**KENNETH DE COSTER**

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

Vitória

05/2021

**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:

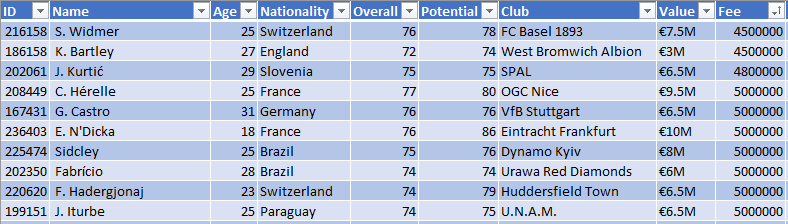
1. SELL, I. **UTILIZAÇÃO DA REGRESSÃO LINEAR COMO FERRAMENTA DE DECISÃO NA GESTÃO DE CUSTOS**. IX Congresso Internacional de Custos. Florianópolis, SC, Brasil. p.14. 2005.
2. BASTOS, E.V.P. **MODELO DE REGRESSÃO LINEAR PARA ANÁLISE DE INVESTIMENTOS EM UMA EMPRESA DO RAMO**.Centro de Instrução Almirante 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 2019-2020 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.

3.1 CRONOGRAMA

Tabela 2 - Cronograma

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Atividades** | jul | ago | set | out | nov | dec |
| Limpeza e Tratamento de Dados | **X** | **X** |  |  |  |  |
| Calcular a correlação linear |  | **X** | **X** |  |  |  |
| Obter o coeficiente de determinação |  |  | **X** |  |  |  |
| Aplicar regressão linear múltipla |  |  | **X** |  |  |  |
| Prever os preços dos jogadores usando o modelo |  |  |  | **X** |  |  |
| Validar os resultados |  |  |  | **X** | **X** |  |
| Desenvolver aplicação de linha de comando |  |  |  |  | **X** | **X** |

Fonte: o autor

**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.