise de correlação, com a adição das variáveis `draft_round_group`, `conference_group`, `age`, `height` e `weight`.

Três métodos diferentes de modelagem foram aplicados: regressão logística (com funções de ligação logito e probito), *Random Forest* e *XGBoost*. Também foi realizada uma seleção de covariáveis, com o objetivo de eliminar aquelas que não contribuíam de forma significativa para o desempenho preditivo dos modelos. A escolha do modelo final baseou-se no equilíbrio entre parcimônia e na acurácia dos valores preditos. Os resultados obtidos por cada abordagem estão apresentados na Tabela 5.

O modelo selecionado foi o *XGBoost*, com um total de cinco variáveis explicativas. Esse número foi determinado a partir da aplicação do método de RFE, realizando a remoção sequencial das variáveis menos relevantes e avaliando a acurácia a cada etapa.

Tabela 5 – Métricas de avaliação para os modelos candidatos para a classificação de QBs calouros.

Modelo	Acurácia	N° de Variáveis
modelo_log_logito	0,7689	15
modelo_log_logito_stepwise	0,7813	6
modelo_log_probito	0,7771	15
modelo_log_probito_stepwise	0,7813	6
modelo_rf	0,7858	17
modelo_rf_rfe	0,7970	12
modelo_xg	0,8503	17
modelo_xg_rfe	**0,8173**	**5**

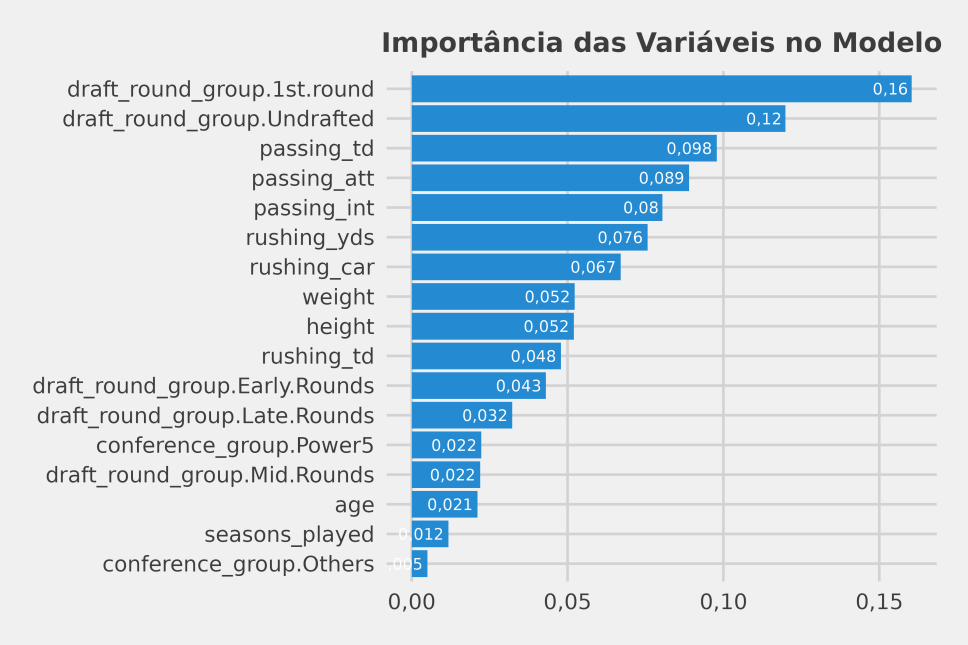

Figura 4 – Importância relativa pela métrica *Gain* das variáveis explicativas do modelo de classificação *XGBoost* para os QBs calouros.

As importâncias relativas pela métrica *Gain* é apresentada na Figura 4, e na sequência, o desempenho do modelo ao longo do processo de RFE, ilustrado na Figura 5.

As cinco variáveis selecionadas foram: `passing_td`, `passing_att`, `passing_int` e dois níveis da variável `draft_round_group`: `1st round` e `Undrafted`.

Os valores otimizados para os hiperparâmetros do modelo estão apresentados na Tabela 6.

Na matriz de confusão, Tabela 7, podemos observar detalhadamente o desempenho preditivo do modelo. Dentre os *Quarterbacks* que pontuaram, o modelo foi capaz de classificar corretamente 68,3% dos atletas, a taxa de acerto entre os positivos (sensibilidade). Já entre os que não pontuaram, 95,16% foram corretamente identificados, refletindo a especificidade do modelo. A acurácia geral foi de 84,5%.

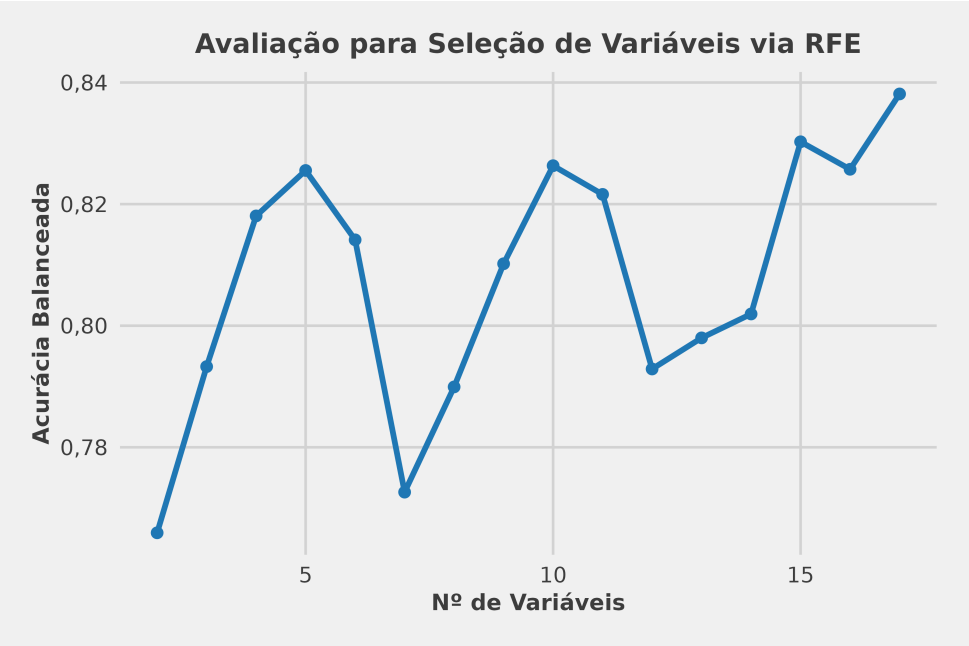

Figura 5 – Acurácia do modelo em cada passo do RFE do modelo de classificação de QBs calouros.

Tabela 6 – Hiperparâmetros do modelo de classificação *XGBoost* de QBs calouros.

Hiperparâmetro	Valor
nrounds	57,00
verbose	0,00
max_depth	10,00
eta	0,05
subsample	0,80
colsample_bytree	0,66

Tabela 7 – Matriz de confusão para o modelo de classificação de QBs calouros.

Predito / Real	Negativo	Positivo
Negativo	59	13
Positivo	3	28

Tabela 8 – Métricas de avaliação para os modelos candidatos para a regressão de QBs calouros.

Modelo	MAE	N° de Variáveis
modelo_gamma_log	74,50	15
modelo_gamma_log_stepwise	**57,65**	**6**
modelo_gamma_inverse	73,71	15
modelo_gamma_inverse_stepwise	57,90	4
modelo_rf1	63,61	17
modelo_rf_rfe	62,85	10
modelo_xg1	65,58	17
modelo_xg_rfe	60,84	2

4.1.1.2 Regressão

Para o modelo de regressão, foram utilizados os mesmos dados do modelo de classificação, com duas alterações principais. A primeira refere-se à variável resposta, que passa a representar, de fato, o valor da pontuação no *Fantasy Football*, a segunda consiste em um filtro que mantém apenas as observações em que o jogador efetivamente pontuou.

Os métodos utilizados foram: MLG Gamma com funções de ligação logarítmica e inversa, *Random Forest* e *XGBoost*. Foi conduzida uma seleção das covariáveis com o objetivo de simplificar o modelo sem perder capacidade preditiva. A escolha do modelo final baseou-se no equilíbrio entre parcimônia e o MAE de cada modelo. Os resultados obtidos por cada abordagem estão apresentados na Tabela 8.

O modelo com melhor desempenho foi o modelo Gamma com função de ligação logarítmica, ajustado com três variáveis explicativas: `draft_round_group`, `passing_att` e `passing_int`. Importante notar, que dentre as variáveis de estatísticas da carreira universitária, somente aquelas de passe foram consideras significativas, e nenhuma relacionada ao jogo terrestre. A equação do modelo está apresentada em (4.1), e o resumo dos parâmetros estimados encontra-se na Tabela 9.

$$\begin{aligned}\hat{\mathbb{E}}(Y_i) = \exp\Big(&4{,}8447 - 0{,}9801 \cdot I(\text{draft\_round\_group} = \text{'Early Rounds'}) \\ &- 1{,}3835 \cdot I(\text{draft\_round\_group} = \text{'Mid Rounds'}) \\ &- 1{,}5287 \cdot I(\text{draft\_round\_group} = \text{'Late Rounds'}) \\ &- 1{,}8895 \cdot I(\text{draft\_round\_group} = \text{'Undrafted'}) \\ &+ 0{,}00136 \cdot \text{passing\_att} - 0{,}0453 \cdot \text{passing\_int}\Big)\end{aligned} \tag{4.1}$$

Os resultados do modelo indicam que todos os níveis da variável `draft_round_group` apresentaram médias esperadas significativamente inferiores em comparação ao grupo de referência, formado por jogadores do nível `1st Round`. Fixando os valores das demais variáveis do modelo (`passing_att` e `passing_int`), é esperado que jogadores de `Early`

Tabela 9 – Resumo dos parâmetros estimados do modelo de regressão de QBs calouros.

Parâmetro	Estimativa	Erro Padrão	P-Valor
Intercepto	4,8447	0,2877	0,0000
draft_round_group = "Early Rounds"	-0,9801	0,2856	0,0009
draft_round_group = "Mid Rounds"	-1,3835	0,3056	0,0000
draft_round_group = "Late Rounds"	-1,5287	0,3529	0,0000
draft_round_group = "Undrafted"	-1,8895	0,3124	0,0000
passing_att	0,0014	0,0005	0,0046
passing_int	-0,0453	0,0182	0,0145

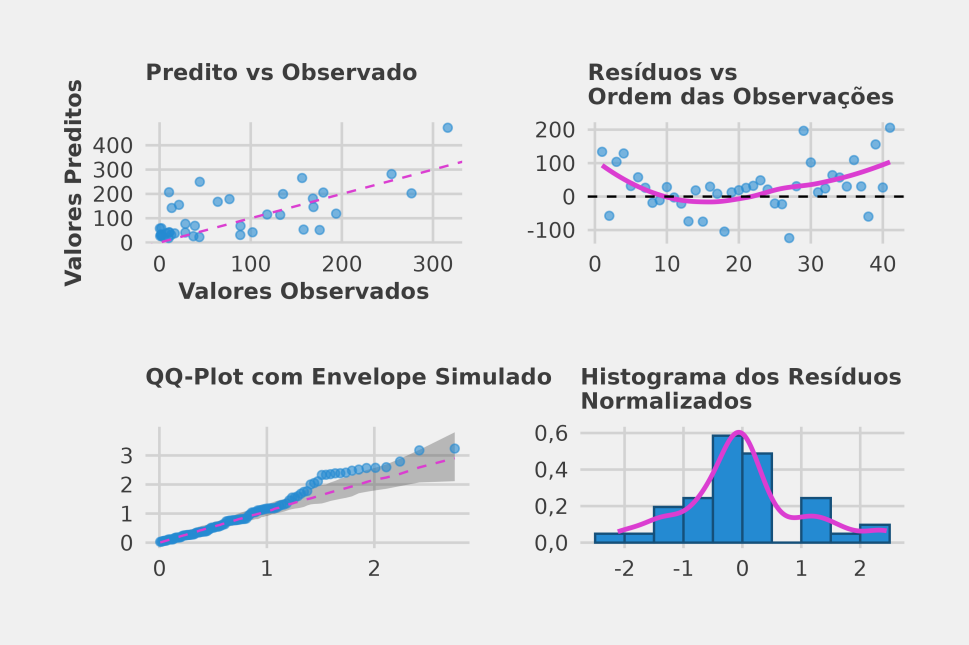

Figura 6 – Gráficos de avaliação dos resíduos no modelo de regressão dos QBs calouros.

`Rounds` tenham uma pontuação de 62,5% ($(1 - e^{-0,9801} \approx 0,625)$) menor em relação a jogadores de `1st Round`. Para jogadores de `Late Rounds`, a queda esperada é de 78,3% $(1 - e^{-1,5287} \approx 0,783)$. Por fim, jogadores `Undrafted` apresentam um valor esperado de cerca de 84,9% $(1 - e^{-1,8895} \approx 0,849)$ inferior em relação ao grupo de referência.

Para a variável `passing_att`, fixando as demais variáveis, temos que a cada jarda passada pelo jogador, espera-se que a resposta aumente em 0,14% ($e^{0,0014} \approx 1,0014$), já para `passing_int`, espera-se que a cada interceptação, a resposta decresça em 4,4% ($e^{-0,0453} \approx 0,956$).

A análise de resíduos, neste caso, pode ter sua eficácia comprometida devido ao tamanho reduzido da base de teste. Observou-se um desvio em relação ao comportamento esperado dos resíduos, conforme ilustrado na Figura 6. No gráfico QQ-Plot com envelopes simulados, nota-se que, para valores mais altos, os resíduos extrapolam as bandas de confiança, indicando possíveis inconsistências. Os principais *outliers* identificados são apresentados na Tabela 10.

Tabela 10 – Principais *outliers* presentes no modelo de regressão de QBs calouros.

Jogador	Pontuação Observada	Pontuação Predita	Resíduo Absoluto
Michael Penix Jr.	44,10	249,75	205,65
Patrick Mahomes	10,36	206,84	196,48
Bo Nix	316,20	472,06	155,86
Alex Smith	21,30	155,03	133,73
JaMarcus Russell	13,32	142,19	128,87

Entre os casos mais notáveis estão os jogadores *Michael Penix Jr.* (Pro Football Reference, 2025p), *Patrick Mahomes* (Pro Football Reference, 2025r) e *Jamarcus Russell* (Pro Football Reference, 2025i), todos selecionados entre as dez primeiras escolhas no *Draft* da NFL em seus respectivos anos, e com histórico de destaque no futebol universitário. No entanto, esses atletas atuaram como reservas durante a maior parte da temporada de estreia, entrando em campo apenas em alguns jogos. Como resultado, embora tenham pontuado, seus desempenhos foram significativamente inferiores ao previsto.

No caso do atleta *Michael Penix Jr.*, outro fator relevante diz respeito à sua longa permanência na liga universitária, realidade também observada com *Bo Nix* (Pro Football Reference, 2025e). Enquanto *Penix* atuou por seis temporadas no *College Football*, *Nix* participou por cinco. Esse tempo estendido permitiu que acumulassem estatísticas expressivas, o que pode ter supervalorizado suas predições de desempenho na NFL.

Outro exemplo é o jogador *Alex Smith* (Pro Football Reference, 2025s), que foi selecionado como a primeira escolha geral em seu ano, mas que apresentou um desempenho bem abaixo das expectativas, e sofreu com lesões em seu primeiro ano.

4.1.2 Veteranos

Para o caso de QBs veteranos, também temos uma proporção significativa de jogadores que são reservas durante toda a temporada e por isso não acumulam pontuação no *Fantasy*, a distribuição da variável resposta pode ser apreciada na Figura 7. Temos 27,1% das observações com a pontuação zerada, com isso, também será necessário a inclusão de um modelo de classificação antes do modelo de regressão.

Como variáveis explicativas, foram utilizadas estatísticas semelhantes às empregadas nos modelos para *Quarterbacks* calouros, porém adaptadas ao contexto da NFL. Em vez de considerar o desempenho dos atletas no *College Football*, utilizaram-se as estatísticas de passe e corrida referentes à temporada anterior na NFL. Além disso, foram incluídas variáveis relacionadas às características físicas dos jogadores, como idade, altura e peso, bem como os anos de experiência na liga profissional.

Na Figura 8, observamos, no gráfico superior, a distribuição da pontuação no *Fantasy Football* de acordo com a faixa etária dos jogadores. No grupo de 20 a 24 anos, composto majoritariamente por atletas em início de carreira, nota-se maior variabilidade

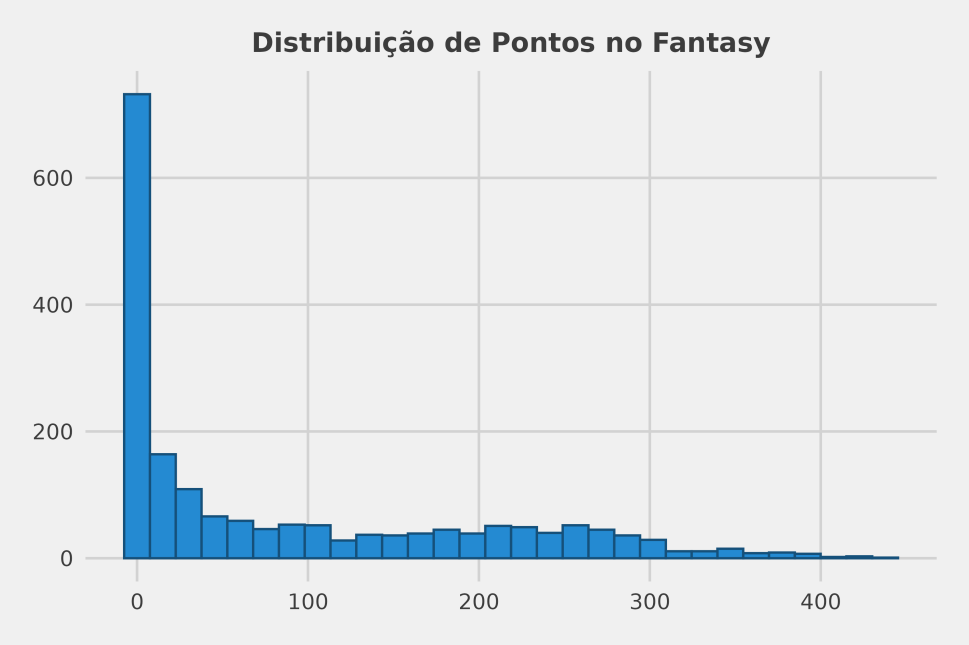

Figura 7 – Distribuição da pontuação no *Fantasy* para QBs veteranos da NFL.

Tabela 11 – Análise de correlação para as variáveis de passe para os QBs veteranos.

	passing_completions	passing_att	passing_yds	passing_td	passing_int
passing_completions	1,00	1,00	0,99	0,95	0,86
passing_att	1,00	1,00	0,99	0,93	0,89
passing_yds	0,99	0,99	1,00	0,96	0,86
passing_td	0,95	0,93	0,96	1,00	0,78
passing_int	0,86	0,89	0,86	0,78	1,00

na distribuição, com muitos jogadores concentrados entre 100 e 300 pontos. Na faixa de 25 a 28 anos, há uma concentração de pontuações elevadas, embora também se verifique uma proporção significativa de desempenhos muito baixos. A partir dos 29 anos, a distribuição torna-se mais homogênea, com a maioria dos jogadores registrando entre 200 e 300 pontos, embora com menos variabilidade, o que pode estar associado à maior experiência ou a papéis mais definidos na equipe.

Nos gráficos inferiores, à esquerda, observa-se a relação entre altura (em polegadas) e pontuação. Embora haja picos de desempenho em diferentes alturas, destaca-se uma maior concentração de pontuações entre 75 e 77 polegadas. Em alturas muito baixas, o número de observações é reduzido, dificultando generalizações. À direita, o gráfico relaciona peso (em libras) com pontuação. Nota-se uma tendência crescente até aproximadamente 250 libras, a partir desse ponto, observa-se uma queda na pontuação esperada, indicando um possível efeito de excesso de peso no desempenho.

Para a análise de correlação entre as variáveis explicativas, temos as tabelas 11 para as estatísticas de passe e 12 para as estatísticas de corrida. As variáveis filtradas para utilizar na modelagem foram: `passing_td`, `passing_int`, `rushing_yds` e `rushing_td`.

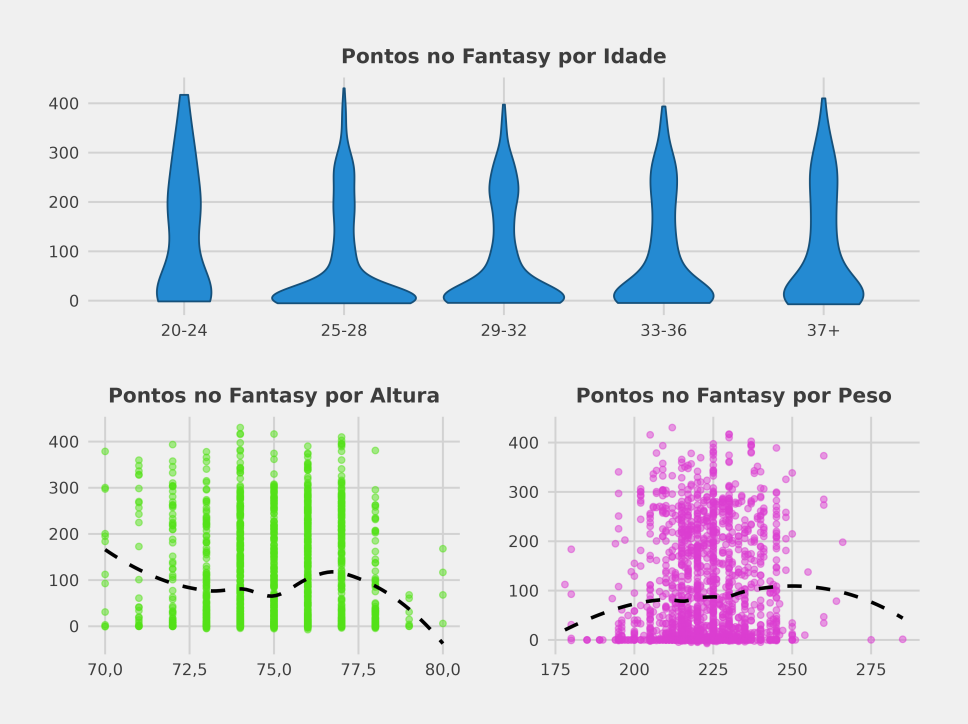

Figura 8 – Distribuição na pontuação no *Fantasy* para as características pessoais dos QBs veteranos da NFL.

Tabela 12 – Análise de correlação para as variáveis de corrida para os QBs veteranos.

	rushing_car	rushing_yds	rushing_td
rushing_car	1,00	0,94	0,77
rushing_yds	0,94	1,00	0,76
rushing_td	0,77	0,76	1,00

4.1.2.1 Classificação

O modelo de classificação seguiu a mesma metodologia aplicada aos *Quarterbacks* calouros. As variáveis utilizadas incluíram aquelas selecionadas na análise de correlação, acrescidas de `sacks_suffered`, `games`, `age`, `height`, `weight` e `years_exp`.

Os resultados de cada abordagem estão apresentados na Tabela 13. O modelo selecionado foi o MLG Binomial com função de ligação probito, com 7 variáveis explicativas, que são: `age`, `years_exp`, `games`, `passing_td`, `passing_int`, `sacks_suffered` e `rushing_yds`. A equação do modelo é apresentada em (4.2), e os parâmetros estimados são apresentados na Tabela 14.

$$\begin{aligned}\hat{\mathbb{P}}(Y_i = 1) = \Phi\Big(&1{,}6469 - 0{,}0749 \cdot \text{age} + 0{,}0898 \cdot \text{years\_exp} \\ &+ 0{,}1140 \cdot \text{games} + 0{,}0543 \cdot \text{passing\_td} \\ &- 0{,}0503 \cdot \text{passing\_int} - 0{,}0135 \cdot \text{sacks\_suffered} \\ &+ 0{,}0014 \cdot \text{rushing\_yds}\Big)\end{aligned} \tag{4.2}$$

Tabela 13 – Métricas de avaliação para os modelos candidatos para a classificação de QBs veteranos.

Modelo	Acurácia	N° de Variáveis
modelo_log_logito	0,7323	10
modelo_log_logito_stepwise	0,7318	6
modelo_log_probito	0,7323	10
modelo_log_probito_stepwise	**0,7352**	**7**
modelo_rf	0,7171	10
modelo_rf_rfe	0,7175	8
modelo_xg	0,7387	10
modelo_xg_rfe	0,7310	5

Tabela 14 – Resumo dos parâmetros estimados do modelo de classificação de QBs veteranos.

Parâmetro	Estimativa	Erro Padrão	P-Valor
(Intercept)	1,6469	0,8842	0,0625
age	-0,0749	0,0376	0,0462
years_exp	0,0898	0,0380	0,0181
games	0,1140	0,0248	0,0000
passing_td	0,0543	0,0161	0,0008
passing_int	-0,0503	0,0192	0,0089
sacks_suffered	-0,0135	0,0087	0,1233
rushing_yds	0,0014	0,0009	0,1121

Tabela 15 – Matriz de confusão para o modelo de classificação de QBs calouros.

Predito / Real	Negativo	Positivo
Negativo	155	176
Positivo	14	218

A matriz de confusão do modelo está apresentada na Tabela 15. Observa-se que a acurácia geral foi de 66,3%, a sensibilidade atingiu 55,3% e a especificidade foi de 91,7%. Esses resultados indicam que o modelo apresentou boa capacidade preditiva para identificar os casos em que o jogador não pontuaria no *Fantasy Football*. No entanto, seu desempenho foi inferior na identificação dos casos em que a pontuação foi maior que zero, refletido no baixo valor de sensibilidade.

4.1.2.2 Regressão

Os resultados do modelo de regressão estão apresentados na Tabela 16. O modelo selecionado foi o *XGBoost*, utilizando apenas três variáveis explicativas: `passing_td`, `games` e `rushing_yds`. Conforme ilustrado no gráfico de importâncias, Figura 9, essas

Tabela 16 – Métricas de avaliação para os modelos candidatos para a regressão de QBs veteranos.

Modelo	MAE	N° de Variáveis
modelo_gamma_log	116,85	10
modelo_gamma_stepwise	64,48	5
modelo_gamma_inverse	68,39	10
modelo_gamma_inverse_stepwise	68,51	4
modelo_rf	61,14	10
modelo_rf_rfe	61,01	9
modelo_xg	60,67	10
modelo_xg_rfe	**60,62**	**3**

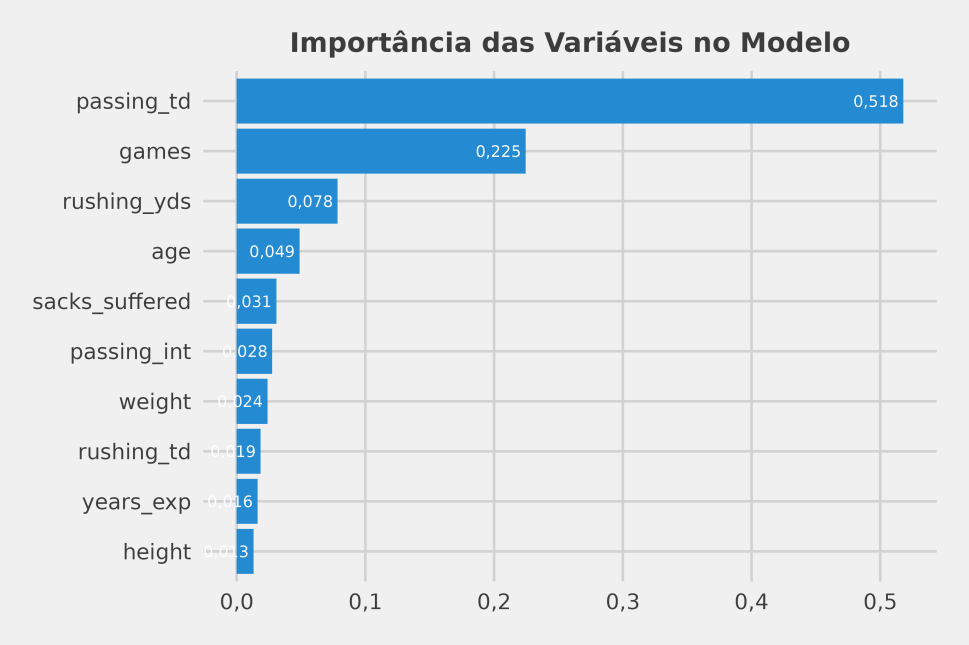

Figura 9 – Importância relativa pela métrica *Gain* das variáveis explicativas do modelo de regressão *XGBoost* para os QBs veteranos.

três variáveis se destacam em relação às demais, especialmente as duas primeiras. Isso demonstra que, mesmo com um conjunto reduzido de preditores, o modelo é capaz de alcançar um bom desempenho preditivo.

Os hiperparâmetros ajustados são apresentados na Tabela 17.

A Figura 10 apresenta a análise dos resíduos do modelo. De modo geral, observa-se um comportamento dentro do esperado: os resíduos aparentam seguir uma distribuição aproximadamente normal e não evidenciam correlação entre si, indicando a ausência de padrões sistemáticos não capturados pelo modelo. Identificam-se apenas alguns poucos *outliers*, que são discutidos de forma mais detalhada na Tabela 18.

Entre os principais *outliers*, destacam-se cinco casos notáveis. O primeiro é *Daunte Culpepper*, na temporada de 2000, sua segunda na NFL. Na temporada anterior, ele não chegou a entrar em campo, o que comprometeu a capacidade do modelo de predizer seu

Tabela 17 – Hiperparâmetros do modelo de regressão *XGBoost* de QBs veteranos.

Hiperparâmetro	Valor
nrounds	199,00
verbose	0,00
max_depth	3,00
eta	0,02
subsample	0,69
colsample_bytree	0,86

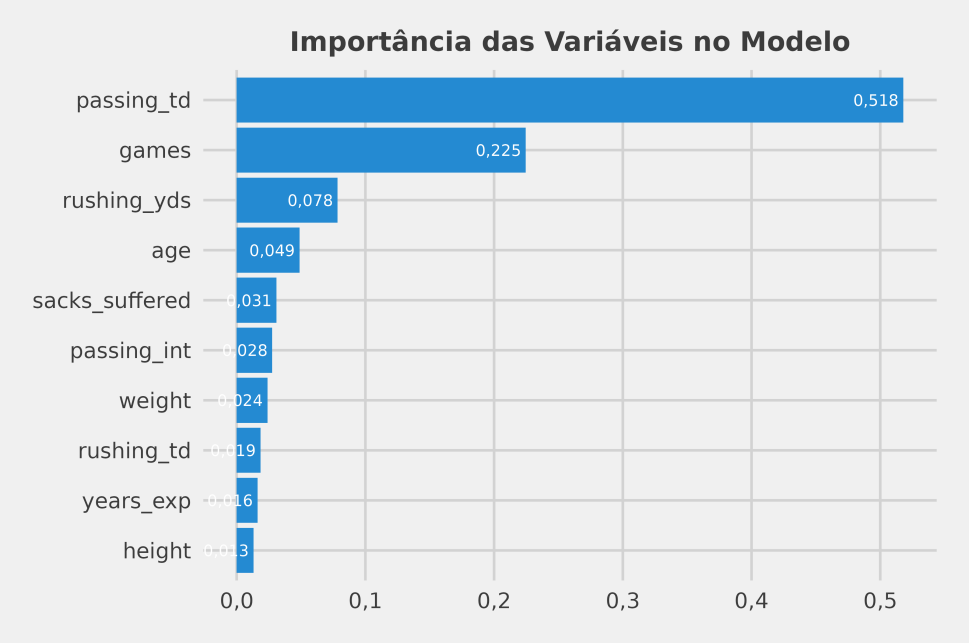

Figura 10 – Gráficos de avaliação dos resíduos no modelo de regressão dos QBs veteranos.

Tabela 18 – Principais *outliers* presentes no modelo de regressão de QBs veteranos.

Jogador	Ano	Pontuação Observada	Pontuação Predita	Resíduo Absoluto
Daunte Culpepper	2000	338,48	41,95	296,53
Jordan Love	2023	319,06	39,47	279,59
Tom Brady	2008	3,04	271,95	268,91
Aaron Rodgers	2014	354,14	95,52	258,62
Jameis Winston	2020	2,40	246,65	244,25

desempenho elevado naquele ano (Pro Football Reference, 2025g).

Outro caso é o de *Jordan Love*, que permaneceu como reserva durante suas três primeiras temporadas. Quando finalmente assumiu a titularidade, apresentou um desempenho expressivo, contudo, o histórico limitado de jogos anteriores prejudicou a precisão da predição (Pro Football Reference, 2025l).

Já *Tom Brady*, eleito o melhor jogador da temporada de 2007, sofreu uma lesão no primeiro jogo de 2008 que o afastou de toda a temporada, tornando seu desempenho naquele ano uma exceção nos dados (Pro Football Reference, 2025y).

No caso de *Aaron Rodgers*, observa-se o efeito oposto: após sofrer uma lesão em 2013, ele teve um desempenho destacado em 2014, sendo inclusive eleito o melhor jogador daquela temporada, um resultado que não foi adequadamente captado pelo modelo (Pro Football Reference, 2025a).

Por fim, *Jameis Winston* teve uma temporada atípica em 2019, acumulando altos volumes de jardas e touchdowns, mas também um número elevado de interceptações. Apesar do bom desempenho no *Fantasy Football*, ele foi transferido para outra equipe e tornou-se reserva em 2020, o que gerou um descompasso entre sua pontuação e o contexto subsequente (Pro Football Reference, 2025j).

4.2 Running Backs

Enquanto os *Quarterbacks* são geralmente considerados os jogadores mais importantes dentro dos times da NFL, os *Running Backs* ocupam essa posição de destaque no *Fantasy Football*. Isso se deve, em grande parte, ao fato de que um time da NFL costuma contar com um RB titular que concentra a maior parte das estatísticas de corrida da equipe, sendo que jogadores de elite nessa posição costumam apresentar alto volume e produção nos quesitos que geram pontuação no *Fantasy*. Dessa forma, identificar atletas valiosos nessa posição pode ser uma tarefa bastante desafiadora, especialmente considerando que esses RBs de elite estão entre os mais cobiçados nos *drafts* das ligas de *Fantasy Football* (Pro Football Focus, 2024).

Essa posição também possui uma característica marcante: os jogadores frequentemente chegam da universidade prontos para impactar seus times na NFL. No entanto, devido à alta fisicalidade envolvida em sua atuação, com constantes pancadas e contato físico intenso, a carreira dos RBs tende a ser significativamente mais curta em comparação a outras posições.

A modelagem realizada para essa posição seguiu os mesmos princípios adotados no caso dos *Quarterbacks*, sendo utilizadas variáveis relacionadas às estatísticas de corrida e de recepção, bem como características individuais e contextuais de cada atleta.

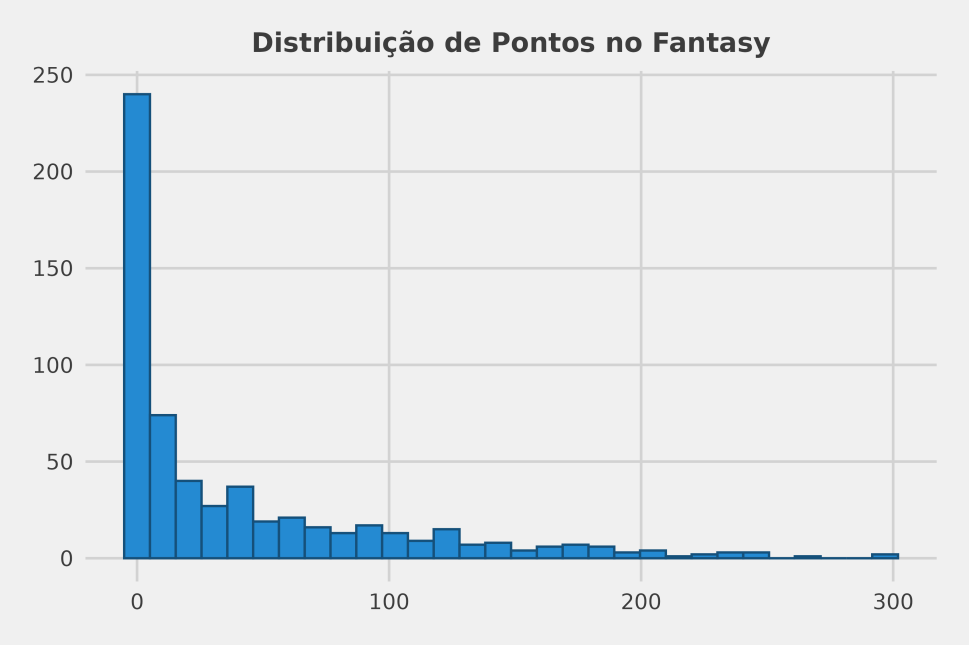

Figura 11 – Distribuição da pontuação no *Fantasy* para *Running Backs* em seu primeiro ano na NFL.

4.2.1 Calouros

Os *Running Backs* calouros costumam impactar significativamente seus times já na primeira temporada na NFL, diferentemente do que foi observado para os *Quarterbacks* (ESPN, 2025). No entanto, o número de jogadores oriundos do futebol universitário nessa posição é, consideravelmente maior do que nas demais. Isso faz com que haja, por um lado, atletas com desempenho comparável ao de veteranos e, por outro, jogadores com pontuação muito baixa ou até mesmo nula no *Fantasy Football*. A distribuição da variável resposta pode ser observada na Figura 11, com uma proporção de 25,3% de valores iguais a zero. Diante desse cenário, optou-se, novamente, pela aplicação de um modelo de duas etapas.

As variáveis `draft_round_group` e `conference_group` são apresentadas na Figura 12. A variável `draft_round_group` apresenta uma dinâmica semelhante à observada para os *Quarterbacks*, embora de forma menos acentuada: a diferença entre a primeira rodada e os demais grupos é menor, sendo possível identificar pontuações elevadas também entre jogadores provenientes de `Early Rounds`, `Mid Rounds` e até mesmo `Late Rounds`. Quanto à variável `conference_group`, a distribuição da resposta mostra-se bastante similar entre os dois níveis considerados.

Como mencionado anteriormente, entre as posições pontuadoras no *Fantasy Football*, os *Running Backs* são os que mais dependem da fisicalidade dos atletas. Essas características estão representadas na Figura 13. No gráfico superior, observa-se uma relação inversa entre idade e pontuação: quanto mais jovem o atleta deixa o *College*, maior tende a ser sua pontuação. No entanto, essa variável pode estar correlacionada com

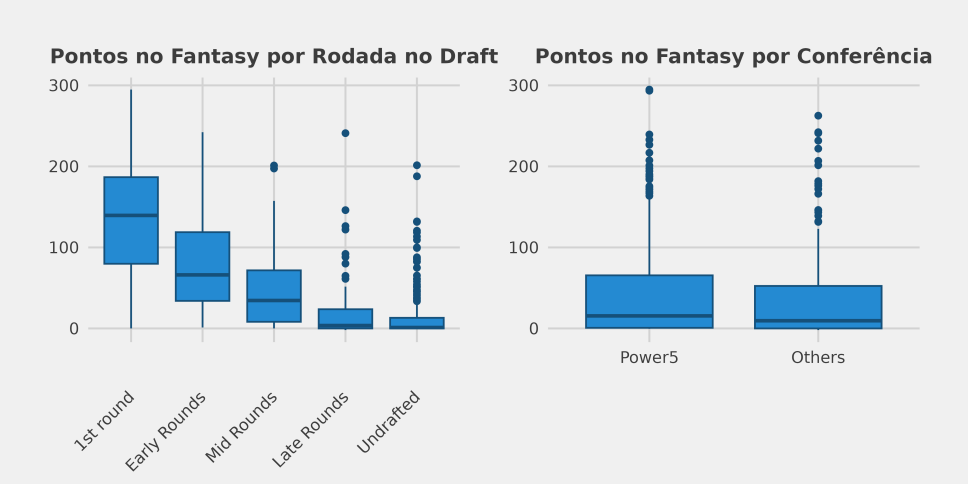

Figura 12 – Distribuição da pontuação no *Fantasy* por rodada em que o *Running Back* foi selecionado no *Draft* da NFL, e de acordo com a conferência do *College Football* ele atuou.

Tabela 19 – Análise de correlação para as variáveis de corrida para os RBs calouros.

	rushing_car	rushing_yds	rushing_td
rushing_car	1,00	0,97	0,86
rushing_yds	0,97	1,00	0,89
rushing_td	0,86	0,89	1,00

Tabela 20 – Análise de correlação para as variáveis de recepção para os RBs calouros.

	receiving_rec	receiving_yds	receiving_td
receiving_rec	1,00	0,95	0,69
receiving_yds	0,95	1,00	0,78
receiving_td	0,69	0,78	1,00

`draft_round_group`, uma vez que jogadores mais talentosos, geralmente, saem mais cedo do universitário e tendem a ser selecionados nas primeiras rodadas do `Draft` da NFL.

À esquerda, a relação entre altura e pontuação é relativamente estável até aproximadamente 74 polegadas, ponto a partir do qual jogadores mais altos tendem a apresentar queda de desempenho. À direita, a relação com o peso mostra uma tendência crescente até certo ponto, indicando que jogadores muito leves enfrentam maior dificuldade para pontuar. No entanto, após um ponto de inflexão, jogadores mais pesados também apresentam declínio na pontuação esperada.

Por fim, foi conduzido as análises de correlação para colunas de corrida, Tabela 19, e de recepção, Tabela 20. As variáveis selecionadas foram: `rushing_car`, `rushing_td`, `receiving_rec` e `receiving_td`.

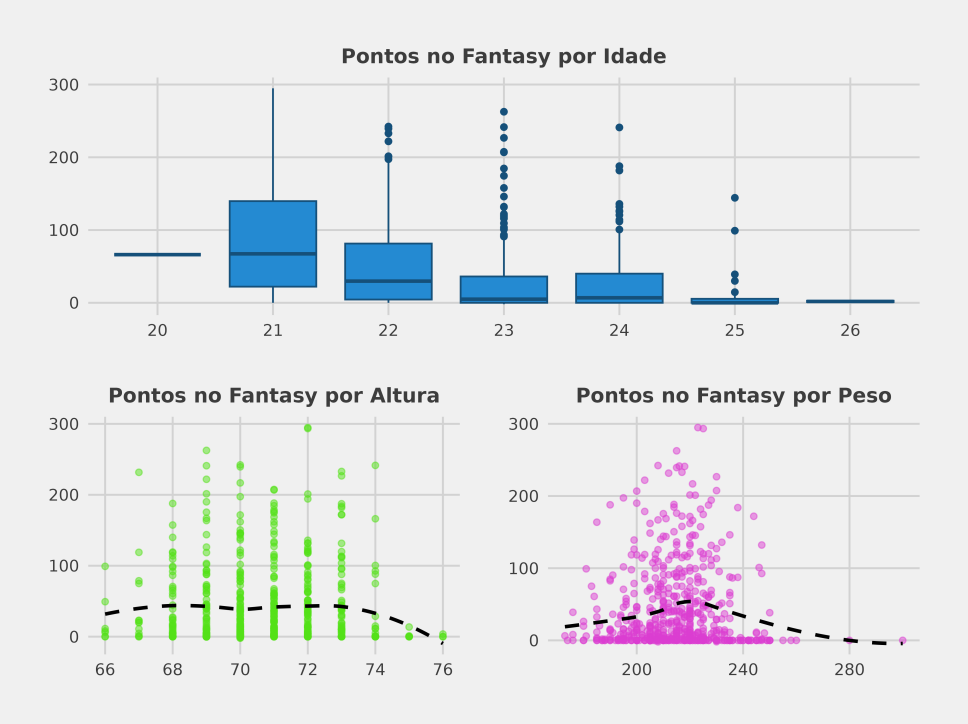

Figura 13 – Distribuição da pontuação no *Fantasy* por características pessoais dos RBs calouros.

Tabela 21 – Métricas de avaliação para os modelos candidatos para a classificação de RBs calouros.

Modelo	Acurácia	N° de Variáveis
modelo_log_logito	0,7506	13
modelo_log_logito_stepwise	0,7544	8
modelo_log_probito	0,7561	13
modelo_log_probito_stepwise	0,7568	7
modelo_rf	0,7694	15
modelo_rf_rfe	0,7616	13
modelo_xg	0,7812	15
modelo_xg_rfe	**0,7844**	**8**

4.2.1.1 Classificação

Os resultados para o modelo de classificação pode ser observada na Tabela 21, o modelo preterido foi o *XGBoost*, com 8 variáveis explicativas. Esse modelo foi escolhido por obter a melhor acurácia balanceada entre os modelo propostos. As variáveis selecionadas foram: `weight`, `height`, `rushing_car`, `rushing_td`, `receiving_rec` e 3 níveis de `draft_round_group` - `early_rounds`, `late_rounds` e `undrafted`.

As respectivas importâncias relativas dessas variáveis estão apresentadas na Figura 14, podemos observar que a variável `weight` e também o nível `undrafted` são as que obtiveram os maiores valores, porém as demais variáveis ainda possuem importâncias

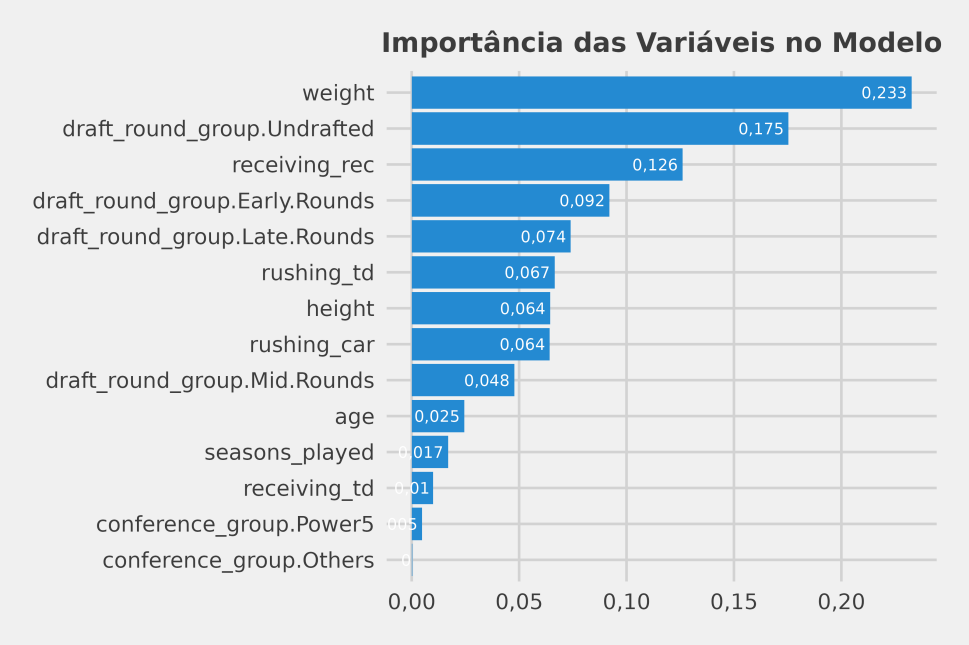

Figura 14 – Importância relativa pela métrica Gain das variáveis explicativas do modelo de classificação *XGBoost* para os RBs calouros.

Tabela 22 – Hiperparâmetros do modelo de classificação *XGBoost* de RBs calouros.

Hiperparâmetro	Valor
nrounds	105,00
verbose	0,00
max_depth	2,00
eta	0,07
subsample	0,92
colsample_bytree	0,62

Tabela 23 – Matriz de confusão para o modelo de classificação de RBs calouros.

Predito / Real	Negativo	Positivo
Negativo	46	38
Positivo	7	89

consideráveis.

Os hiperparâmetros ajustados são apresentados na Tabela 22.

Por fim, a matriz de confusão 23 apresenta a capacidade preditiva do modelo. A acurácia alcançada foi de 75,0%, com sensibilidade de 70,1% e especificidade de 86,8%. Esses resultados indicam que o modelo tem boa capacidade preditiva geral, com destaque para a predição de casos negativos (não pontuadores) com alta especificidade.

Tabela 24 – Métricas de avaliação para os modelos candidatos para a regressão de RBs veteranos.

Modelo	MAE	N° de Variáveis
modelo_gamma_log	54,31	13
modelo_gamma_log_stepwise	43,26	6
modelo_gamma_inverse	72,47	13
modelo_gamma_inverse_stepwise	83,10	6
modelo_rf1	40,43	15
modelo_rf_rfe	**40,82**	**11**
modelo_xg1	40,52	15
modelo_xg_rfe	41,70	6

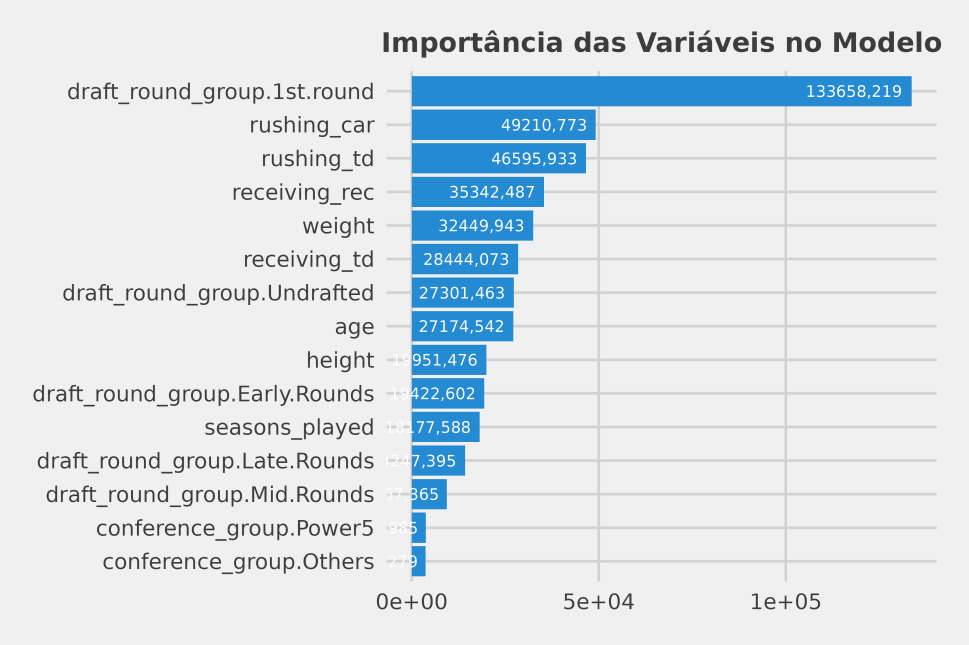

Figura 15 – Importância relativa pela métrica MDI das variáveis explicativas do modelo de regressão *Random Forest* para os RBs calouros.

4.2.1.2 Regressão

Já para o modelo de regressão, os resultados são apresentados na Tabela 24. O modelo selecionado foi o *Random Forest*, com um total de 11 variáveis explicativas, escolhido por apresentar o menor MAE entre os modelos com seleção de variáveis. As variáveis selecionadas foram: `draft_round_group` (nos níveis `1st Round`, `Early Rounds` e `Undrafted`), `age`, `height`, `weight`, `rushing_car`, `rushing_td`, `receiving_rec`, `receiving_td` e `seasons_played`.

As respectivas importâncias relativas dessas variáveis estão apresentadas na Figura 15, sendo que o nível `1st Round` destacou-se com valor expressivamente superior aos demais, evidenciando-se como um fator crucial para a estimativa da pontuação dos jogadores calouros.

Tabela 25 – Hiperparâmetros do modelo de classificação *Random Forest* de RBs calouros.

Hiperparâmetro	Valor
mtry	4,00
sample.fraction	0,63
replace	0,00
nodesize	10,00
num.trees	1000,00

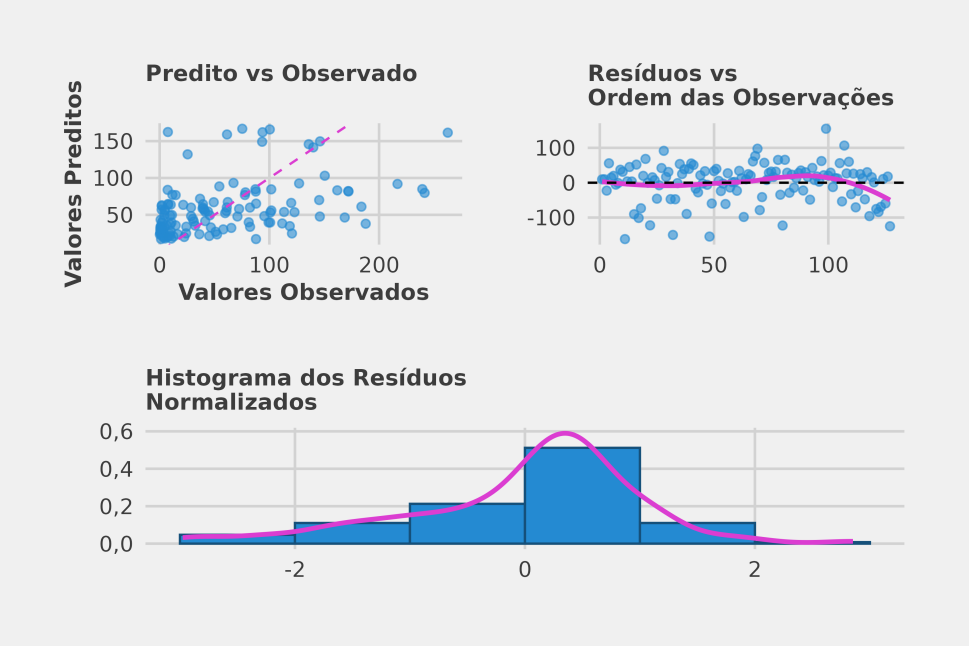

Figura 16 – Gráficos de avaliação dos resíduos no modelo de regressão dos RBs calouros.

Tabela 26 – Principais *outliers* presentes no modelo de regressão de RBs calouros.

Jogador	Pontuação Observada	Pontuação Predita	Resíduo Absoluto
Matt Forte	241,5	79,77	161,73
Rashard Mendenhall	7,5	162,34	154,84
Alvin Kamara	239,4	84,88	154,52
Phillip Lindsay	187,8	37,97	149,83
Jonathan Taylor	216,8	91,91	124,89

Os hiperparâmetros ajustados são apresentados na Tabela 25.

O desempenho do modelo é apresentado na Figura 16, na qual observa-se que os resíduos se comportam conforme o esperado, aparentando seguir uma distribuição aproximadamente normal. No gráfico de "Predito vs Observado", nota-se que o modelo tende a subestimar os valores mais altos da variável resposta, indicando limitação na predição dos casos de maior desempenho. Diante disso, os principais *outliers* destacados pelo modelo foram analisados e estão apresentados na Tabela 26.

Os jogadores *Matt Forte*, *Alvin Kamara* e *Johnathan Taylor* foram selecionados

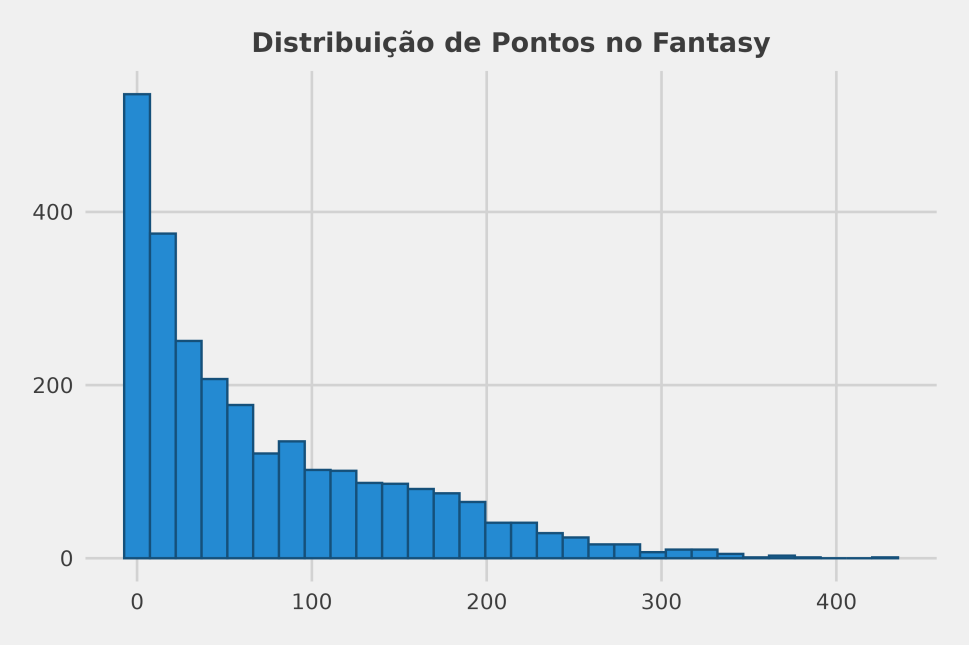

Figura 17 – Distribuição da pontuação no *Fantasy* para RBs veteranos da NFL.

nas `Mid Rounds` do *Draft* e apresentaram um impacto muito acima do esperado já em suas temporadas de estreia. É relevante destacar que esses atletas são conhecidos pela sua versatilidade, com capacidade tanto para corridas quanto para recepções, o que os torna especialmente valiosos no *Fantasy Football* (Pro Football Reference, 2025o; Pro Football Reference, 2025c; Pro Football Reference, 2025k).

Um caso semelhante é o de *Phillip Lindsay*, que não foi selecionado no *Draft* (`Undrafted`), mas teve uma temporada de calouro extremamente atípica e de destaque para alguém do seu grupo (Pro Football Reference, 2025t).

Por outro lado, *Rashard Mendenhall* era um jogador altamente cotado ao sair do *College*, mas sofreu uma lesão logo em sua primeira partida como titular, o que o afastou do restante da temporada de calouro, impactando negativamente seu desempenho (Pro Football Reference, 2025v).

4.2.2 Veteranos

Para os *Running Backs* veteranos, observou-se uma baixa incidência de pontuação zerada, com proporção de apenas 5,1%. Diante disso, optou-se por ajustar apenas um modelo de regressão. A distribuição da variável resposta está apresentada na Figura 17.

Espera-se que a idade seja um fator determinante na pontuação dos *Running Backs* veteranos. Conforme ilustrado na Figura 18, observa-se uma queda acentuada na pontuação a partir da faixa etária de 29 a 32 anos, indicando que o auge desses atletas tende a ocorrer nos primeiros anos após a entrada na NFL. Nos gráficos de altura e peso, identifica-se uma relação relativamente constante ao longo de toda a amplitude dessas variáveis, com uma leve tendência de aumento na pontuação a partir de 72 polegadas de

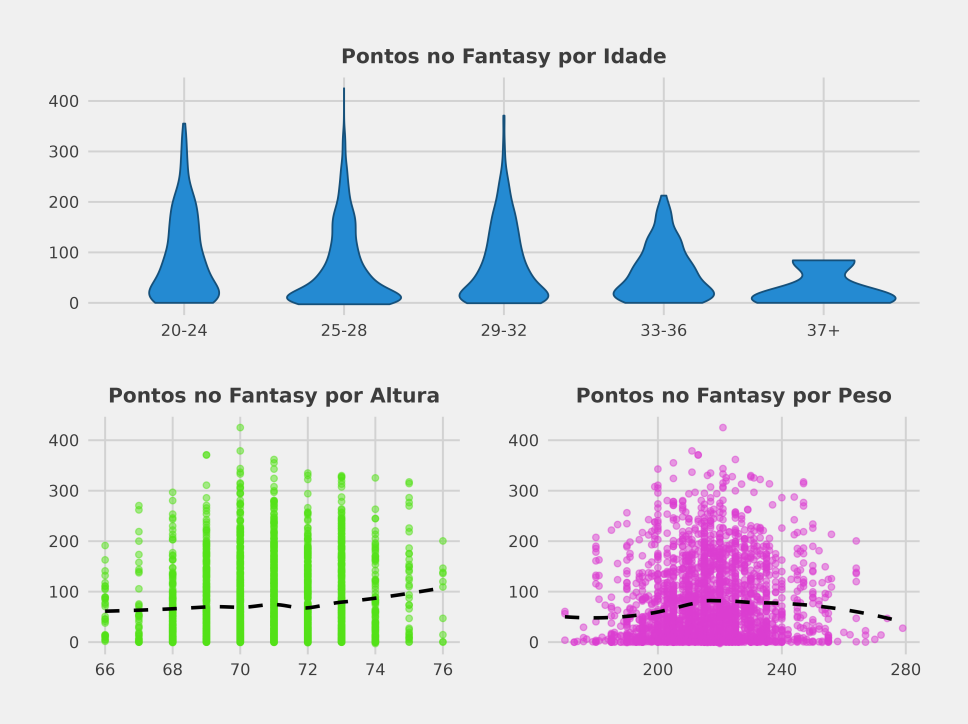

Figura 18 – Distribuição na pontuação no *Fantasy* para as características pessoais dos RBs veteranos da NFL.

Tabela 27 – Análise de correlação para as variáveis de corrida para os RBs veteranos.

	rushing_car	rushing_yds	rushing_first_downs	rushing_td
rushing_car	1,00	0,98	0,97	0,82
rushing_yds	0,98	1,00	0,98	0,84
rushing_first_downs	0,97	0,98	1,00	0,87
rushing_td	0,82	0,84	0,87	1,00

Tabela 28 – Análise de correlação para as variáveis de recepção para os RBs veteranos.

	targets	receiving_rec	receiving_yds	receiving_td	receiving_yards_after_catch	receiving_first_downs
targets	1,00	0,81	0,79	0,58	0,70	0,78
receiving_rec	0,81	1,00	0,97	0,65	0,75	0,95
receiving_yds	0,79	0,97	1,00	0,69	0,76	0,97
receiving_td	0,58	0,65	0,69	1,00	0,56	0,70
receiving_yards_after_catch	0,70	0,75	0,76	0,56	1,00	0,75
receiving_first_downs	0,78	0,95	0,97	0,70	0,75	1,00

altura. Ainda assim, as maiores pontuações concentram-se em torno de 70 polegadas.

A análise de correlação, apresentada nas Tabelas 27 e 28, foi conduzida entre as variáveis relacionadas às jogadas de corrida e recepção. No caso dos jogadores da NFL, foram incluídas as variáveis adicionais `rushing_first_downs` para corrida, e `targets`, `receiving_yards_after_catch` e `receiving_first_downs` para recepção. Ao final da análise, optou-se pela remoção das variáveis `rushing_yds`, `rushing_first_downs`, `receiving_rec` e `receiving_yds`, com base em sua alta correlação com outras covariáveis.

Tabela 29 – Métricas de avaliação para os modelos candidatos para a regressão de RBs veteranos.

Modelo	MAE	N° de Variáveis
modelo_rf	44,28	11
modelo_rf_rfe	44,20	8
modelo_xg	43,88	11
modelo_xg_rfe	**43,57**	**9**

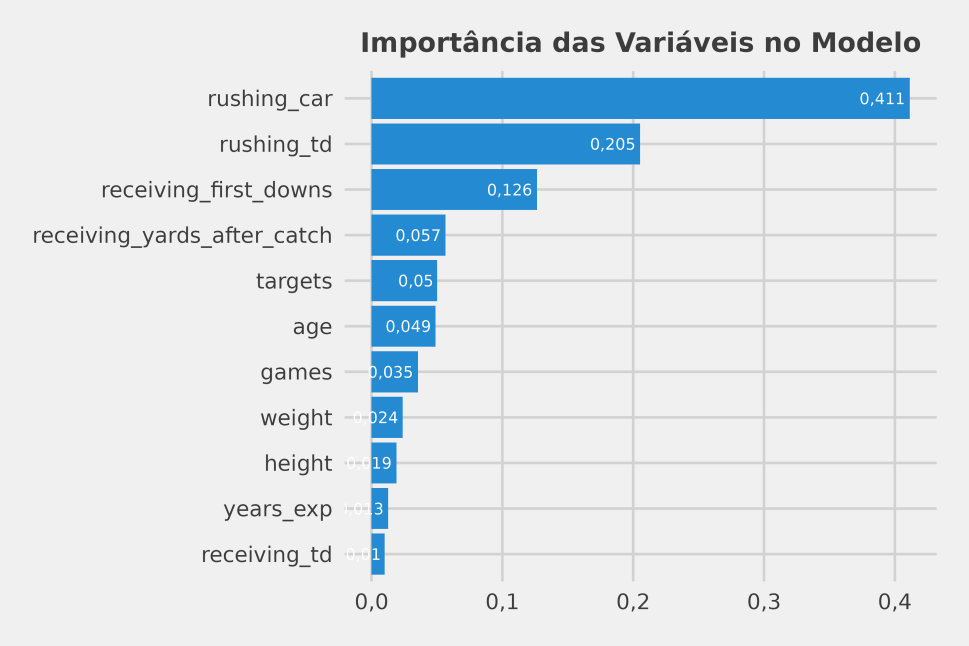

Figura 19 – Importância relativa pela métrica *Gain* das variáveis explicativas do modelo de regressão *XGBoost* para os RBs veteranos

4.2.2.1 Regressão

Partindo diretamente para o modelo de regressão, os resultados são apresentados na Tabela 29. Neste caso, não foram considerados os modelos com distribuição Gamma, uma vez que esta admite apenas variáveis resposta estritamente positivas, o que não se aplica ao presente cenário. O modelo selecionado foi o *XGBoost*, com um total de nove covariáveis, por ter apresentado o menor MAE entre as alternativas avaliadas.

As variáveis utilizadas foram: `games`, `age`, `weight`, `rushing_car`, `rushing_td`, `receiving_first_downs`, `receiving_yards_after_catch` e `targets`. O gráfico de importância relativa das covariáveis está apresentado na Figura 19, e observa-se que as variáveis relacionadas às estatísticas de jogo foram as mais relevantes, com destaque para aquelas associadas às jogadas de corrida.

Os hiperparâmetros ajustados para o modelo estão descritos na Tabela 30.

Por final, foi conduzido a análise de resíduos conforme a Figura 20, notou-se um certa assimetria nos resíduos, com uma cauda inferior mais longa, sendo possível fazer essa observação pelo histograma.

Tabela 30 – Hiperparâmetros do modelo de regressão *XGBoost* de RBs veteranos.

Hiperparâmetro	Valor
nrounds	181,00
verbose	0,00
max_depth	3,00
eta	0,03
subsample	0,61
colsample_bytree	0,60

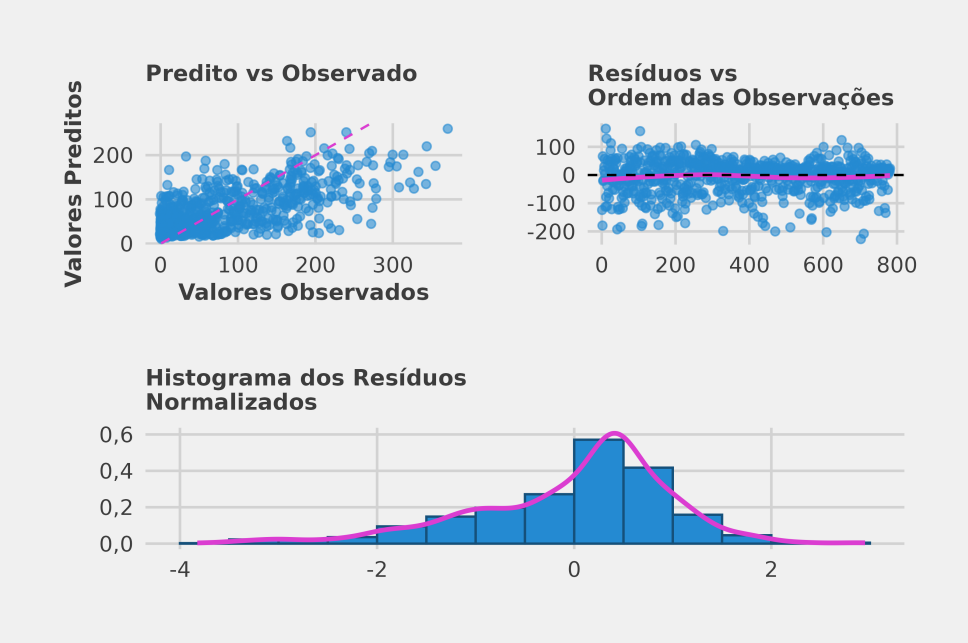

Figura 20 – Gráficos de avaliação dos resíduos no modelo de regressão dos RBs calouros.

Os principais *outliers* são apresentados na Tabela 31. Em todos os casos, o modelo subestimou a pontuação real dos jogadores. Para *Chris Johnson* e *Larry Johnson*, por exemplo, foram temporadas atípicas para *Running Backs* da NFL. Embora o modelo tenha previsto uma pontuação relativamente alta, *Chris Johnson* superou todas as expectativas ao registrar mais de 2.000 jardas corridas, o que, na época, representou o quinto melhor desempenho da história da liga (Pro Football Reference, 2025f). Já *Larry Johnson* marcou 20 *touchdowns*, figurando entre os dez melhores desempenhos da história nesse quesito (Pro Football Reference, 2025n).

O caso de *Michael Turner* se destaca pelo fato de que era um jogador jovem que nunca havia tido uma temporada particularmente produtiva anteriormente, o que influenciou na subestimação de sua pontuação (Pro Football Reference, 2025q). Por fim, *Adrian Peterson* perdeu toda a temporada de 2014 por conta de uma lesão e retornou em 2015 já com 30 anos de idade. Apesar disso, teve um desempenho notável, com uma pontuação elevada que não foi adequadamente captada pelo modelo (Pro Football Reference, 2025b).

Tabela 31 – Principais *outliers* presentes no modelo de regressão de RBs veteranos.

Jogador	Ano	Pontuação Observada	Pontuação Predita	Resíduo Absoluto
Michael Turner	2008	272,0	44,75	227,25
Chris Johnson	2009	342,9	134,73	208,17
Larry Johnson	2005	327,3	124,36	202,94
Adrian Peterson	2015	230,7	30,53	200,17

4.3 Wide Receivers

Os *Wide Receivers* também são considerados uma posição *premium* no *Fantasy Football*. No entanto, diferentemente dos *Running Backs*, um time da NFL costuma contar com três WRs titulares, e há formações em que até cinco jogadores dessa posição podem estar em campo simultaneamente. Por esse motivo, a pontuação não fica concentrada em um pequeno grupo de jogadores de elite, sendo mais distribuída entre os atletas da posição.

Esses jogadores atuam de forma quase exclusiva no jogo aéreo das equipes, sendo os principais alvos dos passes lançados pelos *Quarterbacks*. Para desempenhar essa função com eficiência, os WRs geralmente são atletas mais leves, ágeis e rápidos, características essenciais para escapar da marcação dos defensores. Ainda que, em média, os principais WRs toquem menos na bola do que RBs titulares, são capazes de acumular um grande volume de jardas e apresentar uma maior longevidade na carreira, quando comparados aos *Running Backs* (Bleacher Report, 2022).

Na modelagem aplicada a essa posição, foram utilizadas apenas variáveis relacionadas ao desempenho em recepções, além das variáveis de características físicas e contexto dos jogadores, também utilizadas nas modelagens anteriores.

4.3.1 Calouros

Assim como os *Running Backs*, os *Wide Receivers* também apresentam um período de adaptação mais rápido à liga profissional, especialmente quando comparados aos *Quarterbacks*. Dessa forma, atletas que se destacaram no *College Football* tendem a ser bastante valorizados nos *Drafts* das ligas de *Fantasy Football* (Bleacher Report, 2024).

Ainda assim, observa-se uma elevada proporção de jogadores que não pontuam em suas temporadas de estreia. A Figura 21 apresenta a distribuição da variável resposta para os WRs calouros, revelando que 26,6% dos casos correspondem a pontuações zeradas. Diante disso, aplicou-se a modelagem em duas etapas.

As variáveis `draft_round_group` e `conference_group` são apresentadas na Figura 22. A variável `draft_round_group` exibe uma dinâmica semelhante à observada para as demais posições: jogadores selecionados na primeira rodada tendem a apresentar pontuações mais elevadas no *Fantasy Football* em comparação aos demais grupos. Ainda assim, identificam-se pontuações relevantes entre atletas selecionados em rodadas mais

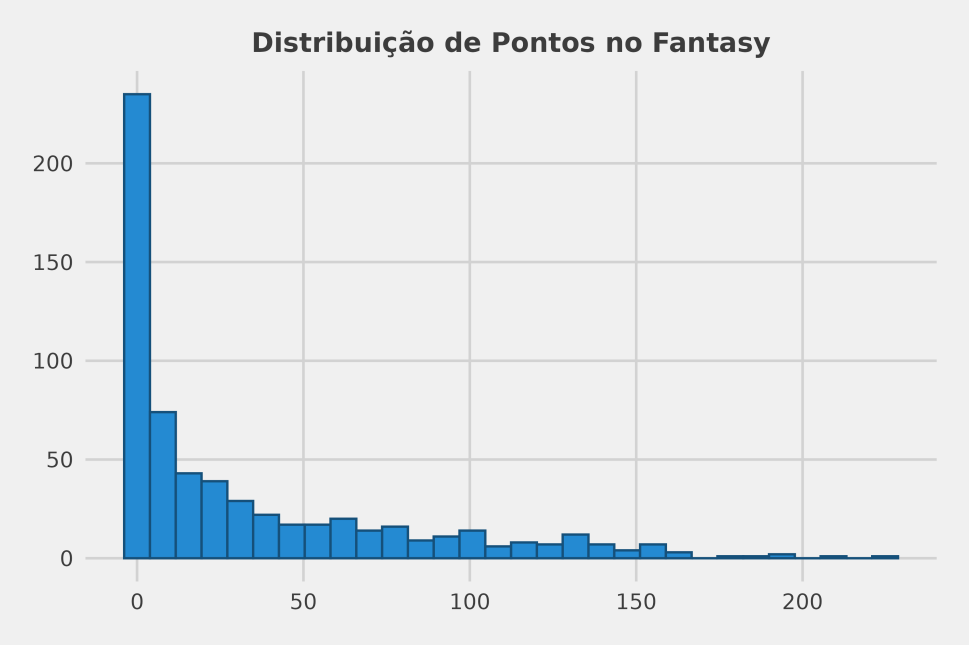

Figura 21 – Distribuição da pontuação no *Fantasy* para *Wide Receivers* em seu primeiro ano na NFL.

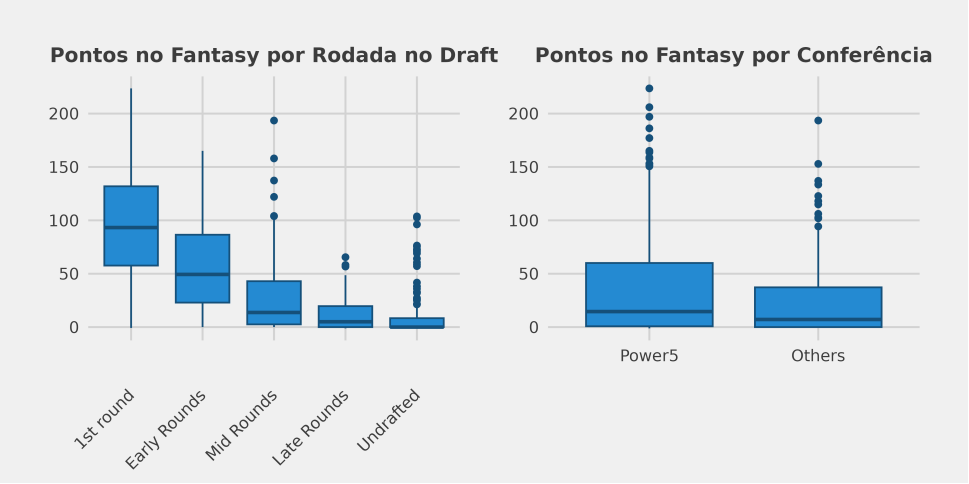

Figura 22 – Distribuição da pontuação no *Fantasy* por rodada em que o *Wide Receiver* foi selecionado no *Draft* da NFL, e de acordo com a conferência do *College Football* ele atuou.

tardias do *Draft*, o que indica que o impacto imediato dos WRs calouros não está restrito apenas aos jogadores mais valorizados na seleção.

Quanto à variável `conference_group`, observa-se que o nível `Power5` apresenta média e terceiro quartil levemente superiores aos das demais conferências. No entanto, a diferença é pouco expressiva, sugerindo que essa variável exerce influência limitada sobre a pontuação dos WRs em suas temporadas de estreia.

O comportamento das variáveis `age`, `height` e `weight` é explorado na Figura 23. No gráfico superior, observa-se novamente uma relação decrescente entre a variável resposta e a idade dos jogadores, o que pode estar confundido com a variável `draft_round_group`,

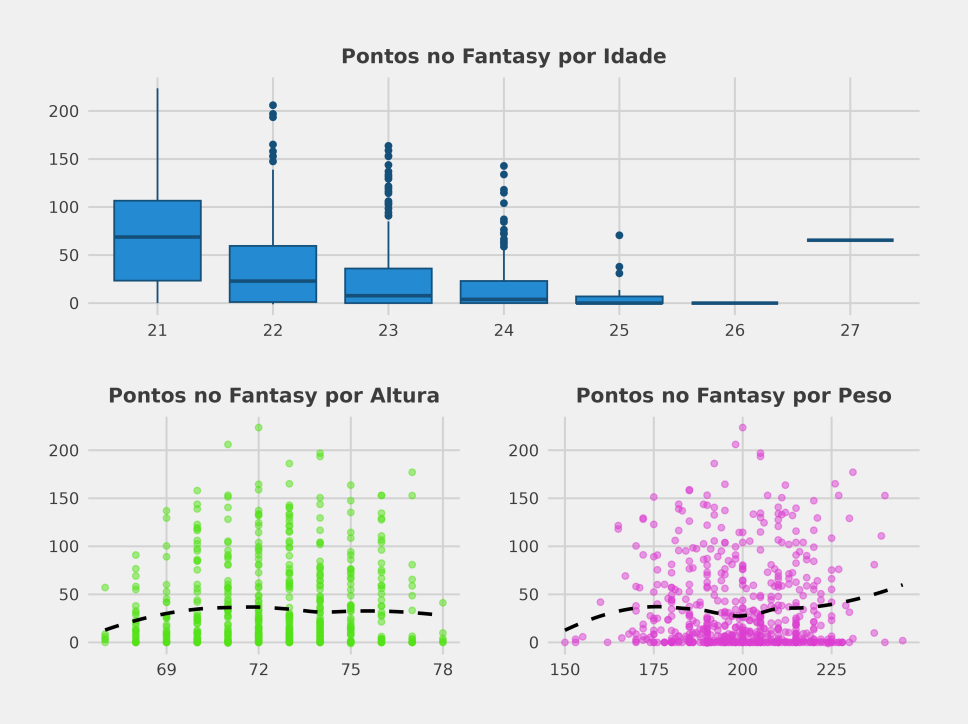

Figura 23 – Distribuição da pontuação no *Fantasy* por características pessoais dos WRs calouros.

Tabela 32 – Análise de correlação para as variáveis de recepção para os WRs calouros.

	receiving_rec	receiving_yds	receiving_td
receiving_rec	1,00	0,95	0,77
receiving_yds	0,95	1,00	0,87
receiving_td	0,77	0,87	1,00

visto que atletas mais talentosos tendem a ingressar na NFL mais cedo.

Nos gráficos inferiores, à esquerda (altura) e à direita (peso), identificam-se relações relativamente constantes ao longo da maior parte da distribuição. As variações observadas nas extremidades podem ser atribuídas à menor frequência de observações nessas faixas, o que limita a robustez das inferências para esses valores extremos.

Para a análise de correlação, foram consideradas exclusivamente as variáveis relacionadas às recepções, conforme apresentado na Tabela 32. A variável `receiving_yds` foi removida por não atender aos critérios previamente estabelecidos para seleção de covariáveis.

4.3.1.1 Classificação

Partindo para a modelagem, os resultados dos modelos propostos para classificação, são apresentados na Tabela 33, o modelo escolhido foi o *XGBoost*, com 12 variáveis explicativas selecionadas pelo método RFE, a única que ficou de fora foi o nível `Others` da variável `conference_group`.

Tabela 33 – Métricas de avaliação para os modelos candidatos para a classificação de WRs calouros.

Modelo	Acurácia	N° de Variáveis
modelo_log_logito	0,7519	11
modelo_log_logito_stepwise	0,7529	8
modelo_log_probito	0,7556	11
modelo_log_probito_stepwise	0,7529	8
modelo_rf	0,7613	13
modelo_rf_rfe	0,7566	11
modelo_xg	0,7709	13
modelo_xg_rfe	**0,7609**	**12**

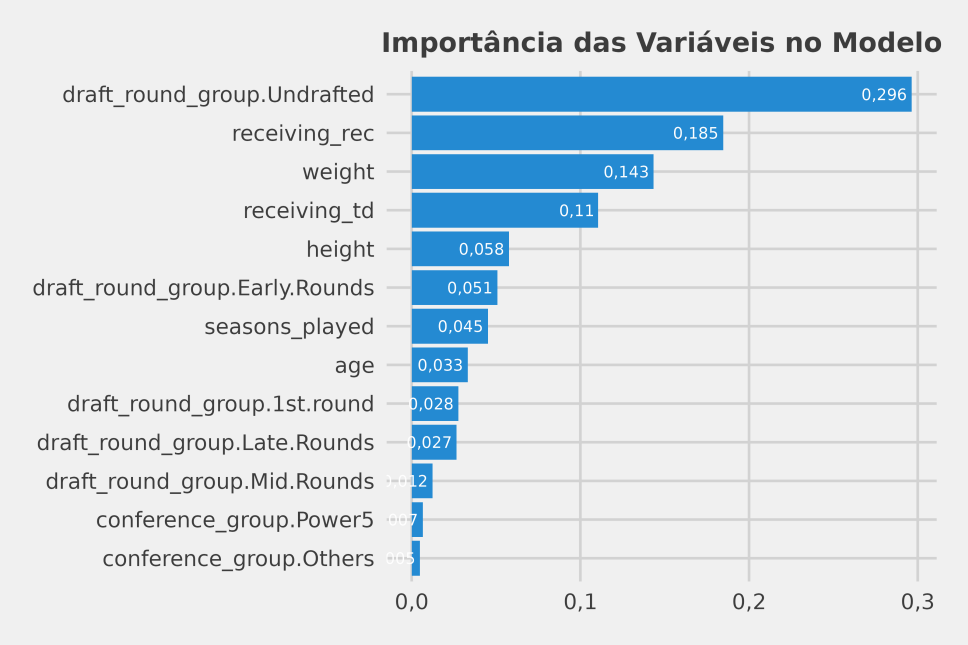

Figura 24 – Importância relativa pela métrica Gain das variáveis explicativas do modelo de classificação *XGBoost* para os WRs calouros.

As importâncias relativas são apresentadas na Figura 24, a variável com maior importância foi o nível `Undrafted` da variável `draft_round_group`, que pode nos indicar um jogador ser ou não draftado é crucial para sabermos se ele vai pontuar ou não. As variáveis de recepção (`receiving_rec` e `receiving_td`) também se destacaram, o que é um comportamento esperado, já que são as únicas estatísticas de jogo presentes para esse posição.

Os hiperparâmetros ajustados são apresentados na Tabela 34.

Por fim, é apresentado a matriz de confusão na Tabela 35. A acurácia alcançada foi de 71,5%, com sensibilidade de 66,2% e especificidade de 86,0%. Os resultados obtidos indicam que o modelo tem bom desempenho na predição de casos negativos e um desempenho razoável para os positivos. Um comportamento similar com o observado nos modelos de classificação das demais posições.

Tabela 34 – Hiperparâmetros do modelo de classificação *XGBoost* de WRs calouros.

Hiperparâmetro	Valor
nrounds	70,00
verbose	0,00
max_depth	6,00
eta	0,03
subsample	0,84
colsample_bytree	0,55

Tabela 35 – Matriz de confusão para o modelo de classificação de WRs calouros.

Predito / Real	Negativo	Positivo
Negativo	43	46
Positivo	7	90

Tabela 36 – Métricas de avaliação para os modelos candidatos para a regressão de WRs calouros.

Modelo	MAE	N° de Variáveis
modelo_gamma_log	38,27	11
modelo_gamma_log_stepwise	29,65	7
modelo_gamma_inverse	30,06	11
modelo_gamma_inverse_stepwise	32,54	5
modelo_rf1	31,47	13
modelo_rf_rfe	31,17	6
modelo_xg1	30,09	13
modelo_xg_rfe	**29,37**	**4**

4.3.1.2 Regressão

Seguindo a abordagem de modelagem em duas etapas, os resultados obtidos para os modelos candidatos à etapa de regressão são apresentados na Tabela 36. O modelo com melhor desempenho, segundo a métrica de MAE, foi o *XGBoost*, que selecionou apenas quatro variáveis explicativas. As variáveis incluídas no modelo final foram `draft_round_group` (nos níveis `1st Round`, `Early Rounds` e `Late Rounds`) e `receiving_td`.

As respectivas importâncias relativas dessas variáveis estão apresentadas na Figura 25, sendo que o nível `1st Round` destacou-se com valor expressivamente superior aos demais, evidenciando-se como um fator crucial para a estimativa da pontuação dos jogadores calouros.

Os hiperparâmetros ajustados são apresentados na Tabela 37.

Por fim, realizando a análise dos resíduos do modelo, temos a Figura 26, em que podemos observar um comportamento dentro do esperado para os resíduos, seguindo uma distribuição bem próxima da normal. O único ponto que se sobressai é um `outlier`, que

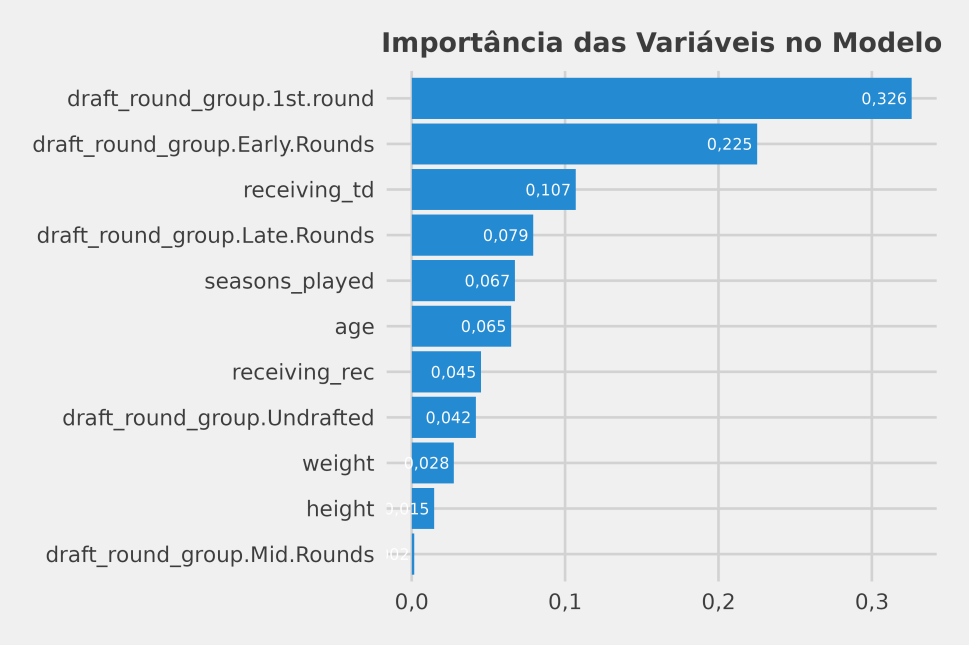

Figura 25 – Importância relativa pela métrica *Gain* das variáveis explicativas do modelo de regressão *XGBoost* para os WRs calouros.

Tabela 37 – Hiperparâmetros do modelo de regressão *XGBoost* de WRs calouros.

Hiperparâmetro	Valor
nrounds	69,00
verbose	0,00
max_depth	2,00
eta	0,04
subsample	0,87
colsample_bytree	0,88

Tabela 38 – Principais *outliers* presentes no modelo de regressão de WRs calouros.

Jogador	Pontuação Observada	Pontuação Predita	Resíduo Absoluto
Puka Nacua	193,5	31,02	162,48

foi melhor explorado na Tabela 38.

Esse *outlier* foi o jogador *Puka Nacua*, selecionado na quinta rodada do *Draft* de 2023. Ele atingiu a maior marca de jardas recebidas por um calouro na história da NFL, o que representa uma exceção notável e uma tendência extremamente difícil de ser prevista pelo modelo (Pro Football Reference, 2025u).

4.3.2 Veteranos

Os *Wide Receivers* veteranos, obtiveram um baixo índice de pontuação zerada, apenas 7,2% dos valores. Portanto, não foi necessário um modelo para lidar com a inflação

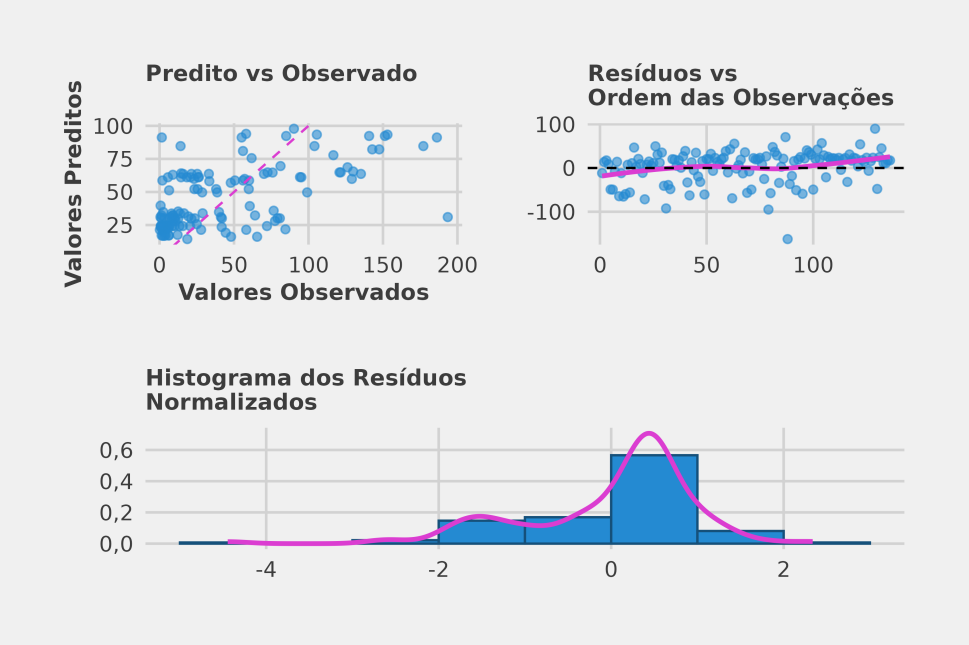

Figura 26 – Gráficos de avaliação dos resíduos no modelo de regressão dos WRs calouros.

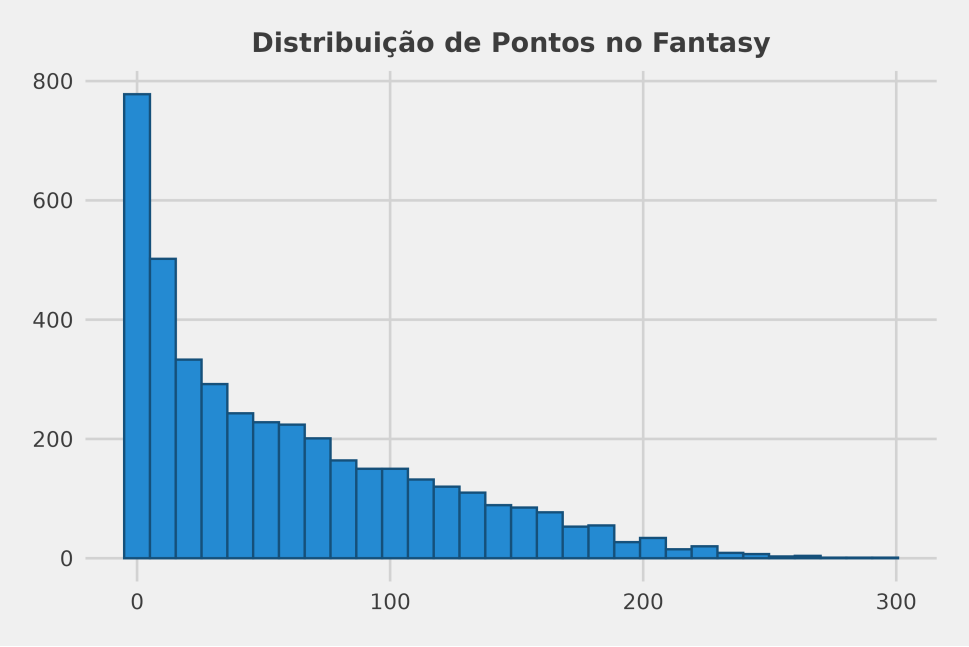

Figura 27 – Distribuição da pontuação no *Fantasy* para WRs veteranos da NFL.

de zeros. A distribuição da variável resposta para os WRs veteranos é explorada na Figura 27.

Na Figura 28 são apresentadas as relações entre as variáveis de características pessoais e a pontuação dos atletas. No gráfico superior, observa-se a distribuição por idade, com maior concentração de jogadores entre 20 e 32 anos, faixa que representa o auge de desempenho, com as maiores pontuações registradas. Ainda assim, pontuações relevantes também são observadas em atletas mais velhos. À esquerda, na linha inferior, encontra-se o gráfico de altura, que apresenta crescimento até aproximadamente 70 polegadas, mantendo-

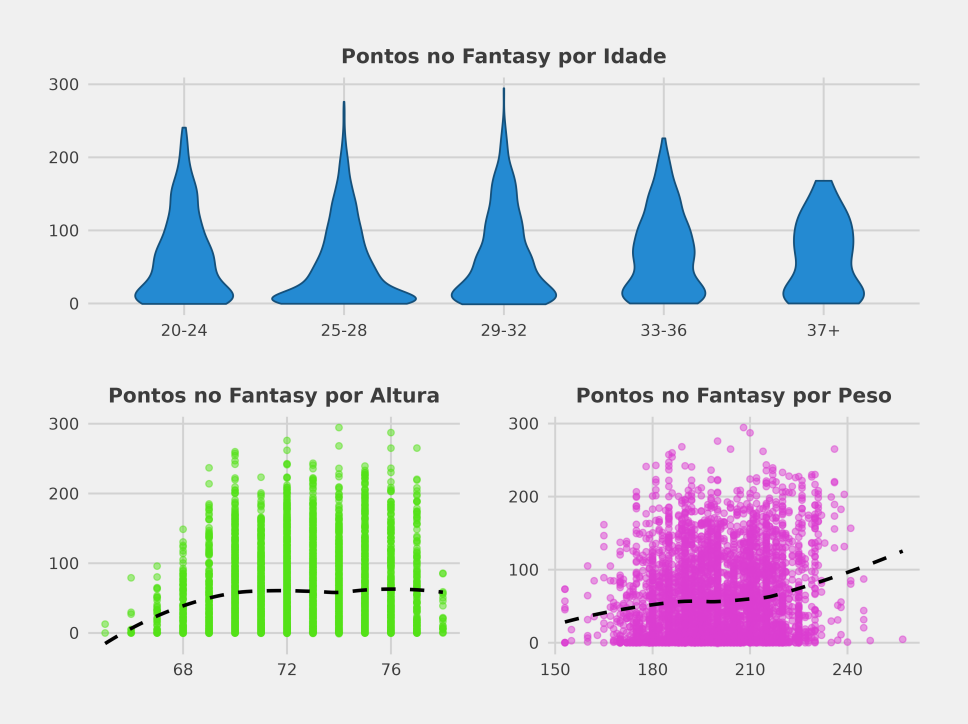

Figura 28 – Distribuição na pontuação no *Fantasy* para as características pessoais dos WRs veteranos da NFL.

Tabela 39 – Análise de correlação para as variáveis de recepção para os WRs veteranos.

	targets	receiving_rec	receiving_yds	receiving_td	receiving_yards_after_catch	receiving_first_downs
targets	1,00	0,77	0,75	0,61	0,64	0,75
receiving_rec	0,77	1,00	0,97	0,81	0,71	0,98
receiving_yds	0,75	0,97	1,00	0,85	0,69	0,98
receiving_td	0,61	0,81	0,85	1,00	0,57	0,84
receiving_yards_after_catch	0,64	0,71	0,69	0,57	1,00	0,68
receiving_first_downs	0,75	0,98	0,98	0,84	0,68	1,00

se estável a partir desse ponto. À direita, o gráfico de peso revela uma tendência crescente ao longo de quase toda a sua amplitude, com uma faixa constante de desempenho entre 180 e 210 libras.

Na análise de correlação, foram consideradas apenas as variáveis relacionadas à recepção, incluindo `targets`, `receiving_yards_after_catch` e `receiving_first_downs`, além daquelas previamente utilizadas na análise dos calouros. Ao final do processo, as variáveis `receiving_rec` e `receiving_yards` foram removidas por apresentarem alta correlação com outras covariáveis, conforme demonstrado na Tabela 39.

4.3.2.1 Regressão

Os resultados para os modelos propostos de regressão estão apresentados na Tabela 40. Observa-se que o modelo *XGBoost* obteve desempenho superior ao *Random Forest*, sendo, portanto, o modelo selecionado para esta etapa da análise.

Foram selecionadas seis variáveis explicativas: `games`, `age`, `receiving_first_downs`, `receiving_yards_after_catch`, `targets` e `receiving_td`. A Figura 29 apresenta o grá-

Tabela 40 – Métricas de avaliação para os modelos candidatos para a regressão de WRs veteranos.

Modelo	MAE	N° de Variáveis
modelo_rf	31,01	9
modelo_rf_rfe	30,96	8
modelo_xg	30,86	9
modelo_xg_rfe	**30,87**	**6**

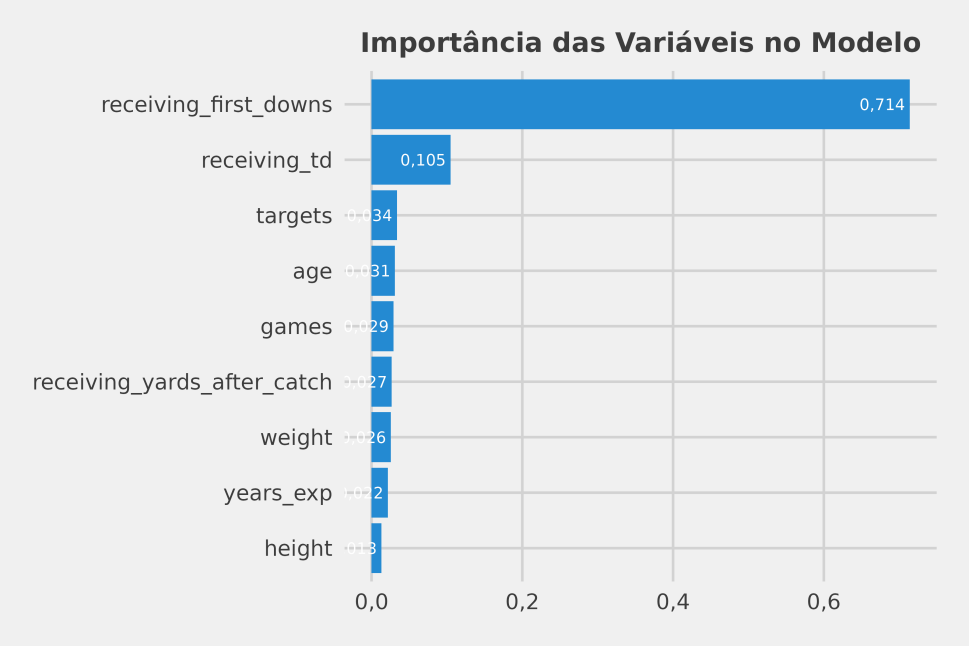

Figura 29 – Importância relativa pela métrica *Gain* das variáveis explicativas do modelo de regressão *XGBoost* para os WRs veteranos.

Tabela 41 – Hiperparâmetros do modelo de regressão *XGBoost* de WRs veteranos.

Hiperparâmetro	Valor
nrounds	51,00
verbose	0,00
max_depth	4,00
eta	0,08
subsample	0,75
colsample_bytree	0,90

fico de importância relativa dessas covariáveis, no qual se observa que `receiving_first_downs` e `receiving_td` se destacam como as mais relevantes para o modelo, com ênfase particular na primeira, que apresenta valor expressivamente superior em relação às demais.

Os hiperparâmetros ajustados para o modelo estão descritos na Tabela 41.

A análise de resíduos é apresentada na Figura 30. Observa-se que a distribuição dos resíduos se aproxima de uma distribuição normal; no entanto, verifica-se novamente uma tendência de subestimação dos valores observados, o que resulta em *outliers* com

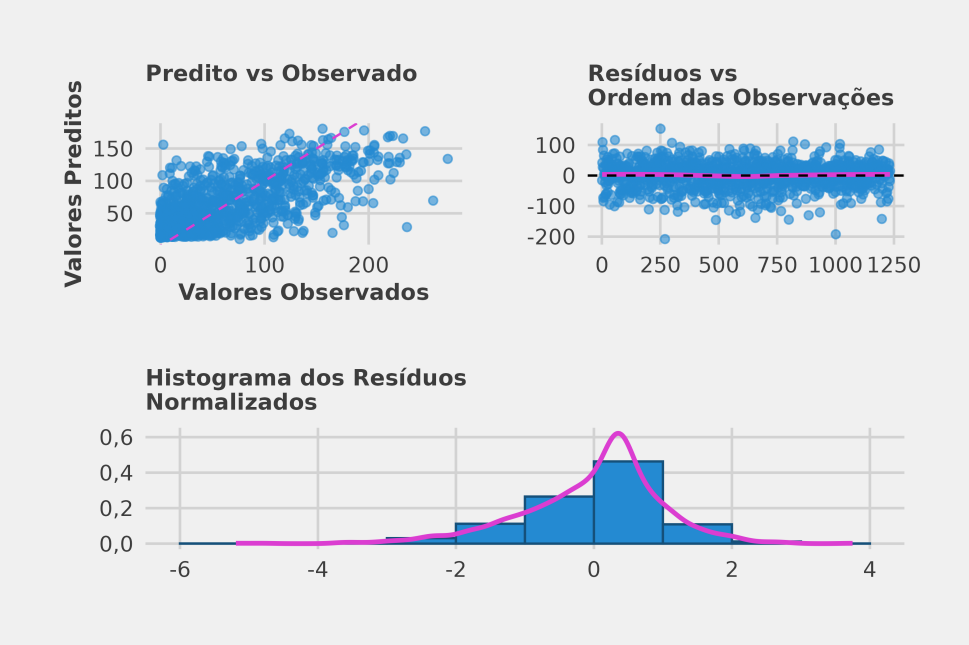

Figura 30 – Gráficos de avaliação dos resíduos no modelo de regressão dos WRs veteranos.

Tabela 42 – Principais *outliers* presentes no modelo de regressão de WRs veteranos.

Jogador	Ano	Pontuação Observada	Pontuação Predita	Resíduo Absoluto
Steve Smith	2005	236,80	28,80	208,00
Deebo Samuel	2021	261,96	69,63	192,33

magnitude superior à esperada. A Tabela 42 explora em detalhe os dois maiores valores identificados.

Os casos de *outliers* que mais se destacaram foram dos atletas *Steve Smith*, em 2005, e *Deebo Samuel*, em 2021. Ambos os jogadores haviam sofrido lesões nas temporadas anteriores, o que levou o modelo a subestimar suas pontuações esperadas, uma vez que as estimativas foram influenciadas por desempenhos anteriores comprometidos (Pro Football Reference, 2025x; Pro Football Reference, 2025h).

4.4 Tight Ends

Os *Tight Ends* desempenham funções semelhantes às dos *Wide Receivers*, atuando como alvos do *Quarterback* no jogo aéreo. No entanto, diferenciam-se por sua função adicional como bloqueadores, sendo frequentemente utilizados tanto em jogadas de passe (protegendo o QB contra a pressão adversária), quanto em jogadas de corrida (auxiliando na abertura de espaços para os *Running Backs*). Em razão dessas exigências físicas, os TEs tendem a ser jogadores mais altos e fortes, embora geralmente menos ágeis e mais lentos do que os WRs. Como consequência, sua produção no *Fantasy Football* costuma ser inferior.

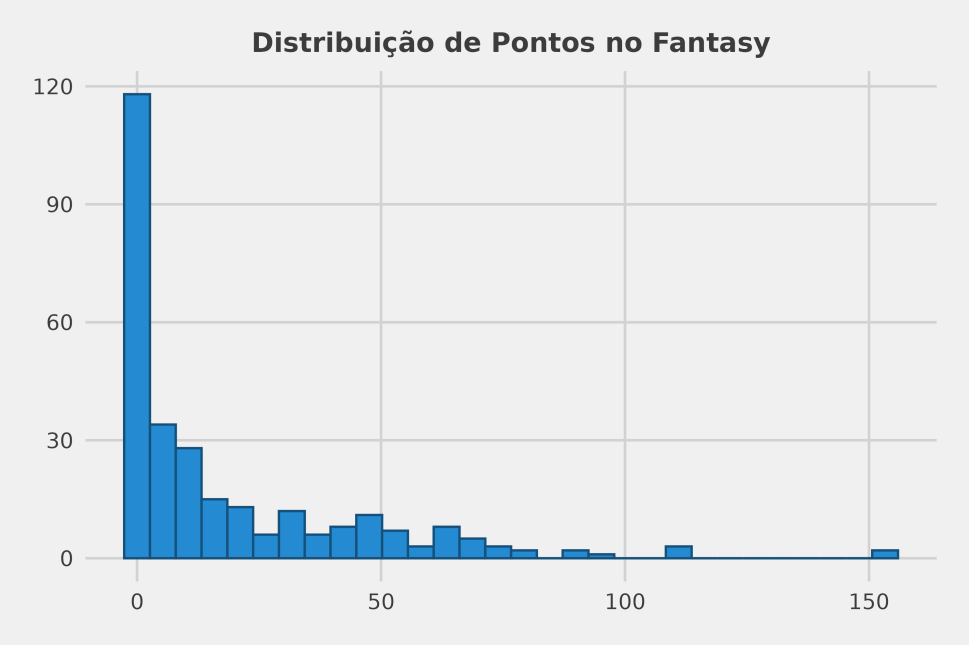

Figura 31 – Distribuição da pontuação no *Fantasy* para *Tight Ends* em seu primeiro ano na NFL.

Diante dessas particularidades, o número de TEs relevantes para o *Fantasy Football* é mais limitado, o que justifica o fato de que as ligas costumam reservar apenas um espaço por time no time titular para essa posição (NFL, 2021).

Para a modelagem estatística, foram utilizadas as mesmas variáveis consideradas no caso dos WRs, restringindo-se às estatísticas de recepção, além da inclusão das variáveis de características físicas e de contexto dos jogadores.

4.4.1 Calouros

Para essa posição, a transição do *College Football* para o Futebol Americano profissional tende a ser mais lenta (Bleacher Report, 2025), os atletas geralmente demoram um tempo para conseguirem se desenvolver e serem bem aproveitados no *Fantasy*.

Essa dinâmica é possível de ser observada na Figura 31, comparando com as posições anteriores, vemos que os TEs, num geral, pontuam menos que WRs e RBs. Além disso, foram observados 27,5% de valores zerados, sendo necessário, portanto, um modelo de duas etapas.

Os *Tight Ends* calouros apresentam uma dinâmica semelhante às demais posições tanto em relação ao *Draft* quanto à conferência em que o atleta atuava no nível universitário, conforme exposto na Figura 32. Observa-se que a distribuição da pontuação tende a ser menor à medida que o jogador é selecionado em rodadas mais tardias do *Draft*. No entanto, é interessante notar que o grupo atletas oriundos das `Early Rounds` apresentaram pontuações tão elevadas quanto daqueles escolhidos na `1st Round`. Com relação ao grupo de conferência, não foram identificadas diferenças significativas entre os

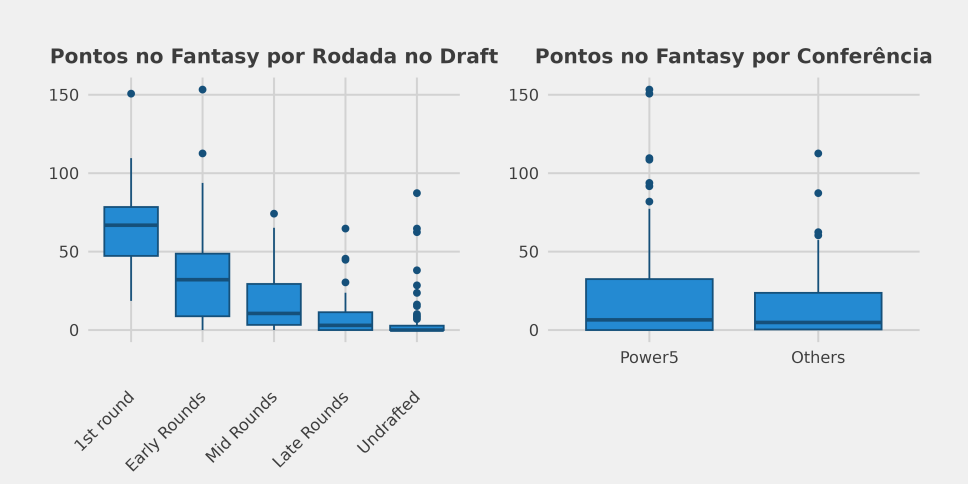

Figura 32 – Distribuição da pontuação no *Fantasy* por rodada em que o *Tight End* foi selecionado no *Draft* da NFL, e de acordo com a conferência do *College Football* ele atuou.

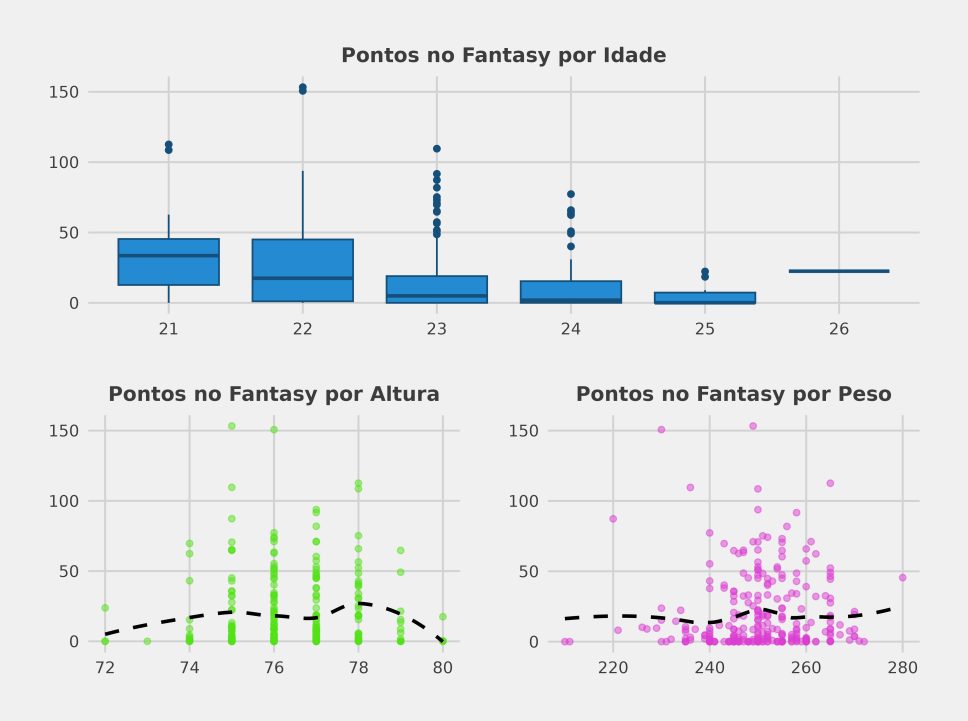

Figura 33 – Distribuição da pontuação no *Fantasy* por características pessoais dos TEs calouros.

dois níveis considerados.

Na Figura 33, observa-se que a relação entre idade e pontuação (gráfico superior) é menos acentuada em comparação com as demais posições, embora atletas mais jovens ainda tendam a apresentar maiores pontuações. Os gráficos inferiores, à esquerda e à direita, mostram as distribuições de altura e peso, respectivamente. Nota-se a presença de picos em torno de 78 polegadas de altura e 250 libras de peso, o que pode sugerir um perfil corporal ideal para jogadores calouros dessa posição.

Tabela 43 – Análise de correlação para as variáveis de recepção para os WRs calouros.

	receiving_rec	receiving_yds	receiving_td
receiving_rec	1,00	0,96	0,75
receiving_yds	0,96	1,00	0,79
receiving_td	0,75	0,79	1,00

Tabela 44 – Métricas de avaliação para os modelos candidatos para a classificação de TEs calouros.

Modelo	Acurácia	N° de Variáveis
modelo_log_logito	0,8003	11
modelo_log_logito_stepwise	0,7802	5
modelo_log_probito	0,8081	11
modelo_log_probito_stepwise	0,7802	5
modelo_rf	0,7517	13
modelo_rf_rfe	0,7541	10
modelo_xg	0,8193	13
modelo_xg_rfe	**0,7976**	**5**

A análise de correlação, apresentada na Tabela 43, foi conduzida com foco nas variáveis de recepção. De acordo com os critérios previamente estabelecidos, a variável `receiving_yds` não atendeu aos requisitos e, portanto, foi excluída das etapas subsequentes de modelagem.

4.4.1.1 Classificação

Os resultados obtidos para a classificação de *Tight Ends* calouros são expostos na Tabela 44. Como resultado temos que o melhor modelo foi o *XGBoost* com 4 variáveis, esse modelo foi o que obteve a melhor acurácia, com o menor número de variáveis.

As variáveis selecionadas para o modelo foram: `draft_round_group` no nível `Undrafted`, `weight`, `receiving_rec` e `receiving_td`. A Figura 34 apresenta a métrica de importância relativa calculada para todas as variáveis inicialmente propostas. Observa-se que as quatro covariáveis selecionadas se destacam significativamente das demais, indicando que, mesmo com um número reduzido de variáveis, o modelo alcança desempenho satisfatório nas métricas de avaliação.

Os hiperparâmetros ajustados para o modelo estão descritos na Tabela 45.

Por fim, a matriz de confusão é apresentada na Tabela 46. O modelo alcançou uma acurácia de 80,5%, com sensibilidade de 81,3% e especificidade de 78,3%. Esses resultados indicam uma boa capacidade preditiva do modelo, tanto para identificar corretamente os jogadores que pontuaram quanto para aqueles que não pontuaram.

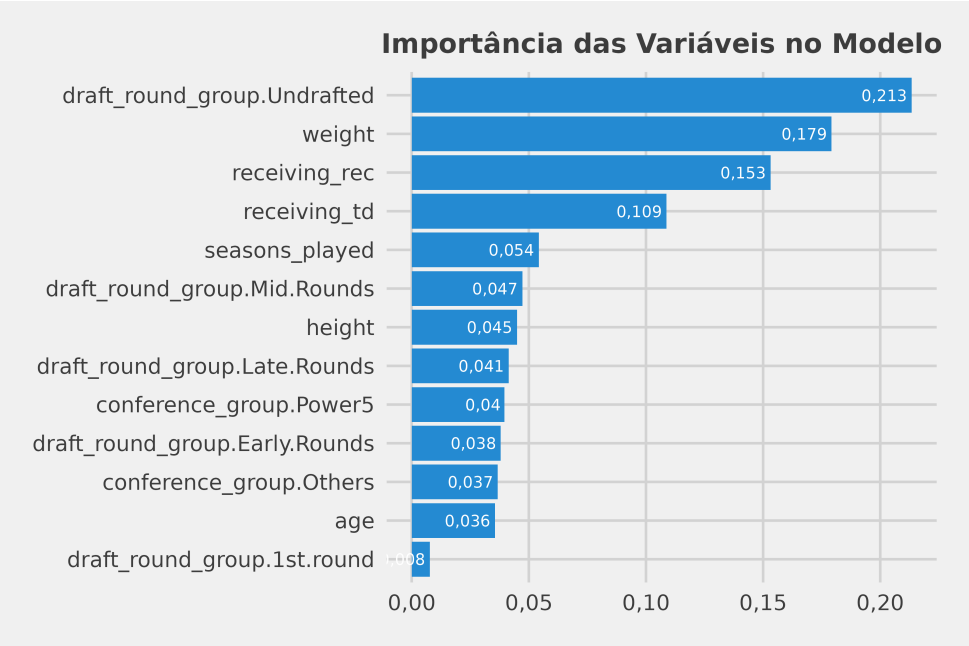

Figura 34 – Importância relativa pela métrica *Gain* das variáveis explicativas do modelo de regressão *XGBoost* para os TEs calouros.

Tabela 45 – Hiperparâmetros do modelo de regressão *XGBoost* de TEs calouros.

Hiperparâmetro	Valor
nrounds	145,00
verbose	0,00
max_depth	4,00
eta	0,02
subsample	0,77
colsample_bytree	0,57

Tabela 46 – Matriz de confusão para o modelo de classificação de TEs calouros.

Predito / Real	Negativo	Positivo
Negativo	18	12
Positivo	5	52

Tabela 47 – Métricas de avaliação para os modelos candidatos para a regressão de TEs calouros.

Modelo	MAE	N° de Variáveis
modelo_gamma_log	20,95	11
modelo_gamma_log_stepwise	18,25	7
modelo_gamma_inverse	18,83	11
modelo_gamma_inverse_stepwise	18,56	8
modelo_rf	17,46	13
modelo_rf_rfe	**17,28**	**12**
modelo_xg	20,06	13
modelo_xg_rfe	19,32	7

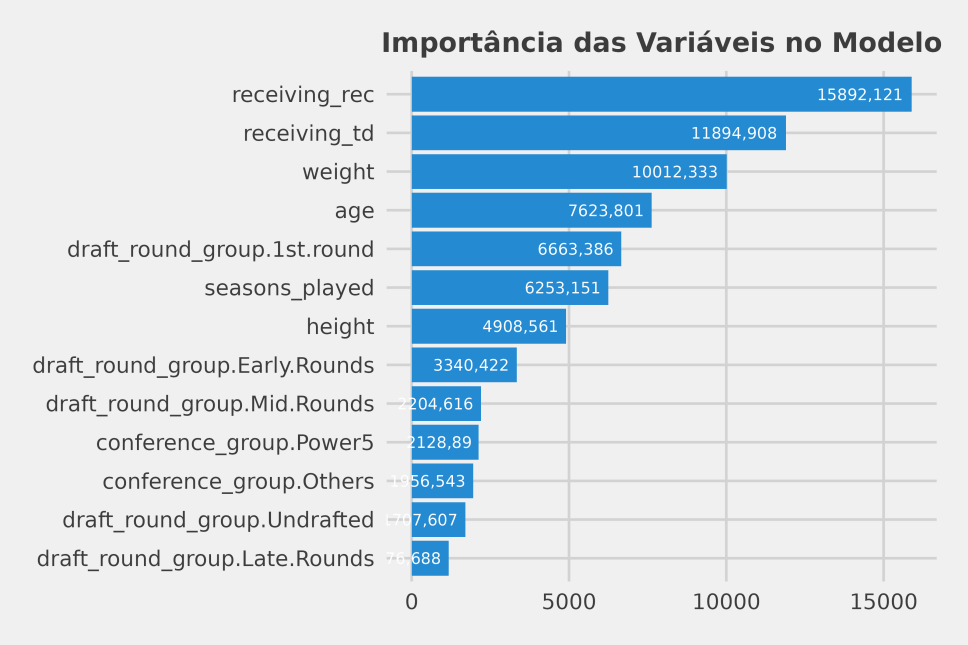

Figura 35 – Importância relativa pela métrica MDI das variáveis explicativas do modelo de regressão *Random Forest* para os TEs calouros.

4.4.1.2 Regressão

Para o modelo de regressão, os resultados obtidos em cada modelagem candidata estão apresentados na Tabela 47. O modelo selecionado foi o *Random Forest*, que incorporou 12 das 13 variáveis candidatas, sendo a única exceção o nível `1st Round` da variável `draft_round_group`, que não foi incluído.

As importâncias relativas de cada covariável utilizadas no modelo estão representadas na Figura 35. A partir da variável `seasons_played` a importância decaí significativamente, mas essa variável e as subsequentes ainda contribuem positivamente para o desempenho do modelo, exceto o nível `1st Round` de `draft_round_group`, que demonstrou uma importância quase nula, e portanto foi removida do modelo final.

Os hiperparâmetros ajustados são apresentados na Tabela 48.

Tabela 48 – Hiperparâmetros do modelo de classificação *Random Forest* de TEs calouros.

Hiperparâmetro	Valor
mtry	2,00
sample.fraction	0,80
replace	0,00
nodesize	1,00
num.trees	750,00

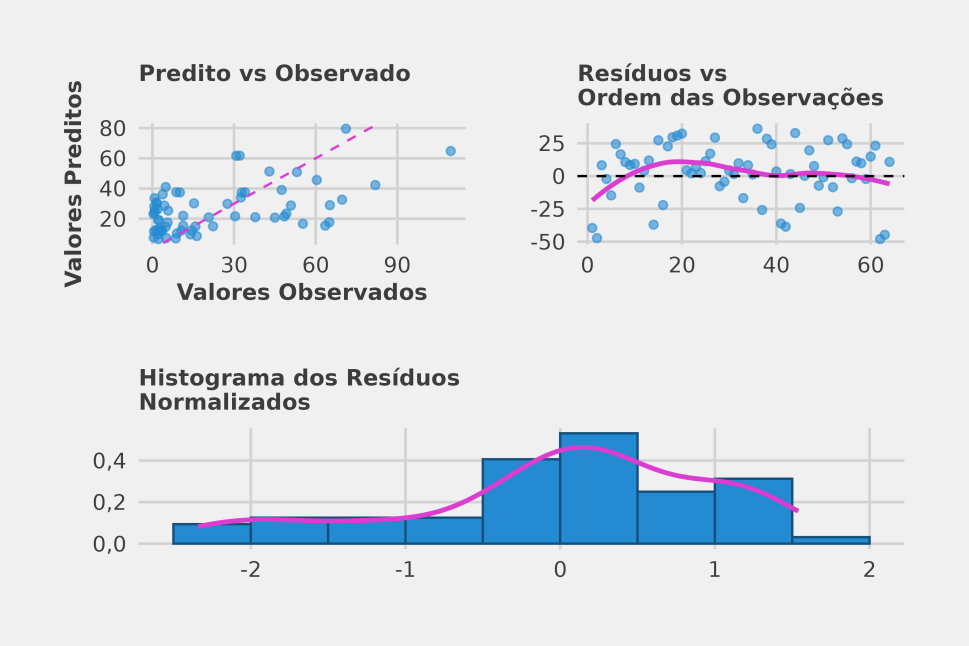

Figura 36 – Gráficos de avaliação dos resíduos no modelo de regressão dos TEs calouros.

O desempenho do modelo final, bem como a distribuição dos resíduos, pode ser observado na Figura 36. A análise, contudo, é limitada pelo número reduzido de observações (64). Ainda assim, verifica-se que a distribuição dos resíduos tende à normalidade, sem a presença de desvios significativos ou *outliers* expressivos.

4.4.2 Veteranos

Para os *Tight Ends* veteranos, o índice de valores zerados foi de 7,6%, não sendo necessário um modelo de duas etapas. A distribuição da resposta é apresentada na Figura 37.

Para esses atletas, as características pessoais são apresentadas na Figura 38. A idade (acima) indica que esta posição tende a ser bastante longeva, com pontuações elevadas sendo registradas até mesmo na faixa dos 33–36 anos. A partir dos 37 anos, no entanto, não há registros de jogadores com mais de 150 pontos. Quanto à altura (linha inferior à esquerda) e ao peso (linha inferior à direita), observa-se uma dinâmica semelhante à dos *Tight Ends* calouros, com picos entre 77 e 78 polegadas de altura e de 250 até 275

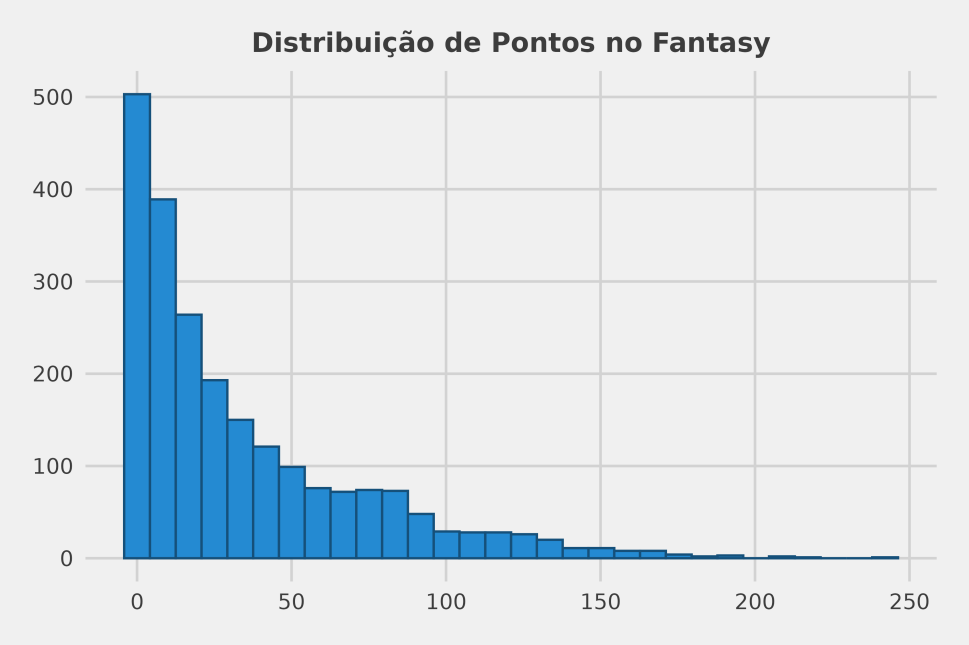

Figura 37 – Distribuição da pontuação no *Fantasy* para TEs veteranos da NFL.

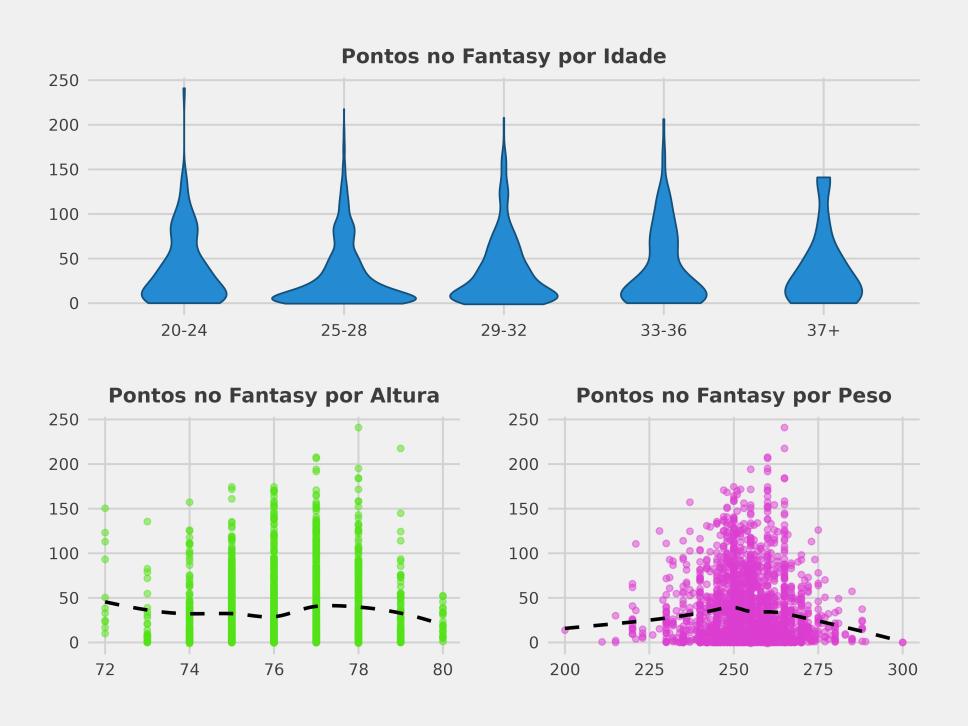

Figura 38 – Distribuição na pontuação no *Fantasy* para as características pessoais dos TEs veteranos da NFL.

libras de peso. Esses padrões reforçam a hipótese da existência de um protótipo físico ideal para atletas dessa posição.

Com a análise de correlação, foi possível remover as variáveis de `receptions` e `receiving_yds`, por apresentarem correlações acima de 0,90, conforme apresentado na Tabela 49.

Tabela 49 – Análise de correlação para as variáveis de recepção para os WRs veteranos.

	targets	receiving_rec	receiving_yds	receiving_td	receiving_yards_after_catch	receiving_first_downs
targets	1,00	0,82	0,81	0,65	0,75	0,80
receiving_rec	0,82	1,00	0,98	0,77	0,80	0,97
receiving_yds	0,81	0,98	1,00	0,79	0,81	0,99
receiving_td	0,65	0,77	0,79	1,00	0,62	0,81
receiving_yards_after_catch	0,75	0,80	0,81	0,62	1,00	0,79
receiving_first_downs	0,80	0,97	0,99	0,81	0,79	1,00

Tabela 50 – Métricas de avaliação para os modelos candidatos para a regressão de TEs veteranos.

Modelo	MAE	N° de Variáveis
modelo_rf	18,73	9
modelo_rf_rfe	18,92	7
modelo_xg	18,65	9
modelo_xg_rfe	**18,85**	**7**

Tabela 51 – Hiperparâmetros do modelo de regressão *XGBoost* de TEs veteranos.

Hiperparâmetro	Valor
nrounds	185,00
verbose	0,00
max_depth	5,00
eta	0,02
subsample	0,76
colsample_bytree	0,67

4.4.2.1 Regressão

Como a classificação não foi necessária, focamos diretamente nos modelos de regressão. A Tabela 50 apresenta as métricas de desempenho para os modelos propostos. Entre os modelos avaliados, o *XGBoost* demonstrou o menor MAE. Por essa razão, optamos por seguir com o modelo *XGBoost* que utiliza 7 variáveis para as análises subsequentes.

As variáveis selecionadas foram: `games`, `age`, `weight`, `receiving_first_downs`, `receiving_yards_after_catch`, `targets` e `receiving_td`. A Figura 39 apresenta o gráfico de importância relativa dessas covariáveis, no qual se observa que receiving_first_downs se destaca significativamente, sendo a principal variável preditiva com ampla vantagem em relação às demais.

Os hiperparâmetros ajustados para o modelo estão descritos na Tabela 51.

Dada a análise dos resíduos apresentada na Figura 40, observa-se que o modelo subestimou algumas predições, resultando em resíduos normalizados que se desviam da distribuição normal esperada. Nota-se esse comportamento no gráfico do histograma, com caudas inferiores mais pesadas, indicando uma maior ocorrência de valores extremos negativos do que o esperado.

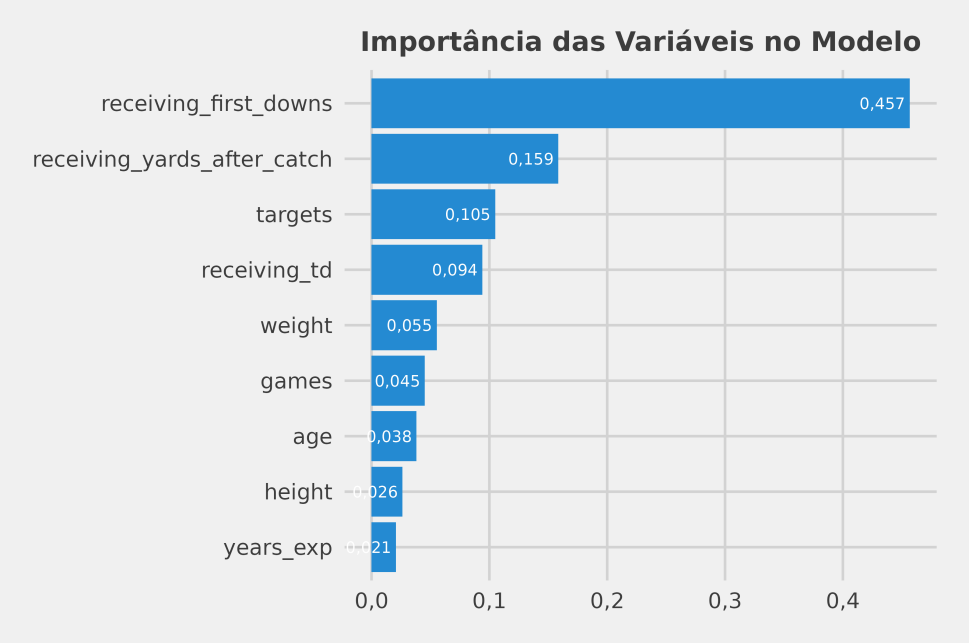

Figura 39 – Importância relativa pela métrica *Gain* das variáveis explicativas do modelo de regressão *XGBoost* para os TEs veteranos.

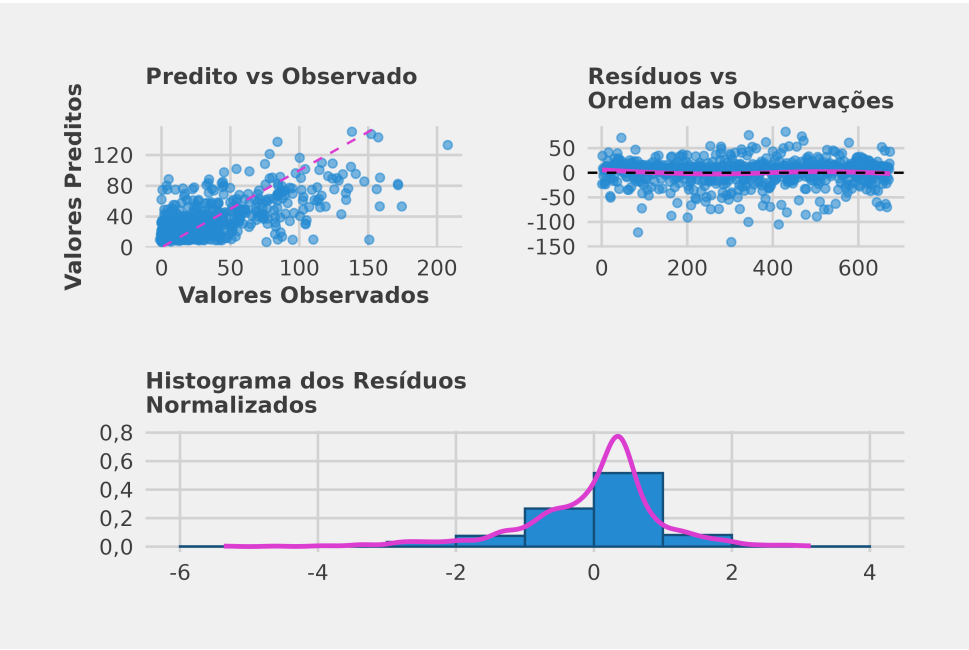

Figura 40 – Gráficos de avaliação dos resíduos no modelo de regressão dos TEs veteranos.

Tabela 52 – Principais *outliers* presentes no modelo de regressão de TEs veteranos.

Jogador	Ano	Pontuação Observada	Pontuação Predita	Resíduo Absoluto
Julius Thomas	2013	150,8	9,86	140,94
Antonio Gates	2004	174,4	53,14	121,26
Rob Gronkowski	2017	158,4	53,26	105,14
Travis Kelce	2014	110,2	10,08	100,12

Investigando as observações *outliers*, expostas na Tabela 52, temos os 4 principais resíduos, que foram os jogadores *Julius Thomas* em 2013, *Antonio Gates* em 2004, *Rob Gronkowski* em 2017 e *Travis Kelce* em 2014.

O primeiro deles, *Julius Thomas*, é um *Tight End* que ingressou na NFL em 2011, mas teve produção bastante limitada em suas duas primeiras temporadas, com apenas nove jogos disputados e uma recepção registrada. No entanto, em 2013, integrou um dos ataques mais produtivos da história da liga, liderado pelo *Quarterback* que estabeleceu os recordes de jardas passadas e *touchdowns* lançados em uma única temporada (ESPN, 2023). Soma-se a isso o fato de *Thomas* apresentar características físicas próximas ao protótipo da posição, altura, peso, agilidade e velocidade superiores à média, o que lhe garantiu protagonismo ofensivo e resultou em uma pontuação significativamente superior àquela estimada pelo modelo (Pro Football Reference, 2025m).

Situação semelhante ocorreu com *Antonio Gates* em 2004 e *Travis Kelce* em 2014. Ambos os atletas não tiveram grande destaque como calouros, mas tornaram-se peças fundamentais em seus respectivos ataques já em suas segundas temporadas. Além disso, ambos estão entre os cinco TEs com maior número de jardas e *touchdowns* recebidos na história da NFL (Pro Football Reference, 2025d; Pro Football Reference, 2025z).

Por fim, *Rob Gronkowski*, sétimo TE com mais jardas e terceiro com mais *touchdowns* na história da liga, sofreu com recorrentes lesões ao longo da carreira (Pro Football Reference, 2025w). Em 2016, atuou em apenas oito partidas, o que levou o modelo a subestimar sua pontuação para a temporada de 2017, quando voltou a produzir em alto nível.

4.5 Análise das Variáveis

Para os jogadores calouros, destacaram-se as variáveis relacionadas ao *Draft* da NFL e às características físicas dos atletas. Nos modelos de classificação, o nível `Undrafted` mostrou-se particularmente relevante na predição de pontuação, sendo possível observar esse padrão nos gráficos de importância das posições de RBs (Figura 14), WRs (Figura 24) e TEs (Figura 34). Já nos modelos de regressão, o nível que mais se sobressaiu foi o `1st Round`.

Entre as variáveis estatísticas de jogo, os resultados foram mais variados conforme

a posição analisada. No conjunto de variáveis de recepção, a variável `receiving_rec` apresentou maior destaque, estando presente em todos os modelos aplicados aos calouros em que foi testada, exceto no modelo de regressão para WRs, sempre com altos índices de importância. A variável `receiving_td` também teve relevância considerável. Quanto às variáveis de corrida, utilizadas apenas nos modelos de RBs e QBs, a principal foi `rushing_car`, embora seu desempenho não tenha sido tão expressivo quanto o das variáveis de recepção.

No caso dos jogadores veteranos, as características físicas perderam relevância, dando lugar às estatísticas de desempenho em campo. A única exceção foi a variável `age`, que esteve presente em todos os modelos, exceto no de regressão para QBs.

Para os QBs veteranos, os modelos selecionaram tanto variáveis de passe quanto de corrida, um comportamento distinto dos calouros, cujos modelos priorizaram apenas variáveis de passe. Entre os RBs, as variáveis de corrida demonstraram maior impacto que as de recepção, o que está de acordo com o esperado. Já para WRs e TEs, as variáveis selecionadas foram bastante semelhantes, com a principal diferença sendo a inclusão da variável `weight` nos modelos dos TEs. Em ambas as posições, a variável com maior importância relativa foi `receiving_first_downs`.

5 Considerações Finais

Todas as etapas deste trabalho revelaram-se bastante complexas, desde a fase de extração e processamento dos dados até a análise e modelagem estatística. Logo no início, os ajustes necessários após a extração apresentaram desafios adicionais, uma vez que parte dos dados encontrava-se incompleta, exigindo correções manuais e um elevado grau de atenção e detalhamento.

A etapa de modelagem também se mostrou desafiadora, considerando que cada posição apresenta dinâmicas e variáveis específicas. Foi necessário adaptar cada modelo de forma a capturar as particularidades de cada função em campo. Algumas variáveis, embora apresentassem comportamentos similares entre posições, possuíam magnitudes e influências distintas, o que inviabilizou a construção de um modelo único e geral. Em razão disso, as interpretações, embora recorrentes, precisaram ser tratadas separadamente.

Outro aspecto relevante refere-se à multiplicidade de modelos considerados. Isso limitou a profundidade da análise de cada um, o que, em alguns casos, resultou em desempenhos aquém do esperado. Ainda assim, essa abordagem amplia as possibilidades para investigações futuras, servindo como base para estudos mais aprofundados por posição.

Um fator que impactou significativamente a análise foi a presença de inflação de zeros. Tanto nos modelos para jogadores calouros quanto nos modelos para *Quarterbacks* veteranos, foi necessária a aplicação de uma abordagem de duas etapas. De modo geral, os modelos de classificação apresentaram bom desempenho na predição de jogadores que pontuaram, mas desempenho regular ou inferior para os que não pontuaram. Como sugestão, o ajuste do *threshold* poderia ser explorado visando otimizar métricas que equilibrem melhor sensibilidade e especificidade.

No que diz respeito aos modelos de regressão, os resultados também se mostraram promissores. A principal dificuldade enfrentada foi a assimetria da variável resposta, concentrada em valores próximos de zero. Ainda assim, a maioria dos modelos apresentou resíduos bem comportados. Uma limitação recorrente foi a subestimação da pontuação de jogadores de elite, sobretudo nas temporadas em que eles despontam, bem como nos casos de atletas que retornam de lesão em plena forma. Para mitigar essas limitações, recomenda-se a inclusão de variáveis relacionadas ao desempenho físico, como as medições do *NFL Combine* [1], com o intuito de identificar jogadores com potencial produtivo ainda não manifestado em temporadas anteriores. Além disso, o uso de informações sobre o histórico completo de desempenho e lesões poderia aprimorar a identificação de atletas com alto risco de inatividade ou, ao contrário, com bom retrospecto que justifique projeções mais altas.

Como sugestão final para estudos futuros, propõe-se a adoção de técnicas adicionais voltadas à predição, além do uso de *Random Forest* e *XGBoost*, bem como a concentração

[1] Evento anual em que atletas elegíveis ao *Draft* da NFL são avaliados em testes físicos e técnicos.

dos esforços analíticos em uma única posição. Tal foco pode favorecer análises mais detalhadas e resultados com maior acurácia.

Referências

ABADZIC, A.; CHEUN, J.; PATEL, M. Data analysis on predicting the top 12 fantasy football players by position. *SMU Data Science Review*, v. 8, n. 2, p. 7, 2024.

ALMEIDA, R. B. d.; ALMEIDA, V. M. C. d.; LIMA, D. d. F. P. Comunidades de marca de fantasy sports games: Identificação, engajamento, intenção de continuidade e valor da marca do patrocinador. *ReMark - Revista Brasileira de Marketing*, v. 14, n. 1, p. 33–48, abr. 2015. Disponível em: <https://periodicos.uninove.br/remark/article/view/12074>.

BISCHL, B.; LANG, M.; KOTTHOFF, L.; SCHIFFNER, J.; RICHTER, J.; STUDERUS, E.; CASALICCHIO, G.; JONES, Z. M. mlr: Machine learning in r. *Journal of Machine Learning Research*, v. 17, n. 170, p. 1–5, 2016. Disponível em: <https://jmlr.org/papers/v17/15-066.html>.

Bleacher Report. *What Makes an Ideal WR Corps in Today's NFL?* [S.l.], 2022. Acesso em 06 de Jun de 2025. Disponível em: <https://bleacherreport.com/articles/10041374-what-makes-an-ideal-wr-corps-in-todays-nfl>.

Bleacher Report. *Non-1st-Round WRs Who Could Erupt as NFL Rookies*. [S.l.], 2024. Acesso em 06 de Jun de 2025. Disponível em: <https://bleacherreport.com/articles/10125109-non-1st-round-wrs-who-could-erupt-as-nfl-rookies>.

Bleacher Report. *Ranking Non-First-Round Rookie TEs Most Likely to Be Breakout Weapons in 2025*. [S.l.], 2025. Acesso em 06 de Jun de 2025. Disponível em: <https://bleacherreport.com/articles/25205322-ranking-non-first-round-rookie-tes-most-likely-be-breakout-weapons-2025>.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, p. 123–140, 1996.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, p. 5–32, 2001.

BREIMAN, L.; FRIEDMAN, J.; STONE, C.; OLSHEN, R. *Classification and Regression Trees*. Taylor & Francis, 1984. ISBN 9780412048418. Disponível em: <https://books.google.com.br/books?id=JwQx-WOmSyQC>.

BROOKS, B. *Ranking each position's importance, from quarterback to returner*. [S.l.], 2015. Acesso em 20 de Mai de 2025. Disponível em: <https://www.nfl.com/news/ranking-each-position-s-importance-from-quarterback-to-returner-0ap3000000503855>.

CARL, S.; BALDWIN, B. *nflfastR: Functions to Efficiently Access NFL Play by Play Data*. [S.l.], 2024. R package version 5.0.0. Disponível em: <https://CRAN.R-project.org/package=nflfastR>.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794.

CRAGG, J. G. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, [Wiley, Econometric Society], v. 39, n. 5, p. 829–844, 1971.

DBeaver Community. *DBeaver – Universal Database Tool*. [S.l.], 2025. Versão 25.0.5.

ESPN. *The 2013 Broncos scored an NFL-record 606 points ... and have been forgotten.* [S.l.], 2023. Acesso em 06 de Jun de 2025. Disponível em: <https://www.espn.com/nfl/story/_/id/38492638/denver-broncos-peyton-manning-nfl-record-606-points-2013-season>.

ESPN. *Rookie NFL running back role tiers: Stacking 25 draft picks.* [S.l.], 2025. Acesso em 06 de Jun de 2025. Disponível em: <https://www.espn.com/nfl/story/_/id/45008218/2025-nfl-draft-rookie-running-backs-role-tiers-jeanty-hampton-skattebo>.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001.

GE. *Calendário do futebol brasileiro em 2025: veja as datas.* Rio de Janeiro, 2024. Acesso em 10 de Mar de 2025. Disponível em: <https://ge.globo.com/futebol/noticia/2024/11/12/calendario-do-futebol-brasileiro-em-2025-veja-as-datas.ghtml>.

GE. *Calendário inchado: Jogadores de City e Real Madrid podem bater 79 partidas na temporada.* Rio de Janeiro, 2024. Acesso em 10 de Mar de 2025. Disponível em: <https://ge.globo.com/futebol/futebol-internacional/noticia/2024/09/20/calendario-inchado-jogadores-de-city-e-real-madrid-podem-bater-79-partidas-na-temporada.ghtml>.

GILANI, S.; EASWARAN, A.; LEE, J.; HESS, E. *cfbfastR: The SportsDataverse's R Package for College Football Data.* [S.l.], 2021. R package version 1.9.5. Disponível em: <https://cfbfastR.sportsdataverse.org/>.

GROTHAUS, M. How next-gen data analytics is changing american football. *Knowable Magazine*, 2024. Acesso em 14 de Mai de 2025. Disponível em: <https://knowablemagazine.org/content/article/technology/2024/how-next-gen-data-analytics-changing-american-football>.

GUYON, I.; WESTON, J.; BARNHILL, S.; VAPNIK, V. Gene selection for cancer classification using support vector machines. *Machine learning*, Springer, v. 46, p. 389–422, 2002.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning.* [S.l.], 2009.

HO, T.; CARL, S. *nflreadr: Download 'nflverse' Data.* [S.l.], 2024. R package version 1.4.1. Disponível em: <https://CRAN.R-project.org/package=nflreadr>.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An introduction to statistical learning.* [S.l.]: Springer, 2013. v. 112.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, v. 30, 2017.

LIASHCHYNSKYI, P.; LIASHCHYNSKYI, P. Grid search, random search, genetic algorithm: a big comparison for nas. *arXiv preprint arXiv:1912.06059*, 2019.

LUTZ, R. W. Fantasy football prediction. *CoRR*, abs/1505.06918, 2015. Disponível em: <http://arxiv.org/abs/1505.06918>.

MLB. *Baseball Basics*. Nova York, 2025. Acesso em 7 de Mar de 2025. Disponível em: <https://www.mlb.com/baseball-basics>.

MORGAN, C. D.; RODRIGUEZ, C.; MACVITTIE, K.; SLATER, R.; ENGELS, D. W. Identifying undervalued players in fantasy football. *SMU Data Science Review*, v. 2, n. 2, p. 14, 2019.

MORGAN, J. N.; SONQUIST, J. A. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, ASA Website, v. 58, n. 302, p. 415–434, 1963. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500855>.

MULHOLLAND, J.; JENSEN, S. T. Projecting the draft and nfl performance of wide receiver and tight end prospects. *CHANCE*, ASA Website, v. 29, n. 4, p. 24–31, 2016. Disponível em: <https://doi.org/10.1080/09332480.2016.1263095>.

NBA. *About the NBA*. Nova York, 2025. Acesso em 7 de Mar de 2025. Disponível em: <https://www.nba.com/news/about>.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, Oxford University Press, v. 135, n. 3, p. 370–384, 1972.

NFL. *NFL fantasy football: Anatomy of an elite fantasy football tight end*. [S.l.], 2021. Acesso em 06 de Jun de 2025. Disponível em: <https://www.nfl.com/news/nfl-fantasy-football-anatomy-of-an-elite-fantasy-football-tight-end>.

NFL. *Official Rules – NFL Fantasy 2024*. [S.l.], 2024. Acesso em 14 de Mai de 2025. Disponível em: <https://static.www.nfl.com/image/upload/v1715009484/league/apps/fantasy/media/rules/OfficialRules2024.pdf>.

NFL. *Creating the NFL Schedule*. Nova York, 2025. Acesso em 7 de Mar de 2025. Disponível em: <https://operations.nfl.com/gameday/nfl-schedule/creating-the-nfl-schedule/>.

NFL. *Formations 101*. Nova York, 2025. Acesso em 7 de Mar de 2025. Disponível em: <https://operations.nfl.com/learn-the-game/nfl-basics/formations-101/>.

NFL. *Rookies Guide*. Nova York, 2025. Acesso em 7 de Mar de 2025. Disponível em: <https://operations.nfl.com/learn-the-game/nfl-basics/rookies-guide/>.

NG, V. K.; CRIBBIE, R. A. Using the gamma generalized linear model for modeling continuous, skewed and heteroscedastic outcomes in psychology. *Current Psychology*, Springer, v. 36, n. 2, p. 225–235, 2017.

Oracle Corporation. *MySQL: The world's most popular open-source database*. [S.l.], 2025. Versão 8.0.42-0ubuntu0.22.04.1.

Pro Football Focus. *Fantasy Football: Studying running back utilization*. [S.l.], 2024. Acesso em 06 de Jun de 2025. Disponível em: <https://www.pff.com/news/fantasy-football-studying-running-back-utilization>.

Pro Football Reference. *Aaron Rodgers Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/R/RodgAa00.htm>.

Pro Football Reference. *Adrian Peterson Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/P/PeteAd01.htm>.

Pro Football Reference. *Alvin Kamara Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/K/KamaAl00.htm>.

Pro Football Reference. *Antonio Gates Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/G/GateAn00.htm>.

Pro Football Reference. *Bo Nix Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/N/NixxBo00.htm>.

Pro Football Reference. *Chris Johnson Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/J/JohnCh04.htm>.

Pro Football Reference. *Daunte Culpepper Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/C/CulpDa00.htm>.

Pro Football Reference. *Deebo Samuel Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/S/SamuDe00.htm>.

Pro Football Reference. *Jamarcus Russell Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/R/RussJa00.htm>.

Pro Football Reference. *Jameis Winston Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/W/WinsJa00.htm>.

Pro Football Reference. *Jonathan Taylor Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/T/TaylJo02.htm>.

Pro Football Reference. *Jordan Love Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/L/LoveJo03.htm>.

Pro Football Reference. *Julius Thomas Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/T/ThomJu00.htm>.

Pro Football Reference. *Larry Johnson Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/J/JohnLa00.htm>.

Pro Football Reference. *Matt Forte Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/F/FortMa00.htm>.

Pro Football Reference. *Michael Penix Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/P/PeniMi00.htm>.

Pro Football Reference. *Michael Turner Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/T/TurnMi00.htm>.

Pro Football Reference. *Patrick Mahomes Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/P/PeniMi00.htm>.

Pro Football Reference. *Patrick Mahomes Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/S/SmitAl03.htm>.

Pro Football Reference. *Phillip Lindsay Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/L/LindPh00.htm>.

Pro Football Reference. *Puka Nacua Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/N/NacuPu00.htm>.

Pro Football Reference. *Rashard Mendenhall Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/M/MendRa00.htm>.

Pro Football Reference. *Rob Gronkowski Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/G/GronRo00.htm>.

Pro Football Reference. *Steve Smith Sr. Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/S/SmitSt01.htm>.

Pro Football Reference. *Tom Brady Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/B/BradTo00.htm>.

Pro Football Reference. *Travis Kelce Stats, Height, Weight, Position, Draft, College*. [S.l.], 2025. Acesso em 21 de Mai de 2025. Disponível em: <https://www.pro-football-reference.com/players/K/KelcTr00.htm>.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A.-L. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, Wiley Online Library, v. 9, n. 3, p. e1301, 2019.

QUINLAN, J. R. Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research*, v. 4, p. 77–90, 1996.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2025. Disponível em: <https://www.R-project.org/>.

ROŽANEC, J. M.; PETELIN, G.; COSTA, J.; CERAR, G.; BERTALANIČ, B.; GUČEK, M.; PAPA, G.; MLADENIĆ, D. Dealing with zero-inflated data: Achieving state-of-the-art with a two-fold machine learning approach. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 149, p. 110339, 2025.

Schneider Downs. *Data Analytics and the NFL Draft*. [S.l.], 2023. Acesso em 17 de Mai de 2025. Disponível em: <https://schneiderdowns.com/our-thoughts-on/data-analytics-and-the-nfl-draft/>.

SMOLA, J. *Things I Learned Doing Fantasy Football Projections*. [S.l.], 2025. Acesso em 20 de Mai de 2025. Disponível em: <https://www.draftsharks.com/article/things-i-learned-doing-fantasy-football-projections>.

WICKHAM, H.; AVERICK, M.; BRYAN, J.; CHANG, W.; MCGOWAN, L. D.; FRANçOIS, R.; GROLEMUND, G.; HAYES, A.; HENRY, L.; HESTER, J.; KUHN, M.; PEDERSEN, T. L.; MILLER, E.; BACHE, S. M.; MüLLER, K.; OOMS, J.; ROBINSON, D.; SEIDEL, D. P.; SPINU, V.; TAKAHASHI, K.; VAUGHAN, D.; WILKE, C.; WOO, K.; YUTANI, H. Welcome to the tidyverse. *Journal of Open Source Software*, v. 4, n. 43, p. 1686, 2019.

WRIGHT, M. N.; ZIEGLER, A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, v. 77, n. 1, p. 1–17, 2017.