**KENNETH DE COSTER**

**Identificação e previsão da taxa de transação de jogadores no mercado internacional de futebol**

Vitória

05/2021

**Resumo**

O mercado de transferências de futebol tem crescido exponencialmente nos últimos anos com o aumento das taxas de transferência para jogadores de futebol. Hoje em dia, o preço que um clube paga por um jogador pode variar de milhões a centenas de milhões de dólares. Este estudo apresenta o uso de regressão linear para identificar a relação entre o preço de transferência de jogadores de futebol e os atributos do jogador como valor, idade, habilidade, potencial, duração do contrato e posição. O objetivo deste projeto é criar um modelo que faça previsões precisas sobre as taxas de transferência e entregar uma ferramenta baseada na entrada de dados pelo usuário. A pesquisa científica é feita por meio da aplicação de conceitos estatísticos como correlação, coeficiente de determinação e regressão linear e usando tecnologias como Python e o Notebook Jupyter para criar visualizações e realizar cálculos. Ao final deste trabalho, o modelo é utilizado e validado em um dataset onde atingiu uma precisão de 108% (no qual 100% é uma previsão perfeita). Por fim, este estudo aponta algumas melhorias que podem ser feitas para aumentar ainda mais a precisão do modelo e os seus resultados.

**Palavras-chave:** Regressão Linear. Taxas de Transferência. Previsões.

**1. INTRODUÇÃO**

Segundo a plataforma de dados Statista (2019), o valor do mercado europeu de futebol é estimado em 28.9 bilhões de dólares, tornando-o o esporte mais popular do planeta. Alguns dos maiores clubes da Europa valem de 1 bilhão a quase 5 bilhões de dólares. A Forbes (2021) colocou o FC Barcelona no topo da lista como o time mais valorizado do mundo, com um valor total de $4.76 bilhões, seguido de perto pelo rival Real Madrid ($4.75 bilhões).

É seguro dizer que a indústria do futebol está crescendo. A maior parte desse dinheiro vem de patrocínios, contratos de televisão, prêmios e principalmente de **transferências de jogadores**. Os preços dos jogadores aumentaram rapidamente na última década. Para dar um exemplo, o Cristiano Ronaldo foi vendido em 2009 por *94 milhões de euros*. Oito anos depois, em 2017, o Neymar mudou-se do FC Barcelona para o Paris Saint-Germain por *222 milhões de euros*, comprovando o aumento do valor das transferências de jogadores entre os clubes.

Segundo a Statista (2019), os clubes gastaram cerca de $7.35 bilhões em jogadores. Comparando com anos anteriores, existe uma *tendência de crescimento* no mercado a qual continuará crescendo na próxima década.

Por meio desses dados disponíveis, qual será o motivo pelo qual esses preços estão crescendo? Seria possível identificar e estabelecer uma relação entre certas variáveis? Ajudaria a prever preços futuros? Hoje não existe um modelo, ou uma linha de base que defina o preço de um jogador. *Não há uma maneira real* de saber se um clube está pedindo o preço correto ou se está superestimando os preços dos seus jogadores.

O **objetivo geral** desse estudo é identificar uma possível relação entre o preço que os clubes pagam por um jogador e outros atributos, tais como: *idade, o valor atual, habilidades, potencial, posição em campo e duração do contrato*. Caso haja uma relação, prever os preços futuros usando os mesmos atributos. Para alcançar o objetivo, será realizada uma **pesquisa científica**.

Os seguintes **objetivos específicos** são necessários:

* Usar a Correlação Linear, para entender a *força da associação* entre o preço do jogador (variável dependente) e cada atributo individualmente ou atributos combinados (variáveis ​​independentes).
* Calcular o coeficiente de determinação para expressar a *porcentagem de variação* das variáveis.
* Usar a regressão linear múltipla, para *modelar uma associação* entre o preço do jogador e cada atributo ou atributos combinados (caso haja uma correlação forte).
* *Prever os preços* dos jogadores no mercado atual por meio dos seus atributos e modelos estabelecidos.

Com a indústria do futebol crescendo rapidamente, é interessante entender a ideia por trás de uma das principais fontes de receita (compra e venda de jogadores) e possivelmente modelá-la. Não existe uma linha de base real para a compra de jogadores no mercado de futebol e fica difícil para um clube saber se está pagando o preço certo (ou não). No caso de criar um modelo confiável, ele será transformado em um aplicativo desktop para ser usado por qualquer clube para determinar um preço preciso para os jogadores. Ao inserir as variáveis, o aplicativo retornará uma estimativa de preço.

Esperançosamente, o modelo forneceria conhecimento necessário para *a decisão dos clubes* e assim auxiliando na gestão financeira.

**2. REFERENCIAL TEÓRICO**

Para a resolução do problema apresentado na seção anterior, serão aplicados os seguintes conceitos: *Regressão Linear, Coeficiente de Correlação e Coeficiente de Determinação*.

Primeiramente, é importante entender o que significa **Regressão Linear** e o seu propósito. Segundo Sandra Rodrigues (2012, p. 17), a Regressão Linear define técnicas estatísticas usadas para modelar as relações entre variáveis e realizar previsões de uma ou mais variáveis. A equação representativa do modelo de regressão linear simples:

onde:

representa o valor da variável dependente Y.

representa o valor da variável independente X.

representa o intercepto entre a linha de regressão e eixo Y.

representa o coeficiente linear de regressão da variável dependente Y em função da variável independente X.

O segundo conceito do estudo é o **coeficiente de correlação**:

Os coeficientes de correlação são métodos estatísticos para se medir as relações entre variáveis e o que elas representam.

O que a correlação procura entender é como uma variável se comporta em um cenário onde outra está variando, visando identificar se existe alguma relação entre a variabilidade de ambas. Embora não implique em causalidade, o coeficiente de correlação exprime em números essa relação, ou seja, quantifica a relação entre as variáveis. (BRUNO OLIVEIRA, 2019, p.1)

É importante conhecer as diferentes técnicas para a aplicação do coeficiente de correlação. A forma que os dados são distribuídos vão determinar a técnica. Algumas técnicas são coeficiente de correlação de Pearson, Coeficiente de Spearman e Coeficiente de Contingência. (Paulo Guimarães, 2017, p.1)

O coeficiente de Correlação Linear de Pearson entre duas variáveis quantitativas X e Y é dado por:

“A intensidade do coeficiente de correlação (r) pode variar entre -1 e 1, sendo que quanto mais próximo de -1 ou 1, mais forte será a associação entre as duas variáveis e quanto mais próximo de 0, mais fraca será a associação. ” (OLIVEIRA, 2018, p.160)

Tabela 1 - Escalas de Correlação

|  |  |
| --- | --- |
|  | Relacionamento entre X e Y |
| 1 ou -1 | Perfeito |
| 0.7 ou -0.7 | Forte |
| 0.5 ou -0.5 | Moderado |
| 0.25 ou -0.25 | Fraco |
| 0 | Inexistente |

Fonte: o autor.

Por último, o **coeficiente de Determinação**:

O coeficiente de determinação represente a proporção da variação da variável dependente que é explicada pela variação da variável independente. O R2 exercerá esse papel de modo significativo no caso de relações linear estimadas pelo método dos mínimos quadrados ordinários. (KENNEDY, 2008, p.84)

“O coeficiente de determinação é dado por e é dado pelo valor encontrado para a Correlação Linear do Pearson ao quadrado” (OLIVEIRA, 2018, p.164):

“Quanto mais próximo de 1 estiver o coeficiente de determinação, melhor será o grau de explicação da variação de Y em termos da variável X.” (LIMA FILHO, 2021)

Existem outros trabalhos que também possuem como finalidade a criação de um modelo para a tomada de decisões financeiras, utilizando Regressão Linear:

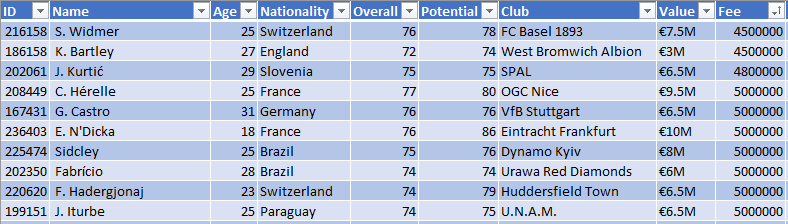
1. SELL, I. **UTILIZAÇÃO DA REGRESSÃO LINEAR COMO FERRAMENTA DE DECISÃO NA GESTÃO DE CUSTOS**. IX Congresso Internacional de Custos. Florianópolis, SC, Brasil. p.14. 2005.
2. BASTOS, E.V.P. **MODELO DE REGRESSÃO LINEAR PARA ANÁLISE DE INVESTIMENTOS EM UMA EMPRESA DO RAMO**.Centro de Instrução Almirante Graça Aranha (CIAGA), Rio de Janeiro, RJ, Brasil. p.12. 2015.

Todos os conceitos apresentados contribuem para a criação do modelo de regressão linear. O próprio modelo dará uma ideia de quanto um jogador pode ser vendido no futuro. O modelo de Regressão Linear é essencial para a resolução do problema apresentado neste trabalho.

**3. METODOLOGIA**

O primeiro passo para a criação do nosso modelo, é ter um dataset que contenha todos os dados ​​necessários. O dataset é essencial para a criação do modelo e para validar se o modelo está alcançando previsões precisas. Para o estudo, será utilizado o dataset do jogo **FIFA19** disponível no site de conjunto de dados *kaggle.com*, o qual é composto por 17866 jogadores de futebol. Esse site contém também mais de 60 atributos (ou colunas) e nem todos eles serão utilizados no modelo, apenas os relevantes como *nome*, *idade*, *habilidades*, *potencial*, *valor atual*, *preço*, *posição*, *duração do contrato,* entre outros. O dataset é salvo no formato Excel e carregado no Python usando a biblioteca Pandas.

Figura 1 – Dataset em Excel



Fonte: o autor

Dando uma primeira olhada no dataset, existem vários problemas e inconsistências a serem resolvidos antes dos dados poderem realmente ser usados. Os principais problemas com o conjunto de dados são:

* Os tipos de dados das colunas '*Value*' e '*Fee*' são inconsistentes. '*Value*' é do tipo string usando o formato '€ xxM / K' (onde 'M' é milhão e 'K' mil). '*Fee*', por outro lado, é do tipo inteiro.
* Nem todos os jogadores possuem um valor na coluna '*Fee*', isso significa que certos jogadores não foram transferidos durante a janela de transferências na temporada 2018-2019 e, por isso, não são relevantes para o nosso modelo.
* Alguns dados da coluna ‘*Contract Valid Until*’ contêm valores ‘N/A’ (ou 'NaN'), que causarão problemas durante os cálculos.
* Alguns dos formatos da coluna *'Contract Valid Until*' são inconsistentes com o resto. A maioria dos valores contém apenas o ano (por exemplo, 2022), embora outros campos usem o formato 'Mês, Dia, Ano'.
* Para usar os atributos ‘​​*Contract Valid Until*’ e '*Position*', eles precisam ser transformados de valores categóricos em valores numéricos.

Para resolver esses problemas do dataset, será necessário preparar, limpar e tratar os dados. Este processo será realizado usando a linguagem de programação **Python** em combinação com o IDE **JupyterLab**.

Após o processamento e padronização dos dados, a correlação entre a variável dependente preço (atributo ‘*fee*’) e as variáveis ​​independentes *(idade*, *valor atual*, *habilidades*, *potencial*, *posição* e *duração do contrato*) é calculada individualmente:

* preço~idade
* preço~valor
* preço~habilidades
* preço~potencial
* preço~duração contrato
* preço~posição

Cada um deles mostrará o quão forte é a relação entre a variável dependente e cada variável independente. Apenas as correlações (r) com valor **entre (-) 0,6 e (-) 1** serão incluídas no modelo, de forma a obter os melhores resultados possíveis. Para calcular e visualizar o **coeficiente de correlação de Pearson**, são utilizadas as bibliotecas Sklearn e Matplotlib em Python. Sklearn fornecerá o valor de correlação (entre -1 e 1), enquanto Matplotlib cria o gráfico de dispersão para uma representação visual de ambas as variáveis ​​e a linha de regressão linear. Uma vez estabelecido o coeficiente de correlação de Pearson entre as duas variáveis, o **coeficiente de determinação** é obtido em Python utilizando a biblioteca Sklearn.

Com o resultado da análise de correlação de cada variável independente com a variável dependente (preço), será realizada uma seleção para escolha dos atributos que contribuem para um modelo confiável. Após, selecionar os atributos um modelo de **regressão linear múltipla** é definido usando as bibliotecas Sklearn e Numpy. Para a validação do modelo, o script fará uma estimativa do preço de cada jogador e como o dataset já contém o preço real dos jogadores, uma comparação será realizada para verificar o quão próximo o modelo chega do preço real. Para organizar a validação, um novo dataframe será criado em Python, que irá receber todos os jogadores do conjunto de dados original com os atributos previamente selecionados. O resultado será armazenado no novo dataframe, próximo à coluna do preço real. A fórmula a seguir nos fornecerá a diferença percentual entre o preço estimado e o preço real, no qual 100% seria uma **predição perfeita**:

Diferença (%) = (Preço Estimado / Preço Real) x 100

Por exemplo, se o preço estimado for *€24,509,149* e o preço real *€23,000,000* então:

107 (%) = (24,509,149/23,000,000) x 100

Neste exemplo, o resultado se aproximou bem do preço real.

Comparando o resultado com o preço real de cada jogador, podemos ter uma ideia de quão preciso é o modelo e se ele pode ser usado para prever preços. Para nosso estudo, uma previsão média aceitável do modelo está dentro de um intervalo de **15%** para mais ou para menos (ou seja, entre 85% e 115%). Caso o modelo consiga prever dentro dessa faixa, será desenvolvido um aplicativo de linha de comando em Python. O aplicativo conterá o modelo e permitirá que o usuário insira as variáveis ​​independentes. Em seguida, ele retornará uma **estimativa de preço** ao usuário, levando em consideração a entrada de dados.

**4. DESENVOLVIMENTO**

A primeira atividade da criação do modelo é realizar **a limpeza e o tratamento dos dados**. Com dados inexistentes (N/A, ou NaN), ou dados com falta de padronização, não será possível treinar o modelo de Regressão Linear. Antes de realizar as operações de limpeza e tratamento, o arquivo .xlsx com os dados é importado no ambiente Jupyter Notebook. A importação é feita através da biblioteca **Pandas**, usando o comando pd.read\_excel().

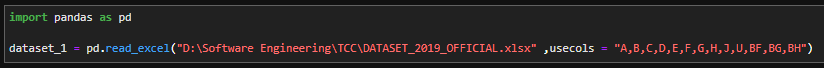


Figura 2 - Importação do dataset

Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

O primeiro campo a ser tratado é o campo ‘**Preço**’ (*Fee*). Este campo contém os valores dos jogadores que foram vendidos na temporada 2018-2019 e representa o campo que queremos prever. No dataset importado, existem vários jogadores que possuem um valor NaN no campo, isso pode ser explicado pelo fato que o jogador não foi vendido nesta temporada (figura 3). Será necessário selecionar apenas os jogadores com um valor no campo ‘Preço’ por meio da função **.notnull()**, pois apenas esses jogadores agregam um valor para o nosso modelo. Os valores da planilha são do tipo String e precisam ser convertidos para o tipo Inteiro, conforme a figura 4.



Figura 3 - O campo 'Fee' com valores NaN

Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

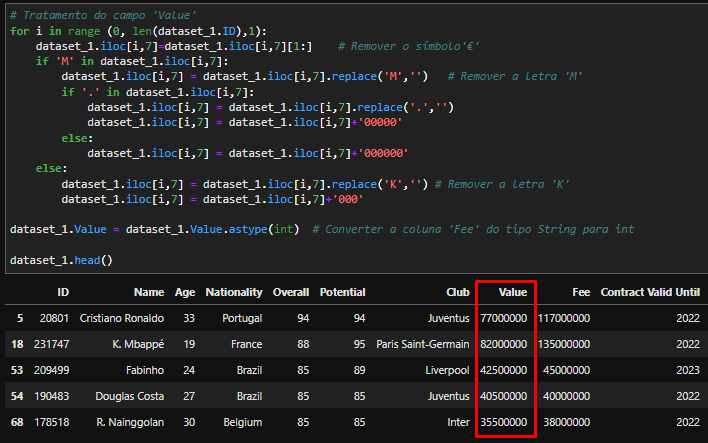


Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

Figura 4 - Tratamento do campo 'Fee' e resultado

O campo ‘**Valor**’ (*Value*) representa o valor atual do jogador. Existem duas operações a serem feitas no campo: o formato '€ xxM / K' deve estar no formato do campo ‘Preço’ (inteiro) e deve ser convertido do tipo String para o tipo Inteiro. Para o tratamento do formato, será necessário tratar cada valor separado, usando o for-loop para a iteração e as estruturas condicionais ‘if’ para realizar as operações especificas, conforme a figura 5. Para converter o tipo do valor, basta utilizar a função da Pandas **.astype()**.

Figura 5 - Tratamento do campo 'Valor' e resultado



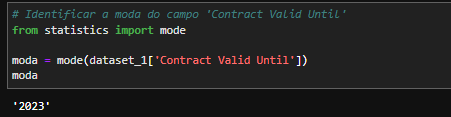
Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

O próximo campo do dataset que exige uma transformação, é o campo ‘**Duração Contrato**’ (*Contract Valid Until*). O campo representa a duração do contrato, ou seja, em qual ano o contrato venceria. Três operações devem ser realizadas neste campo:

1. Tratar os valores inexistentes (NaN), utilizando a moda.
2. Extrair o ano (alguns campos estão com o formato ‘Jun 30, 2019’, outros possuem apenas o ano).
3. Aplicar encoding para transformar os valores categóricos em valores numéricos.

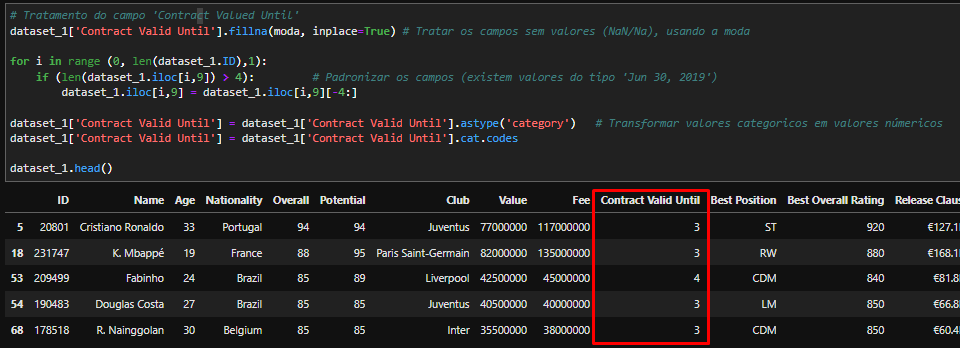
A moda do campo ‘Duração Contrato’ pode ser identificada utilizando a função **.mode()** da biblioteca ‘statistics’ (figura 6). Para a padronização dos valores, os quatro dígitos de cada String são copiados e sempre vão representar o ano. Por fim, os valores categóricos são transformados em valores numéricos com a função **.cat.codes**, conforme a figura 7.

Figura 6 - Identificar a moda



Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

Figura 7 - Tratamento do campo 'Contract Valid Until' e resultado



Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

O último campo do dataset a ser tratado é ‘**Posição**’ (*Best Position*) que representa a posição do jogador no campo de futebol (atacante, zagueiro, goleiro etc.). Como se tratam de valores categóricos, será necessário transforma-lhes em valores numéricos. A transformação é feita através dos métodos **.astype()** e **.cat.codes**, conforme a figura 8.

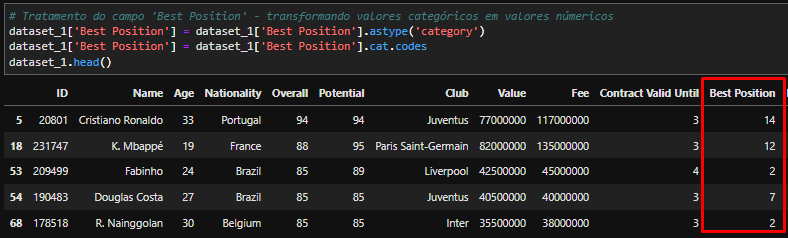


Figura 8 - Tratamento do campo 'Best Position' e resultado

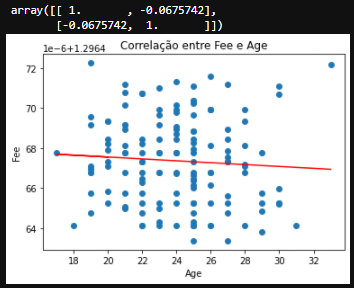
Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

Os dois outros campos, ‘**Habilidades**’ (*Overall*) e ‘**Potencial**’ (*Potential*), já estão no formato numérico correto e não possuem valores inexistentes e consequentemente não precisam ser tratados. Com isso, se encerra a primeira atividade do cronograma, a limpeza e tratamento dos dados.

A segunda atividade para a criação do modelo é **calcular a correlação linear** entre a variável dependente (‘Preço’) e cada variável independente (‘Idade’, ‘Valor’, ‘Habilidades’, ‘Potencial’, ‘Duração Contrato’, ‘Posição’). Por meio da correlação, podemos selecionar as variáveis que possuem uma forte relação com a variável dependente. Apenas as correlações (r) com valor **entre (-) 0,6 e (-) 1** serão incluídas no modelo. Esta etapa á realizada usando as bibliotecas Sklearn (para a criar o modelo linear), Numpy (para calcular o valor da correlação) e Matplotlib (para a representação visual). Um gráfico de dispersão é apresentado para mostrar a correlação de forma visual. Os resultados do algoritmo serão apresentados, o código-fonte se encontra no link <https://github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees>.

* preço~idade

A correlação entre a variável ‘Preço’ e a variável ‘Idade’ possui um valor de **-0.0675**, ou seja, a força de associação entre as duas variáveis é **inexistente**. O gráfico de dispersão apresenta a linha de previsão (em vermelho) que não se aproxima bem dos dados do dataset. Com essas informações podemos afirmar que a variável ‘Idade’ não é uma variável produtiva para realizar previsões de preços de jogadores e **não será incluída** no modelo de Regressão Linear.

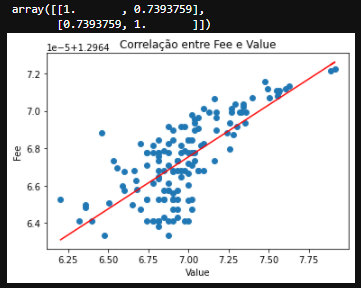
Figura 9 - Correlação entre 'Preço' e 'Idade'

Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

* preço~valor

A correlação entre a variável ‘Preço’ e a variável ‘Valor’ possui um valor de **0.7393**, ou seja, existe uma correlação **positiva e forte** entre as duas variáveis. O gráfico de dispersão apresenta a linha de previsão (em vermelho) que se aproxima bem dos dados do dataset. Com essas informações podemos afirmar que a variável ‘Valor’ é uma variável produtiva para realizar previsões de preços de jogadores e **será incluída** no modelo de Regressão Linear.

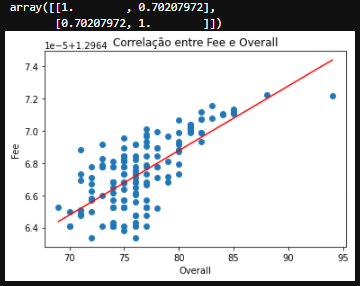
Figura 10 - Correlação entre 'Preço' e 'Valor'



Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

* preço~habilidades

A correlação entre a variável ‘Preço’ e a variável ‘Habilidades’ possui um valor de **0.7020**, ou seja, existe uma correlação **positiva e forte** entre as duas variáveis. O gráfico de dispersão apresenta a linha de previsão (em vermelho) que se aproxima bem dos dados do dataset. Com essas informações podemos afirmar que a variável ‘Habilidades’ é uma variável produtiva para realizar previsões de preços de jogadores e **será incluída** no modelo de Regressão Linear.

Figura 11 - Correlação entre 'Preço' e 'Habilidades'

Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

* preço~potencial

A correlação entre a variável ‘Preço’ e a variável ‘Potencial’ possui um valor de **0.6408**, ou seja, existe uma correlação **positiva e moderada** entre as duas variáveis. O gráfico de dispersão apresenta a linha de previsão (em vermelho) que se aproxima razoavelmente dos dados do dataset. Com essas informações podemos afirmar que a variável ‘Potencial’ é uma variável produtiva para realizar previsões de preços de jogadores e **será incluída** no modelo de Regressão Linear.

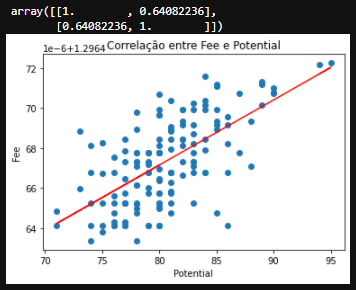


Figura 12 - Correlação entre 'Preço' e 'Potencial'

Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

* preço~duração contrato

A correlação entre a variável ‘Preço’ e a variável ‘Duração Contrato’ possui um valor de **0.3696**, ou seja, existe uma correlação **positiva e fraca** entre as duas variáveis. O gráfico de dispersão apresenta a linha de previsão (em vermelho) que não se aproxima bem dos dados do dataset. Com essas informações podemos afirmar que a variável ‘Duração Contrato’ não é uma variável produtiva para realizar previsões de preços de jogadores e **não será incluída** no modelo de Regressão Linear.

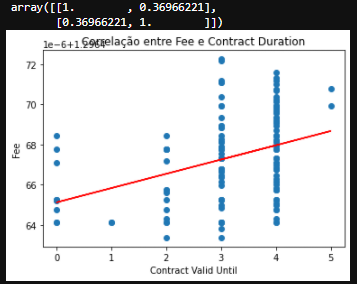


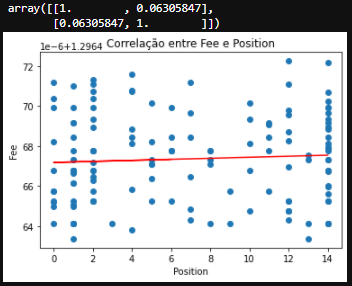
Figura 13 - Correlação entre 'Preço' e 'Duração Contrato'

Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

* preço~posição

A correlação entre a variável ‘Preço’ e a variável ‘Posição’ possui um valor de **0.0630**, ou seja, a força de associação entre as duas variáveis é **inexistente**. O gráfico de dispersão apresenta a linha de previsão (em vermelho) que não se aproxima bem dos dados do dataset. Com essas informações podemos afirmar que a variável ‘Posição’ não é uma variável produtiva para realizar previsões de preços de jogadores e **não será incluída** no modelo de Regressão Linear.

Figura 14 - Correlação entre 'Preço' e 'Posição'

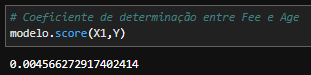


Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

A terceira atividade definida no cronograma é obter **o coeficiente de determinação** entre a variável dependente e as variáveis independentes. O coeficiente de determinação expressa variação dos valores da variável dependente em função das variáveis independentes. O resultado obtido mostra como a linha de previsão se aproxima dos dados do dataset e é representado por um valor **entre 0 e 1**. Quanto mais perto de 1, quanto maior a variação e quanto melhor a linha de previsão representa os dados. Para obter o coeficiente de determinação em Python, utilizamos o método **.score()** da biblioteca Scikit-learn.

* preço~idade

O coeficiente de determinação entre ‘Valor’ e ‘Idade’ é **0,004**, ou seja, apenas **0,4%** dos dados estão próximos da linha de regressão.

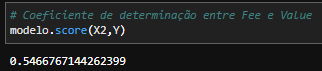
Figura 15 - Coeficiente de Determinação entre 'Preço' e 'Idade'

Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

* preço~valor

O coeficiente de determinação entre ‘Preço’ e ‘Valor’ é **0,5466**, ou seja, **54,66%** dos dados estão próximos da linha de regressão.

Figura 16 - Coeficiente de Determinação entre 'Preço' e 'Valor'

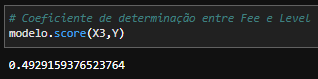


Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

* preço~habilidades

O coeficiente de determinação entre ‘Preço’ e ‘Habilidades’ é **0,4929**, ou seja, **49,29%** dos dados estão próximos da linha de regressão.

Figura 17 - Coeficiente de Determinação entre 'Preço' e 'Habilidades'

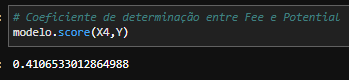


Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

* preço~potencial

O coeficiente de determinação entre ‘Preço’ e ‘Potencial’ é **0,4106**, ou seja, **41,06%** dos dados estão próximos da linha de regressão.

Figura 18 - Coeficiente de Determinação entre 'Preço' e 'Potencial'

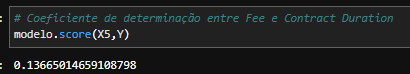


Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

* preço~duração contrato

O coeficiente de determinação entre ‘Preço’ e ‘Duração Contrato’ é **0,1366**, ou seja, apenas **13,66%** dos dados estão próximos da linha de regressão.

Figura 19 - Coeficiente de Determinação entre 'Preço' e 'Duração Contrato'

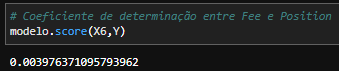


Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

* preço~posição

O coeficiente de determinação entre ‘Preço’ e ‘Posição’ é **0,0039**, ou seja, apenas **0,39%** dos dados estão próximos da linha de regressão.

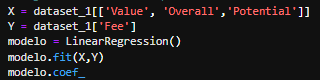
Figura 20 - Coeficiente de Determinação entre 'Preço' e 'Posição'



Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

Nas etapas de correlação linear e coeficiente de determinação foram definidas três variáveis que serão utilizadas na construção do **modelo final de regressão linear múltipla**. As três variáveis independentes com a maior contribuição para o modelo são: ‘*Valor*’, ‘*Habilidades*’, ‘*Potencial*’. Para a construção do modelo, é necessário primeiro guardar as variáveis independentes em uma variável X, e a variável dependente (‘Preço’) em uma variável Y. Em seguida, o modelo de regressão linear é construído e treinado usando ambas variáveis X e Y, conforme a figura 21.

Figura 21 - Criação do modelo de regressão linear múltipla



Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

A próxima etapa é **prever os preços de jogadores usando o modelo**. Com o modelo definido, podemos executá-lo e realizar previsões com base nas três variáveis ​​do dataset. Para capturar as previsões, é criado um novo dataset que recebe a saída do nosso modelo para fins de validação. Adicionamos a coluna '*prev\_1*' ao dataset que contém as previsões de preço para cada jogador. É possível fazer uma comparação entre o campo ‘*Preço*’ (Fee) e ‘*prev\_1*’. Por exemplo, o modelo previu o preço do jogador K.Mbappé como **€123,216,180** enquanto o preço real da venda do jogador foi **€135,000,000**.

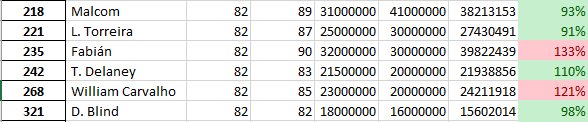
Figura 22 - Previsões de preço de cada jogador



Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

Com as previsões realizadas, podemos fazer a **validação** de quão preciso nosso modelo previu os preços. Para a validação, a diferença entre o preço estimado e o preço real é calculada e expressa em porcentagem. Um resultado de 100% seria uma **previsão perfeita** e as previsões dentro de **15%** (ou seja, entre 85% e 115%) seria considerado como um resultado aceitável para nosso modelo. Para facilitar a validação, o dataset é exportado para um arquivo Excel, onde podemos entender melhor o desempenho do modelo. A última coluna na figura 23 representa a diferença (em %) entre o preço estimado e o preço real. Uma validação mais extensiva será apresentada no **capitulo 5** ‘*Resultados Obtidos*’.

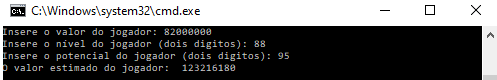
Figura 23 - Resultados do modelo em Excel



Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

Por último, **uma aplicação de linha de comandos** foi desenvolvida para que previsões possam ser realizadas a partir da entrada de dados pelo usuário. Para a execução da aplicação, basta abrir o Windows Command Prompt (cmd), e executar o comando ‘python TCC.py’. A aplicação solicita o valor atual do jogador, o nível do jogador e o potencial do jogador. O **valor estimado** é retornado para o usuário em seguida.

Figura 24 - Aplicação de linha de comandos



Fonte: github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees

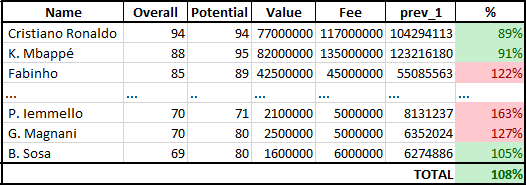
Todos os arquivos utilizados durante o desenvolvimento do modelo e da aplicação se encontram no seguinte repositório: <https://github.com/dacoster90/Predict-Soccer-Players-Transfer-Fees>.

**5. RESULTADOS OBTIDOS**

Para a validação dos resultados do modelo, examinaremos duas métricas: **a média das todas as previsões** e o **número de previsões individuais corretas**.

Conforme definido anteriormente, para o modelo ser considerado preciso, ele deve atingir uma precisão entre **85%** e **115%** (onde 100% é uma previsão perfeita). Cada jogador recebeu uma previsão baseada em seu valor, habilidade e potencial, que é expressa em porcentagem. A média de todas as previsões do modelo é de **108%**. Concluímos que o modelo alcançou **uma precisão aceitável**.

Figura 25 - Média das previsões (108%)



Fonte: o autor

Observando os **casos individuais**, no entanto, apenas 46 dos 146 jogadores tiveram previsões entre 85% e 115%. Isso representa **32%** dos jogadores do dataset. Isso significa que a maioria dos jogadores não teve previsões aceitáveis ​​e precisas obtidas pelo modelo. Um olhar mais atento a essa estatística nos mostra que a **maior concentração** de previsões precisas ainda está na faixa de 85% a 115%. Observando o intervalo de 70%-130%, o total sobe para 82 jogadores, o que representa um aumento de **25%**. O intervalo 55%-145% coloca o total para 112 jogadores, representando um aumento de **21%**. 40%-160% aumenta as previsões apenas em **12%** dos jogadores e 25%-175% apenas em **5%**. A figura 26 representa a concentração das previsões dos diferentes intervalos.

Figura 26 – Precisão dos Intervalos

*100%*

*10%*

*0%*

*25%*

*40%*

*55%*

*85%*

*70%*

**2%**

**4%**

**5%**

**12%**

**21%**

**25%**

**32%**

*200%*

*190%*

*175%*

*160%*

*145%*

*130%*

*115%*

Fonte: o autor

Tabela 2 – Previsões por intervalo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Intervalos | | Previsões por  Intervalo  (incremental) | Incremento | Incremento porcentual |
| *100%* | *100%* | 0 | + 0 | 0% |
| *85%* | *115%* | 46 | + 46 | 32% |
| *70%* | *130%* | 82 | + 36 | 25% |
| *55%* | *145%* | 112 | + 30 | 21% |
| *40%* | *160%* | 129 | + 17 | 12% |
| *25%* | *175%* | 137 | + 8 | 5% |
| *10%* | *190%* | 143 | + 6 | 4% |
| *0%* | *200%* | 146 | + 3 | 0% |
| **Total** | |  | 146 | 100% |

Fonte: o autor

Com esses resultados, ficou claro que, embora muitas das previsões não estivessem dentro da faixa ideal (85%-115%), o desempenho geral do modelo ainda é aceitável. Isso significa que **existe uma relação real** entre a taxa de transferência de um jogador e os atributos do jogador (valor, habilidades, potencial, etc.) e podemos usar essa relação para realizar previsões. De certa forma, os clubes de futebol podem fazer uso do modelo para definir os preços de seus jogadores, mas deve-se levar em consideração que o modelo pode apresentar um preço impreciso. Mais pesquisas são necessárias para melhorar o modelo e ter mais previsões dentro da faixa aceitável.

**6. CONSIDERAÇÕES FINAIS**

Com isso podemos concluir o projeto. Utilizamos a correlação para estudar a relação entre o preço de transferência de um jogador de futebol e atributos como valor, habilidade, idade, potencial, duração do contrato e posição. A regressão linear foi usada para fazer previsões com base nesses atributos e no relacionamento.

Python, Jupyter Notebook e Excel foram usadas ​​para realizar todo o processamento de dados, limpeza de dados, cálculos e para apresentar os resultados da pesquisa. Também foi desenvolvida uma ferramenta que retorna a taxa de transferência prevista de um jogador com base na entrada dos dados pelo usuário.

Além disso, o modelo alcançou uma boa precisão geral ao prever preços de transferências de jogadores. Em um nível individual, no entanto, apenas 32% da previsão do jogador estava dentro da faixa aceitável. Ao usar o modelo, deve-se considerar essa margem de erro antes de tirar qualquer conclusão.

Por fim, acredito que a precisão do modelo pode ser melhorada ainda mais adicionando novos atributos ao modelo que possuem uma relação com a taxa de transferência. Ou mesmo combinando ou processando atributos (também conhecido como *Feature Engineering*). Experimentar com diferentes modelos ou parâmetros também pode melhorar as previsões.

**REFERÊNCIAS**

BAESSO, Robson de Souza; SOUZA, Antônio Artur de. **O coeficiente de determinação é uma medida confiável do poder explicativo de modelos de estimação do valor intrínseco das ações negociadas na BOVESPA?** São Paulo-SP: Fundação Instituto de Pesquisas Contábeis, Atuariais e Financeiras. 2010.

FILHO, Luiz Lima. **Correlação e Regressão**. João Pessoa-PR: UNIVERSIDADE FEDERAL

DA PARAÍBA. 2013.

FORBES. Market size of the European professional football market from 2006/07 to 2018/19. **Forbes**. Disponível em: <<https://www.forbes.com/soccer-valuations/list>>. Acesso em: 21 de maio de 2021.

GADIYA, Karan. FIFA 19 complete player dataset. **Kaggle**. Disponível em: <<https://www.kaggle.com/karangadiya/fifa19>>. Acesso em: 30, junho e 2021.

GUIMARAES, Paulo. **Análise de Correlação e medidas de associação**. Curitiba-PB: Universidade Federal do Paraná. 2017.

KENNEDY, Peter. **A Guide to Econometrics**. Massachusetts: Blackwell Publishing, 2008.

LANGE, David. Market size of the European professional football market from 2006/07 to 2018/19. **Statista**, 26 de nov. de 2020. Disponível em: <https://www.statista.com/statistics/261223/european-soccer-market-total-revenue/>. Acesso em: 21 de maio de 2021.

LANGE, David. Spending on men’s professional soccer transfers worldwide from 2012 to 2019. **Statista**, 26 de nov. de 2020. Disponível em: <<https://www.statista.com/statistics/1126884/spending-fifa-transfers-worldwide>>. Acesso em: 21 de maio de 2021.

OLIVEIRA, Bruno. COEFICIENTES DE CORRELAÇÃO. **Operdata**, 23 de ago. de 2019. Disponível em: <<https://operdata.com.br/blog/coeficientes-de-correlacao/>>. Acesso em: 3 de junho de 2021.

OLIVEIRA, Ivnna Gurniski de; CHATALOV, Renata Cristina de Souza. **Estatística** Maringá-Pr: Unicesumar, 2018..

RODRIGUES, Sandra. **Modelo de Regressão Linear e suas Aplicações**. Covilhã: Universidade Da Beira Interior. 2012.