**Projeto 3 – Ciências dos Dados**

**Nome:** João Guilherme 2°B

Samuel Granato

Vitor Pereira

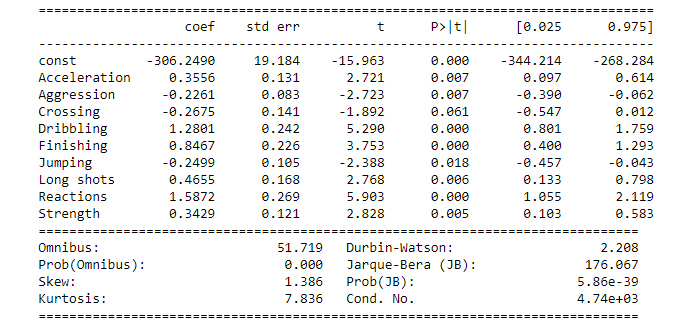
**Introdução**

O projeto em pauta tem como objetivo a análise de um database, composto pelos scout’s (habilidades) dos jogadores do jogo ‘FIFA 18’, a fim de utilizar o método de regressão linear para prever o valor de mercado de cada jogador, baseado em suas características e o apelo que o mesmo possui na mídia.

Para se fazer tal prognóstico, avaliou-se diferentes habilidades de acordo com as posições de cada jogador. Abaixo segue uma tabela que lista quais scout’s foram importantes para prever o preço de cada jogador de acordo com a sua função dentro de campo:

|  |  |
| --- | --- |
| *Acceleration* | *Aggression* |
| *Crossing* | *Dribbling* |
| *Finishing* | *Jumping* |
| *Long shots* | *Reactions* |
| *Strength* |  |

**Análise e Desenvolvimento**

A escolha dos scout´s foram pré-determinadas a partir do valor de P dos resultados da regressão de OLS, de acordo com uma zona de rejeição bicaudal com um α = 5%. Isso significa que, caso P seja > - 0.025 ou < 0.975, H0 não é rejeitado, impossibilitando a regressão linear.

*Fig 1. Resultados da Regressão OLS*

Para estimar o preço do jogador utilizou-se o método da regressão linear, cujo conceito está relacionado a estudar uma variável Y em função de outras N variáveis. Isto é, no caso do projeto, é avaliar o valor de mercado do atleta de acordo com as suas habilidades.

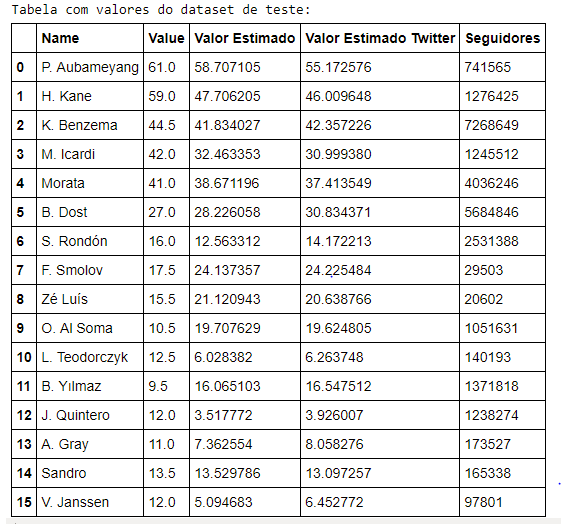
Além dos scout’s, outro fator que ponderamos para considerar o valor estimado é o apelo social que o jogador possui na mídia. A mensuração de tal vertente é avaliada através da quantidade de seguidores que o atleta possui na rede social “Twitter”. Logo, quanto maior o número de seguidores, mais o jogador irá custar na avaliação feita.

Para recolher a quantidade de seguidores de uma lista de mais 10.000 jogadores, foi utilizado um algoritmo conhecido como *Web Crawler*, cujo funcionamento consiste no download de uma página da Web e em uma posterior filtragem para obter a *string* desejada. Para obter o endereço do *Twitter* de cada jogador, foi feito o download da pesquisa do Google com o seguinte título: “nome do jogador (presente no database) + Twitter + Oficial”. Com isso, o algoritmo filtrava as linhas do arquivo html com o endereço da página do Twitter e removia os resultados com "/search", pois esse termo estendia a string a uma pesquisa do Twitter. Da string resultante foi recolhida a primeira palavra que vinha acompanhada com “@”, que indica o nome de usuário do jogador. Logo após, foi feito o download da página do Twitter: "twitter.com/["endereço do jogador"] e, utilizando o segundo algoritmo, houve um filtro pela classe: "ProfileNav-item ProfileNav-item--followers" que contém a string do número de seguidores.

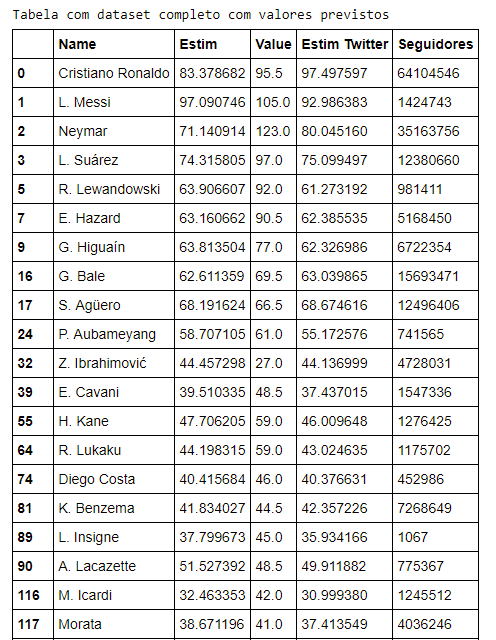
Com as features (scout’s) separadas e o número de seguidores na rede social, o último processo necessário para a análise desejada seria a regressão linear, utilizando o algoritmo presente na biblioteca scikit-learn.

Primeiramente devemos indicar duas partes do dataset: "target", sendo esta a coluna que se deseja prever, e a outra "data", que contém as features que serão usadas no modelo. Após essa divisão, são criados dois set’s, um de treinamento e outro de teste (vale ressaltar a necessidade de treinar e testar o modelo para conquistar uma avaliação consistente). Em seguida, deve-se inicializar o modelo de regressão linear em uma variável, dar fit com o target e data, e por fim, o predict na lista de teste, que contém a matriz com os valores previsto pelo modelo.

**Resultados e Conclusão:**

Os resultados obtidos após a regressão podem ser observados a seguir, em uma tabela contendo o dataset de teste e o dataset com os valores previstos: Value (valor de mercado); Valor estimado (modelo aplicado sem contar os seguidores); Valor Estimado Twitter (modelo aplicado com os seguidores); Seguidores.

*Fig 2. Dataset de Teste*

*Fig 3. Dataset com resultado previsto*

A partir desses dados podemos tomar como conclusão que em alguns casos o valor de mercado real de cada jogador não reflete apenas em suas habilidades. Temos exemplos dos dois extremos, jogadores supervalorizados, ou seja, aqueles que seu valor de mercado é maior do que as qualidades técnicas apresentadas em campo, e jogadores desvalorizados, aqueles que seu valor é menor do que o esperado.

Essa valorização positiva ou negativa, ocorre por variados fatores, entre eles temos: os fatores extracampo, como por exemplo premiações, seja ela de melhor jogador de uma competição, ou de uma temporada, o fator “promissor”, quando o atleta é novo, conquistas coletivas por algum clube ou seleção, entre outros.

E o mais importante para a maioria dos jogadores, o fator “divulgação na mídia”, que pode aumentar o valor de mercado de um jogador caso seu nome seja constantemente citado na mídia, apareça em muitos anúncios, eventos, etc. Isso acaba por aumentar a própria visibilidade do jogador e a do seu clube, aumentando assim o número de torcedores e patrocínios.

Logo, quando um clube se interessa por um atleta com representatividade na mídia, ele terá que desembolsar uma quantia maior que o valor real de mercado do jogador, justamente por causa do impacto que tal contratação causaria na imagem do clube. Da mesma maneira que ocorre quando um jogador é bem falado na mídia, quando seu nome aparece de maneira negativa, prejudica a imagem do clube, sendo assim, seu valor de mercado cai.

Utilizando nossa pesquisa podemos identificar alguns jogadores supervalorizados e outros desvalorizados. No topo da lista dos supervalorizados está o brasileiro Neymar, recém transferido para o Paris Saint Germain, teve seu valor de mercado mais que duplicado para sua venda para o time francês. Especialistas justificam esta supervalorização pelo fato do atacante ter levado o Brasil ao inédito título olímpico, sua idade relativamente baixa, considerando o futebol jogado nas ultimas temporadas, por isso o fator “promissor”. E principalmente pela habilidade apresentada, jornalistas e dirigentes dão como certa a conquista do premio individual de “melhor do mundo”, que se acontecer de fato, trará muita visibilidade para a marca PSG, justificando o investimento feito.

Apesar de apresentar um resultado satisfatório, o modelo vigente contou com duas iterações que foram fundamentais para o desenvolvimento do projeto. A primeira simplificação foi que o valor estimado pela regressão linear agregou apenas jogadores acima de R$1.000.000,00. Isso deve-se ao fato de que como a regressão comporta-se como uma reta de uma equação linear, parte dela apresentou resultados negativos, caso que não é coerente com a realidade.

A segunda simplificação consiste na filtragem por posição. Isso significa que o modelo estimou o preço somente de algumas posições, as quais são:

* RM - Meia Direita CAM - Meio Campo Ofensivo
* LM - Meia Esquerda RW - Ala Direito
* LW - Ala Esquerdo CF – Centro Avante
* ST - Atacante

As posições foram filtradas da maneira supracitada pois as features escolhidas eram mais condizentes com posições de meio-campo/ataque. Ou seja, não seria realista prever o preço de um marcador avaliando o seu *Dribbling, Finishing e Long Shots,* por exemplo.

As discrepâncias observadas eram algo esperado devido a fatores já citados anteriormente. Contudo, adotando apenas habilidades e visibilidade na mídia, o resultado foi contundente, visto que ele foi proporcional com esses dois fatores.

* **Divisão de Trabalho**

Os três contribuíram de maneira uniforme no relatório acima;

Na parte do código, houve uma divisão (não tão uniforme) dos trabalhos. Enquanto João e Vitor trabalharam mais na parte de filtragem dos dados e separar a base de treinamento e teste, Samuel ficou com a parte mais pesada de fazer a regressão Linear. Entretanto, todos foram importantes em alguma parte do trabalho e puderam ajudar a finalizar todo o projeto.