Previsão de valores de transferência de jogadores de futebol

Gabriel Fialho, Leonardo Bhering, Mariano Fernandes, Natan Ventura

26 de julho de 2025

1 Introdução

O mercado de transferências de jogadores de futebol está em constante evolução, movimentando milhões de euros e gerando impactos significativos no desempenho esportivo dos clubes. Além disso, é um assunto bastante abordado pelos torcedores e pela mídia no futebol. Tendo isso em vista, propomos um modelo com o objetivo de avaliar o preço de jogadores, o que permitiria a um clube tomar melhores decisões na contratação, compra e venda de jogadores. Neste trabalho, utilizaremos dados das cinco principais ligas europeias como base para o modelo.

Hipóteses do projeto:

- O valor de mercado de um jogador está diretamente relacionado com sua performance em campo.
- Jogadores mais novos têm um maior potencial de crescimento e, portanto, valores de mercado mais altos.
- Jogadores que atuam em posições mais ofensivas, como atacantes, tendem a ter valores de mercado mais altos.
- Jogadores mais velhos tendem a diminuir a produtividade e estar mais suscetíveis a lesões, tendo, portanto, valores de mercado mais baixos.

A partir do estudo, foram analisadas, por posição, quais as principais característica/estatísticas que são mais relevantes para determinar o valor de um jogador de futebol. Dessa forma, conseguimos descobrir o que mais influencia o valor de venda de um jogador e , assim, construir um modelo que, dado as entradas necessárias (como as estatísticas dentro de jogo, além de idade, altura e outros), seja capaz de predizer o valor de mercado do jogador.

2 Revisão de literatura

Desde a decisão do tema do trabalho até o término do seu desenvolvimento, foram lidos, discutidos e utilizados como referência os seguintes artigos:

- "Modelling football players values on the transfer market and their determinants using robust regression models"[1];
- "When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community"[2] ;
- "Determinants of football players' valuation: A systematic review"[3] e "Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques"[4].

No geral, os trabalhos nessa área utilizam dados de performance de jogador, além de outros extracampos como idade, altura, opinião do público ou/e de jornalistas, performance do time, qual liga o jogador joga e até o ranqueamento do jogador em jogos eletrônicos. Alguns deles utilizam os valores pelos quais os jogadores foram vendidos (como foi feito nesse projeto) para treinar o modelo, e outros utilizam o valor de mercado gerado pelo transfermarket[5]. Por fim, em relação à análise das estatísticas disponíveis no FBREF[7] e transfermarkt[5] mais relevantes que influenciam o valor do jogador, foram obtidas conclusões muito semelhantes aos artigos citados.

3 Metodologia de pesquisa

3.1 Obtenção dos dados

Os dados foram obtidos a partir de técnicas de web scraping no transfermarkt[5] e a utilização da biblioteca soccerdata[6] para obter os dados no FBREF[7] . Foram obtidos dados de transferências do transfermarkt[5] dos anos de 2000 a 2023, e dados de estatísticas de jogadores no FBREF[7] de 2017 até 2023. O transfermarket[5] foi escolhido pois ele possui os valores pelos quais os jogadores foram vendidos, dado esse que usamos para construir nosso modelo, já que queremos encontrar as correlações entre estatísticas de jogadores e outros fatores extracampo (como idade e altura) e ver como isso afeta o valor pelo qual o jogador foi vendido. Como foi dito antes, para construir esse modelo precisamos de diversas estatísticas da performance do jogador em campo, por esse motivo, foi utilizado o FBREF[7],

que é uma fonte de dados bem rica, possuindo diversas estatísticas sobre a performance do jogador.

3.2 Análise exploratória dos dados

Após a obtenção dos dados, o próximo passo foi realizar ajustes e limpeza nas tabelas obtidas, pois a tabela do FBREF[7] contém um modelo diferente do tranfermarkt[5], com colunas multi-index, o que dificulta o trabalho com esses dados. Depois de realizar os ajustes das tabelas, a etapa seguinte foi fazer uma análise exploratória dos dados. Nessa etapa, foram plotados alguns gráficos com o objetivo de observar o comportamento de algumas features e iniciar o processo de testes e descoberta de relações entre features e as taxas de transferências dos jogadores.

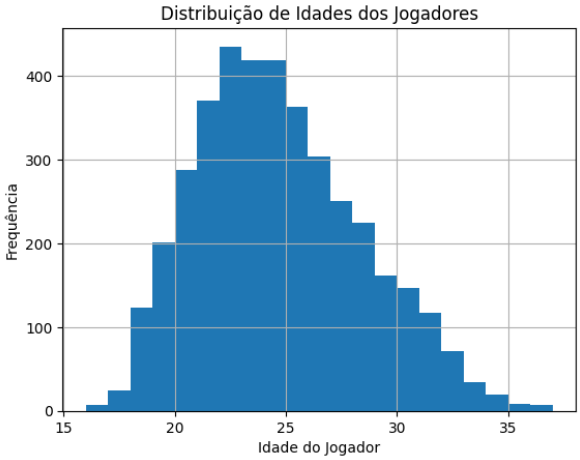

Figura 1: Distribuição de idade das transferências

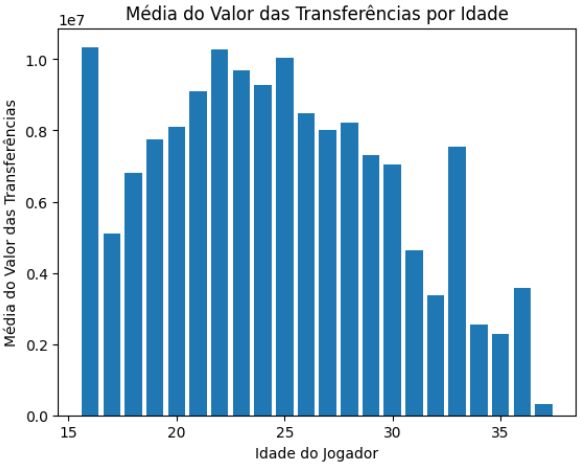

Figura 2: Média por idade

No histograma da figura 1, é possível notar uma distribuição normal nas frequências das idades dos jogadores nas transferências. Dessa forma, foi decidido realizar uma binarização da variável idade, com 5 intervalos. A função qcut da biblioteca pandas foi utilizada para realizar essa segmentação dos dados de idade em 5 intervalos de quantis de tamanhos próximos. Isso garante uma melhoria no modelo, já que a relação entre a idade e o valor de um jogador também é uma relação normal, como é evidenciado na figura 2

Após a binarização da coluna de idades, foram obtidos os seguintes intervalos de idade:

- 16 a 21 anos
- 21 a 23 anos
- 23 a 25 anos
- 25 a 27 anos
- 27 a 37 anos

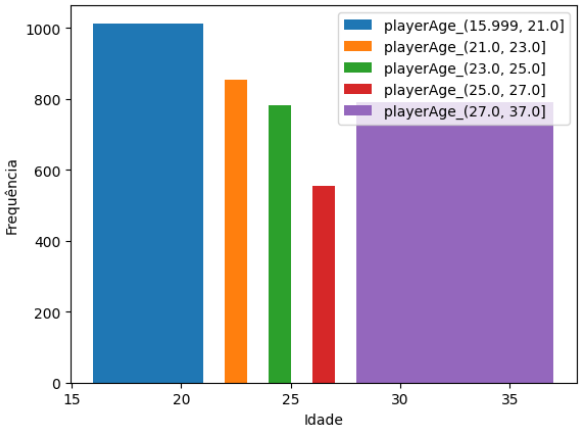

Figura 3: Percentis de idade

O histograma da figura 3 mostra a distribuição de idades após a binarização das colunas

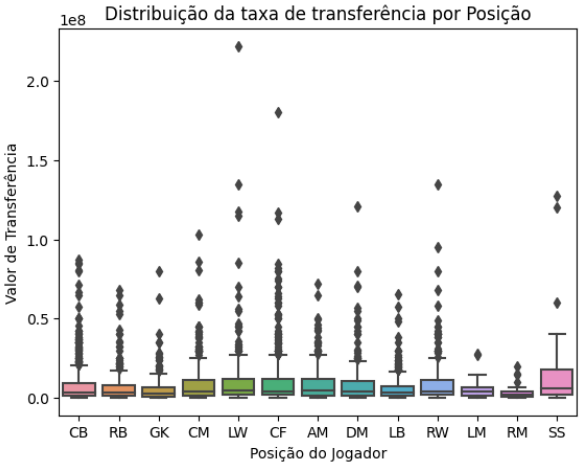

Figura 4: Boxplot de taxas de transferência

A presença de muitos pontos fora dos boxes (outliers) na figura 4 indica que existem transferências com valores excepcionalmente altos em algumas posições específicas. Esses pontos representam transferências muito caras em relação à média das transferências para a respectiva posição.

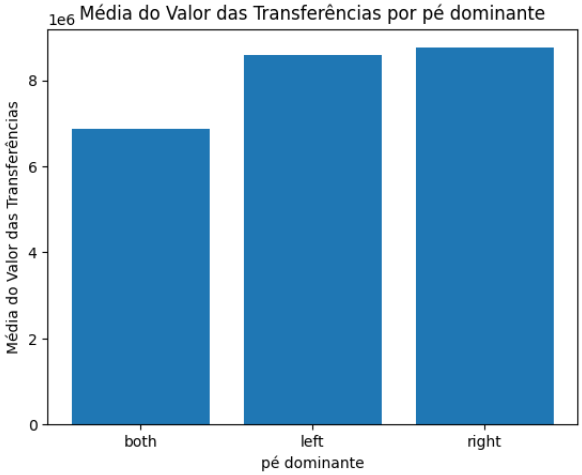

Figura 5: Taxa de transferência por pé dominante

A análise do pé dominante em relação à média de taxa de transferências foi um ponto surpreendente, pois a relação aparente entre um jogador ser ambidestro e o seu valor é negativa. Esse resultado provavelmente aconteceu devido à baixa representatividade de ambidestros ou outros tipos de falhas nos dados do transfermarkt[5]. Por isso, foi decidido ignorar essa feature.

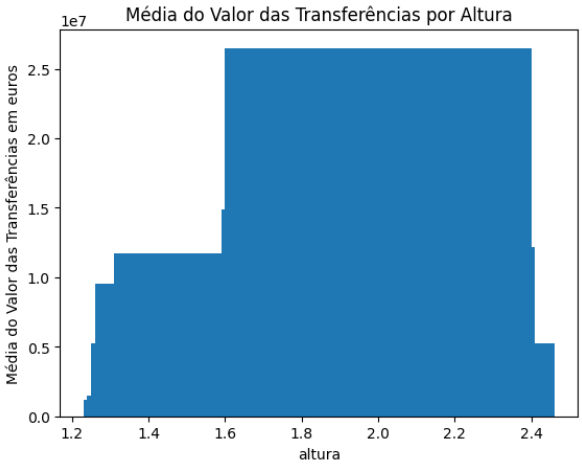

Figura 6: Média por altura

A análise em relação à altura do jogador não foi muito conclusiva, uma vez que entre 1,6m e 2,2m, que é a altura de grande maioria dos jogadores a média de valor de transferência se manteve constante. Entretanto, quando a altura sai do padrão, principalmente quando o jogador tem menos de 1,60 m, seu valor tende a cair, como é possível observar na figura 6.+

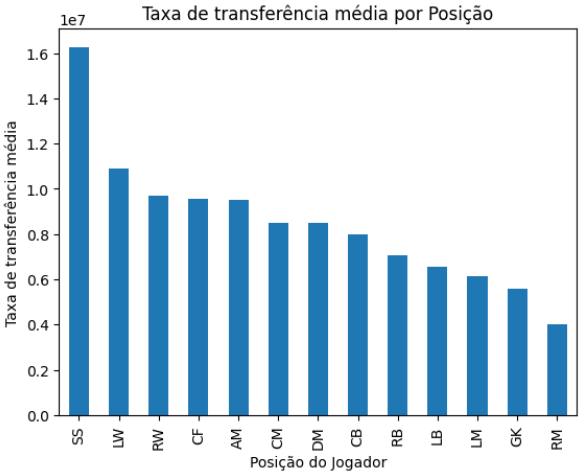

Figura 7: Média por posição

Na figura 7, é possível observar que jogadores das posições mais avançadas costumam ter um valor de transferência maior. Além disso, é possível observar que a posição "SS"(Second Striker) está muito discrepante, devido à baixa robustez da média e poucas transferências envolvendo essa posição.

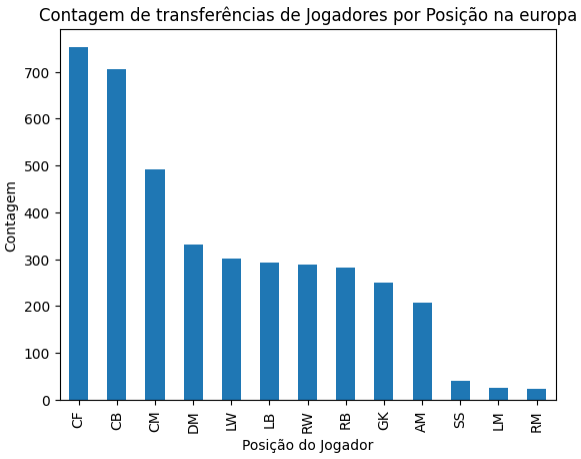

Figura 8: Distribuição das posições

Na figura 8, é possível notar a baixa representatividade das posições "RM", "LM"e "SS". Por isso, jogadores com a posição "RM", "LM"e "SS", tiveram suas posições alteradas, respectivamente, para "RW", "LW"e "AM".

3.3 Integrando dados do FBREF e Transfermarkt

Após a obtenção dos dados de transferências do site transfermarkt[5], a tabela de transferências foi mesclada com as tabelas de estatísticas do FBREF[7]. Para realizar a integração, foram utilizados os nomes dos jogadores e a temporada da transferência. Dessa forma, para cada transferência, foram obtidas as estatísticas do jogador na temporada caso

haja uma correspondência de jogador e temporada nas tabelas de dados do FBREF[7].

3.4 Pré-processamento dos dados

Para a construção do modelo e para os plots de correlação alguns dados tiverem que ser tratados. Alguns exemplos são a idade e a altura, que não se comportam de maneira linear, mas de uma forma similar à uma distribuição Gaussiana. No caso da feature altura, foi realizado um processo de binarização com 5 intervalos. No caso da idade, também foi testada uma transformação polinomial, que pareceu ser melhor em um primeiro momento, já que ajustava melhor os dados. Porém, ao generalizar, a binarização com 5 intervalos acabou funcionando melhor.

Alguns outros pré-processamentos foram feitos:

- Remapeamento das posições "SS", "RM"e "LM"para "AM", "RW"e "LW";
- Binarização das posições dos jogadores com dummies;
- Binarização da preferência de pé dos jogadores;
- Transformação polinomial de grau 2 da altura;

Por fim, as colunas foram normalizadas usando a transformação z-score, implementada na classe StandardScaler, na biblioteca sklearn.preprocessing. A normalização é útil ao buscar um modelo de regressão, pois coloca os valores das features em uma escala comum, mitigando o impacto das diferenças de escala ou valores extremos, permitindo uma análise mais precisa da relação entre as variáveis envolvidas na regressão.

3.5 Seleção de features baseada em correlação

A seleção de features é uma etapa importante no processo de análise de dados, principalmente em tarefas de modelagem preditiva e aprendizado de máquina.

3.5.1 Análise de correlação

Durante a análise de correlação, foram examinadas as correlações entre todas as features e também entre as features e a variável objetivo. Isso permitiu identificar as características que apresentaram uma correlação significativa com a variável objetivo, bem como aquelas que estavam altamente correlacionadas entre si.

Com o objetivo de evitar o problema de multicolinearidade, decidiu-se remover as features com uma correlação maior que 0.9 entre si.

Ao remover as features altamente correlacionadas, optou-se por manter apenas aquela que apresentava a maior correlação com a variável objetivo (Taxa de transferência). Essa abordagem garante que a característica mais relevante em relação à variável objetivo seja mantida, enquanto reduz a redundância de informações entre as features.

Com menos features, a interpretação dos resultados se torna mais simples e compreensível, facilitando a identificação dos principais fatores que afetam a variável objetivo.

Seguem abaixo as melhores correlações entre features selecionadas e a taxa de transferência:

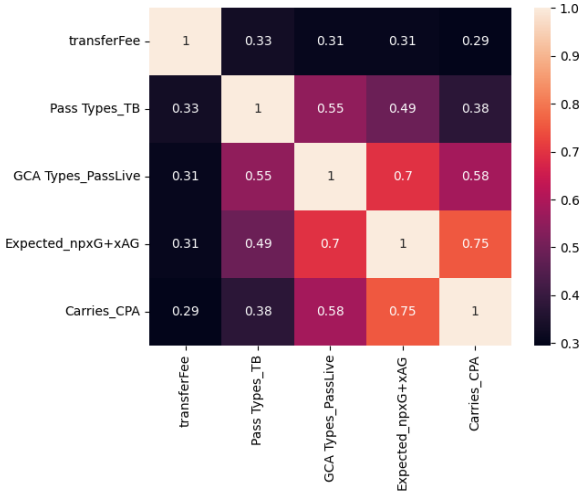

Figura 9: Jogador de qualquer posição / todas estatísticas

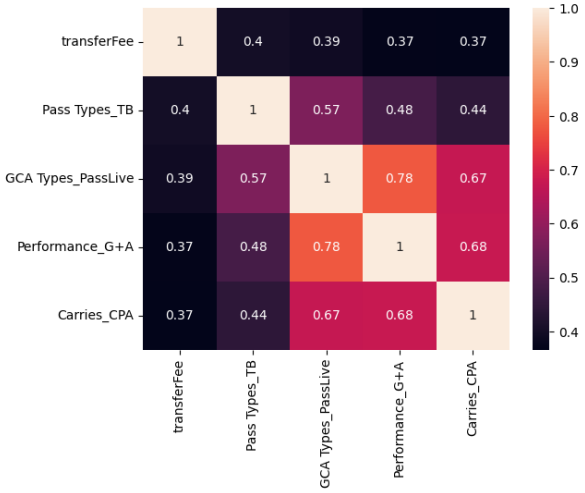

Figura 10: Atacantes / todas estatísticas

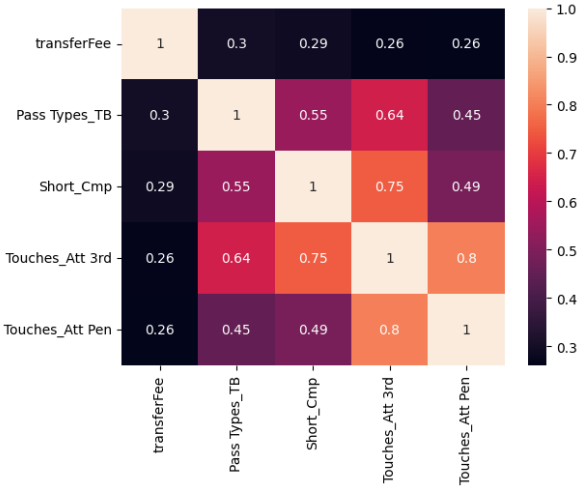

Figura 11: Meias / todas estatísticas

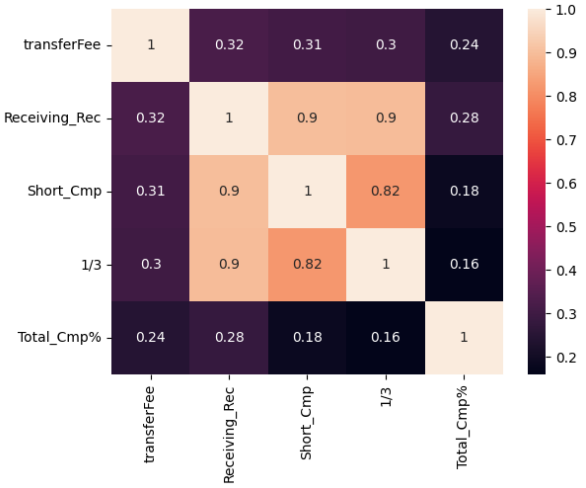

Figura 12: Defensores / todas estatísticas

3.6 Construção e avaliação do modelo

Nesse contexto, foram utilizados dois tipos de modelos para cada grupo de posição analisado (defensores, meio-campistas e atacantes): regressão linear e random forest.

A regressão linear é um algoritmo que busca estabelecer uma relação linear entre as variáveis de interesse. Nesse caso, foram construídos modelos de regressão linear para cada grupo de posição.

A random forest é um algoritmo de aprendizado de máquina que combina várias árvores de decisão para formar um modelo mais preciso e robusto. Também foram construídos modelos de random forest para cada grupo de posição.

Cada modelo foi avaliado em 100 iterações repetidas. Em cada iteração, os dados foram divididos em um grupo de treino, com 80% dos dados, e um grupo de testes, com os outros 20% dos dados. Os modelos de regressão linear e random forest foram gerados utilizando as ferramentas da biblioteca scikit-learn.

Após as iterações, foi calculado o valor médio do R^2 para cada grupo de posição e para cada tipo de modelo.

Os resultados obtidos foram os seguintes:

Regressão Linear:

Para os atacantes, o valor médio do R^2 foi de 0.1112.

Para os meio-campistas, o valor médio do R^2 foi de 0.1638.

Para os defensores, o valor médio do R^2 foi de 0.1590.

Random Forest:

Para os atacantes, o valor médio do R^2 foi de 0.3010.

Para os meio-campistas, o valor médio do R^2 foi de 0.9476.

Para os defensores, o valor médio do R^2 foi de 0.5674.

Esses resultados indicam a qualidade do ajuste dos modelos aos dados. Quanto mais próximo de 1 for o valor do R^2, melhor é o ajuste, pois indica que o modelo é capaz de explicar uma maior parte da variabilidade observada nos dados.

4 Resultados

Como resultados analíticos, conseguimos resultados a respeito das hipóteses que levantamos:

- **O valor de mercado de um jogador está diretamente relacionado com sua performance em campo.**

  Essa hipótese foi testada e confirmada. De fato, a performance do jogador é extremamente essencial para o valor do jogador, mesmo que os fatores mais importantes variam de posição por posição, eles estão relacionados às estatísticas dos jogadores em campo.

- **Jogadores mais novos têm um maior potencial de crescimento e, portanto, valores de mercado mais altos.**

  Percebemos que os valores de venda de jogador, por idade, segue uma distribuição normal. Dessa forma, os valores de venda dos jogadores aumentam entre 15 e 24-26 anos, onde chega no auge do seu valor, e depois começam a cair à medida que a idade aumenta. Isso faz sentido, já que quando o jogador está novo, seu valor está associado com o que ele pode vir a se tornar, quando chega na idade de 24-26 o jogador já desenvolveu bastante, amadureceu a ainda pode melhorar e amadurecer mais, entregando resultado a curto e longo prazo. Por fim, quando a idade avança é esperado que os atributos físicos comecem a se deteriorar e

o resultado a longo prazo seja bastante reduzido.

- **Jogadores que atuam em posições mais ofensivas, como atacantes, tendem a ter valores de mercado mais altos.**

  Conseguimos perceber, através dos plots, que essa hipótese também foi confirmada. É possível perceber que jogadores que atuam em setores mais ofensivos do campo tendem a ter maiores valores de mercado. Isso pode ocorrer porque normalmente são jogadores mais habilidosos, que apresentam mais dribles, mais criação de chances e, principalmente, gols, que é o produto principal do futebol.

- **Jogadores mais velhos tendem a diminuir a produtividade e estar mais suscetíveis a lesões, tendo, portanto, valores de mercado mais baixos.**

  Essa hipótese foi confirmada, como podemos ver pelos plots. Assim como dito anteriormente, o valor de mercado diminui pois à medida que a idade avança, é esperado que as características físicas vão se deteriorando, fazendo com que o jogador fique mais limitado, diminuindo sua produtividade e o deixando mais suscetível a lesões (ainda que essa última parte concluímos pela lógica do corpo humano, uma vez que não foi feito um estudo do número de lesões em relação à idade).

Para além das hipóteses, obtivemos resultados relacionados aos atributos mais relevantes para cada posição e o quanto eles se relacionam com seu valor de mercado. Esse resultado pode ser aplicado no contexto de transferência, para saber se um jogador de uma determinada posição estará supervalorizado ou não. Além disso, pode ser utilizado para encontrar jogadores que tenham boas estatísticas, mas não as melhores naqueles quesitos mais importantes para a função dele, como por exemplo um atacante que é muito bom recompondo e realiza muitos desarmes, talvez ele tenha um valor de mercado baixo, em comparação à outros jogadores, mas pode ser interessante em alguns esquemas táticos.

Por fim, de resultado final, temos nosso modelo que dado os atributos de um jogador, e sua posição ele gera seu valor de mercado. Essa última parte é a mais útil no contexto dos clubes, pois com elas é possível observar se o valor do jogador que o clube vendedor está pedindo vale a pena em relação ao valor gerado pelo modelo, e de forma semelhante para o clube que for vender, ver se o valor que está sendo oferecido pelo jogador é condizente com seu valor de mercado.

5 Discussão

Dado os resultados apresentados, podemos responder algumas questões que foram apresentadas nas hipóteses do projeto. De começo podemos concluir que as estatísticas puramente não são suficientes para prever com assertividade o valor de mercado de um jogador, embora tenha correlações que alterem seu valor. Além disso foi possível perceber que a idade dos jogadores influencia sim em seus valores, principalmente em jogadores mais velhos que tendem a perder muito valor de mercado.

Esse trabalho ainda tem espaço para melhorias, algumas inclusive que foram tentadas durante o nosso processo, mas não encontramos jeito eficaz de implementá-las. Uma dessas possíveis melhorias seria adicionar no modelo o fator popularidade do jogador, no nosso trabalho tentamos implementar através da biblioteca pytrends, entretanto tivemos dificuldade com o API do Google trends, que limita as nossas requisições, entretanto conseguimos para um certo número de jogadores, e nos deparamos com outro problema: jogadores com nomes comuns tem muitas pesquisas om seu nome, e ficaram na frente de nomes extremamente populares, como Cristiano Ronaldo. Outra possível melhoria seria colocar dados de evento, o que foi tentado no nosso trabalho, porem tendo em mãos apenas dados de uma temporada, o merge dos jogadores que tiveram transferências nessa temporada, o dataframe resultante é muito pequeno para trabalhar um modelo, com 381 jogadores somente. Algumas outras melhorias que poderiam ser adicionadas são: incluir os goleiros, realizar análises de popularidades, buscar maneiras de mesclar os dados estatísticos dos jogadores com os dados de transferências sem uma grande perda de informação, levar em consideração no modelo a pontuação/posição do time que o jogador saiu, além também de sua liga, que são fatores que acreditamos influenciar no valor dos jogadores.

6 Conclusão

Com esse trabalho foi possível perceber quais os fatores que mais importam para o valor de venda do jogador e como isso varia de posição por posição, apresentando fatores diferentes para um zagueiro e para um atacante, por exemplo. O que mais faz diferença para as seguintes posições são:

1. Atacante : Número de passes que quebram linha, Expected Goal, Toques na área do Pênalti , Número de vezes que carregou a bola para área do pênalti

2. Meio - Campistas : Número de passes que quebram linha, passes curtos completos e toques no último terço

3. Defensores : Número de vezes que o jogador dominou um passe, passes curtos completos e porcentagem de passes completos

Além disso, como foram construídos modelos utilizando regressão linear e Random Forest, foi possível observar que, considerando o r2 score, o modelo de Random Forest obteve um desempenho superior em relação ao modelo de regressão linear utilizado na análise dos valores das posições. O fato de o modelo Random Forest obter um desempenho melhor diz muito sobre a natureza da relação entre as features e a taxa de transferência, pois a regressão linear assume uma relação linear simples entre as variáveis e tenta ajustar uma reta aos dados, e pode não conseguir capturar bem as relações não lineares, gerando um R2 baixo. Por outro lado, o modelo de Random Forest é capaz de capturar relações não lineares e complexas nos dados, já que é construído a partir de uma combinação de árvores de decisão, onde cada árvore é treinada em um subconjunto dos dados e características selecionadas aleatoriamente. A média das previsões das árvores individuais é usada para fazer a previsão final. Essa abordagem permite que o modelo de Random Forest capture relações não lineares e interações entre as características. Portanto, o modelo de Random Forest foi capaz de se ajustar melhor aos dados, capturando relações mais complexas e, assim, alcançando um R2 mais alto do que o modelo de regressão linear.

Por fim, podemos concluir que com esse trabalho é possível perceber quais atributos fazem um jogador valer mais ou menos. Através da aplicação do modelo é possível identificar jogadores com valores de mercado superestimados ou subestimados. Porém, é importante destacar que, apesar das análises de atributos e construção de modelos, ainda há muitas variáveis envolvendo o valor de um jogador, e ainda não é possível avaliar com precisão o valor real de um jogador apenas com as features utilizadas nesse trabalho. Dessa forma, o trabalho ainda está aberto a diversas melhorias devido à natureza complexa da avaliação de valores de mercado de jogadores.

7 Referências

1. Modelling football players values on the transfer market and their determinants using robust regression models - Rafał Stepień (Bachelor's thesis under scientific supervision of dr hab. Michał Rubaszek, prof. SGH written in Institute of Econometrics)
2. https://www.sciencedirect.com/science/article/abs/pii/S144135231300096X
3. https://ieeexplore.ieee.org/abstract/document/9721908
4. https://onlinelibrary.wiley.com/doi/full/10.1111/joes.12552
5. https://www.transfermarkt.com.br
6. https://github.com/probberechts/soccerdata
7. https://www.fbref.com