Projeto Final – Ciências dos Dados

**FIFA 2018 – Value Prediction**

João Guilherme de Araújo

Victor Demetrio Morales Habib

Victor Jurdi Gomes

Fábio Ayres | 2-A

São Paulo, 2018

**- Introdução e Objetivo**

O projeto final da disciplina de Ciência dos Dados baseia-se em analisar usando qualquer técnica aprendida um tema livre a escolha do grupo, o dataset escolhido para estudo foi de jogadores de futebol presentes no jogo FIFA 2018.

A análise exploratória foi feita com base no salário, idade, pontuação geral do jogador, corrida, etc. com a finalidade de estimar, mediante a essas e a um total de 39 categorias diferentes, o valor de mercado de qualquer jogador presente na base de dados usando diferentes técnicas para melhor precisão.

**- Bibliotecas usadas**

* Pandas:

Usado para manusear e alterar a base de dados quando foi necessário.

* Numpy:

Usado para deixar algumas variáveis em logaritmos para melhor análise dos gráficos.

* Seaborn:

Usado para plotar um Heatmap das habilidades.

* Matplotlib.pyplot:

Usado para plotar os gráficos das análises.

* Statsmodels.api:

Função utilizada para fazer regressão com constante adicionada.

* Sklearn:

Para treinar o modelo.

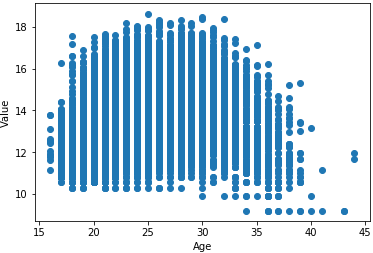
**- Organização do Notebook**

O projeto se organiza primeiramente na importação das bibliotecas necessárias já mostradas anteriormente, em seguida é feita uma “limpeza” removendo itens que não podem ser lidos pelo programa e que não são úteis para a análise feita, a foto do jogador ou sua nacionalidade por exemplo, outra melhoria no dataset foi a remoção do “NaN” (Not a number) que existia em algumas atribuições.

Foram feitas também alterações no dataset a primeira foi a remoção de letras que acompanhavam alguns números no salário dos jogadores estes continham as letras K e M que simbolizam 1.000 e 1.000.000, respectivamente e o símbolo da moeda em uso “€” (euro), em skills foi necessário remover alguns valores que estavam mal formatados, por exemplo uma das células continha o valor “79+2” que não era possível ser rodado pelo programa e por fim as posições preferidas dos jogadores foi alterada para no máximo uma posição, duas ou mais posições causava problemas na análise futura pois foi feito um agrupamento por posição de cada jogador.

Em seguida foi feita uma análise exploratória das variáveis para ver o quão relacionadas elas eram com o nosso Y (preço do jogador ou Value):

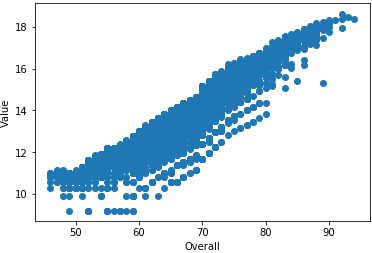
Value X Age:

O gráfico gerado pela correlação das duas variáveis é mostrado ao lado e demonstra que a idade não possui uma correlação forte com o Y como será evidenciado melhor posteriormente.

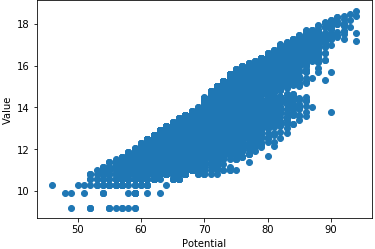
Value X Skills:

Em habilidades existem diversas categorias como chute, reflexos, controle de bola, etc. Foi gerado um gráfico para cada uma das habilidades, que somam 34 no total, que podem ser vistas no Notebook. Nesta categoria é necessário enfatizar que 5 dos gráficos gerados causam curvas estranhas e não se assemelham aos outros gráficos, isso acontece pelo fato dessas habilidades serem exclusivas de goleiros que tem atributos diferentes dos outros jogadores em geral.

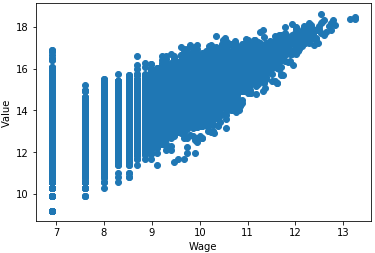
Value X Overall:

Atributo dado após a análise de todas as habilidades do jogador, esta categoria vem predefinida no dataset e com isso estima-se que ela seja derivada de uma regressão feita pelos próprios desenvolvedores do jogo, isso se prova pela forte relação que ela possui com o Y, como mostrado ao lado.

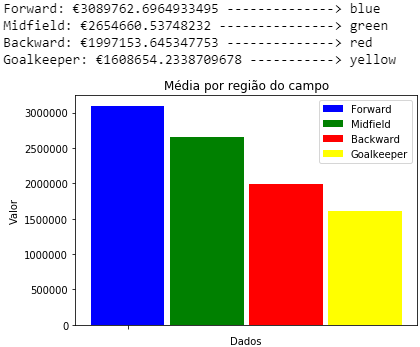
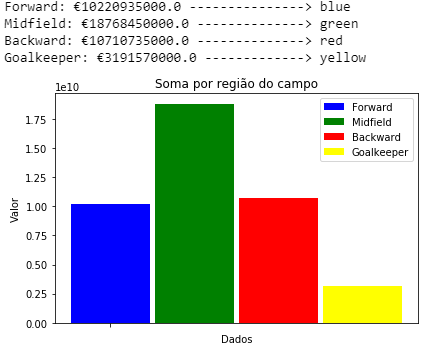
Value X Potential:

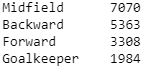
Que é o quanto um jogador ainda pode evoluir, como mostrado no gráfico possui uma boa relação.

Value X Wage:

O gráfico gerado mostra também com algumas exceções uma boa correlação com o Y

Value X Preffered Positions:

Aqui foram feitas outras análises com o intuito de perceber como o valor dos jogadores era distribuido conforme a sua posição de atuação, foram gerados gráficos da soma e da média de seus valores pela sua região no campo.

Para melhor ilustração gerou-se também a quantidade de jogadores de cada uma das posições dos gráficos, como mostrado abaixo.

A partir da análise dos gráficos e dos dados coletados da célula acima, percebe-se que os jogadores mais ofesivos ("Forwards") são, em geral, os mais valorizados do jogo, pois apresentam a média mais alta quando comparada com as outras posições, mesmo tendo menos representantes atuando nessa região do campo. Entretanto, pode-se dizer que há poucos jogadores considerados muito caros e muitos considerados muito baratos. Esse é o mundo do futebol, onde é raro alcançar o sucesso.

Após todas essas análises, o notebook segue para o modelo que será discorrido em outro tópico.

Os procedimentos são discorridos por completo em forma de tópicos abaixo:

1. Importar as bibliotecas python necessárias para a realização do projeto
2. Dataset
   1. Importar o Dataset
   2. Limpar o Dataset

* Retirar as colunas que não são importantes para a realização do projeto:
  + 'Name', 'Photo', 'Nationality', 'Flag', 'Club', 'Club Logo', 'Special', 'ID', 'CAM', 'CB', 'CDM', 'CF', 'CM', 'ID', 'LAM', 'LB', 'LCB', 'LCM', 'LDM', 'LF', 'LM', 'LS', 'LW', 'LWB', 'RAM', 'RB', 'RCB', 'RCM', 'RDM', 'RF', 'RM', 'RS', 'RW', 'RWB', 'ST'.
  1. Fazer as alterações necessárias no Dataset
* Coluna Value e Wage:
  + Retirar o símbolo da moeda (€).
  + Transformar todos os valores em números (multiplicar por 1000 aquels que acompanham a letra K e por 1000000 aqueles que acompanham a letra M).
* Skills:
  + Retirar valores que não eram lidos (Ex.: “79+2”).
* Preffered Postion:
  + Reduzir o número de posições para um.

1. Relações entre variáveis
   1. Value X Age
   2. Value X Skills
   3. Value X Overall
   4. Value X Potencial
   5. Value X Wage
   6. Value X Preffered Position
2. Modelo
   1. Com todas as features  
      4.1.1 Linear Regression  
      4.1.2 Random Forest Regression  
      4.1.3 Importance Features
   2. Sem Overall  
      4.2.1 Linear Regression  
      4.2.2 Random Forest Regression  
      4.2.3 Importance Features
   3. Só com Overall  
      4.3.1 Linear Regression  
      4.3.2 Random Forest Regression
   4. Apenas com Skills  
      4.4.1 Linear Regression  
      4.4.2 Random Forest Regression  
      4.4.3 Importance Features

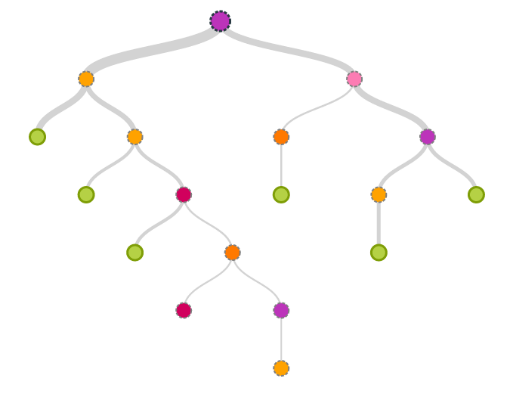
**- Metodologia**

O processo em que se tem o interesse de analisar o comportamento de uma variável a partir de outra é conhecido como regressão. Ele apresenta diversas maneiras de resolução, dentre elas, métodos que levam o nome de regressão linear, múltipla ou Random Forest surgem para suprir os diferentes tipos de problemas. Assim como na maioria dos casos, em nosso classificador é necessário mais de uma variável para obter-se melhor acurácia no resultado final. Dessa forma, descarta-se a ferramenta de regressão linear simples e adere-se aos dois outros métodos: regressão linear múltipla e Random Forest Regression.

Entretanto, existem diferenças tanto conceituais, na forma de resolução, quanto empíricas, no resultado. Para discorrermos a partir dessas diferenças, é válido expor as duas maneiras de resolução que esses métodos aderem e em seguida, junto com o projeto, justificar a diferença nos resultados obtidos.

A Multiple Linear Regression provêm da Simple Linear Regression, com a diferença de que ela trabalha com mais de uma variável. Apesar disso, o método é mantido e realizado da seguinte forma: de forma mais informal e pontual, pode-se dizer que os dados são agrupados em uma reta. Em outras palavras, admitindo-se que X1 , X2 , ..., Xk sejam as variáveis independentes e Y a variável dependente, dada uma amostra de n observações, [x1i, x2i, x3i, … , xki, Yi] com i =1, 2, 3, … , n, o modelo de Multiple Linear Regression será dado por: Yi = **β**0 + **β**1\*x1i + **β**2\*x2i + ... + **β**k\*xki ou seja, Yi = constante + coeficiente 1 \* x1i + coeficiente 2 \* x2i +...+ coeficiente k \* xki.

O Random Forest Regression, por sua vez, utiliza um método diferente, classificado popularmente por árvore de decisões.

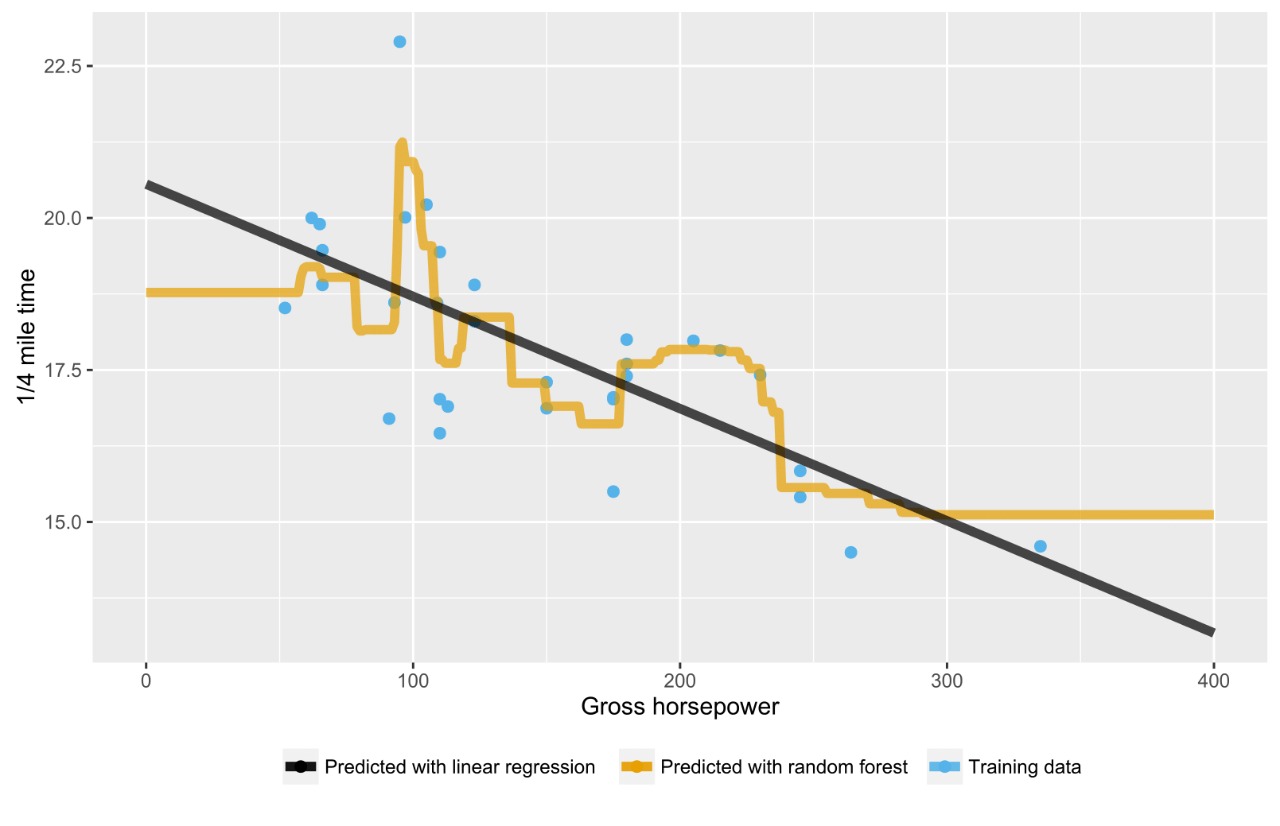


A primeira etapa do processo de Random Forest resume-se em sortear aleatoriamente um valor de sua variável X e recorrer ao método da árvore de decisões para encontrar o respectivo valor de Y.

A árvore de decisões funciona da seguinte forma: em cada ramo da estrutura, o valor sorteado aleatoriamente do modelo pode prosseguir dois caminhos distintos até que seu respectivo valor Y seja encontrado e seja aproximado para seus vizinhos, como uma constante.

Entretanto, vale ressaltar que esse processo ocorre milhares de vezes até alcançar uma função que aproxima ao máximo dos dados reais, impedindo que resulte em uma função afim em que para todo X exista um Y.

Comparando os métodos de resolução de ambos os processos, pressupõe-se que a ferramenta Random Forest Regression é mais eficaz do que a Multiple Linear Regression por motivos, como: a segunda estima uma função retilínea, o que pode ser prejudicial em casos em que os dados não segue uma distribuição regular, enquanto a primeira fornece uma função que parece caminhar junto aos dados reais. Além disso, pode-se dizer que essa característica enaltece a ideia de que este método diminui a variância do modelo, pode-se observar isso no exemplo abaixo.

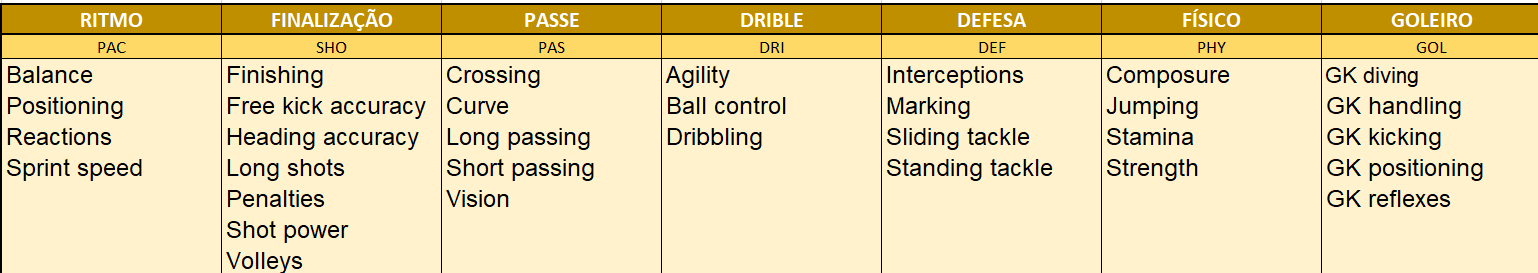


À primeira vista, analisando o problema e percebendo a quantidade de dados e de variáveis que ele apresenta, a conclusão de que o método Random Forest Regression é mais eficaz, se fortalece.

Entretanto, até agora estamos no campo das hipóteses, que serão validadas ou questionadas com a apresentação dos resultados mais à frente.

**- Materiais**

Para alcançarmos da forma mais precisa nosso objetivo, as seguintes categorias de habilidades, que denominamos como o grupo de “skills”, foram utilizadas:



Goalkeeper Card

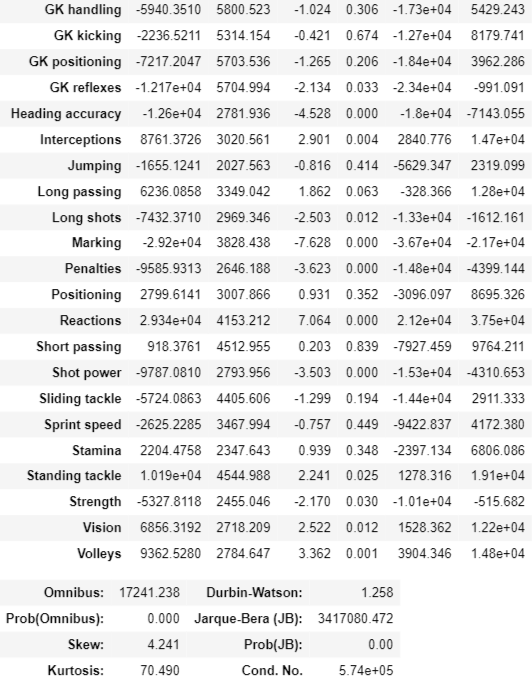
Player Card

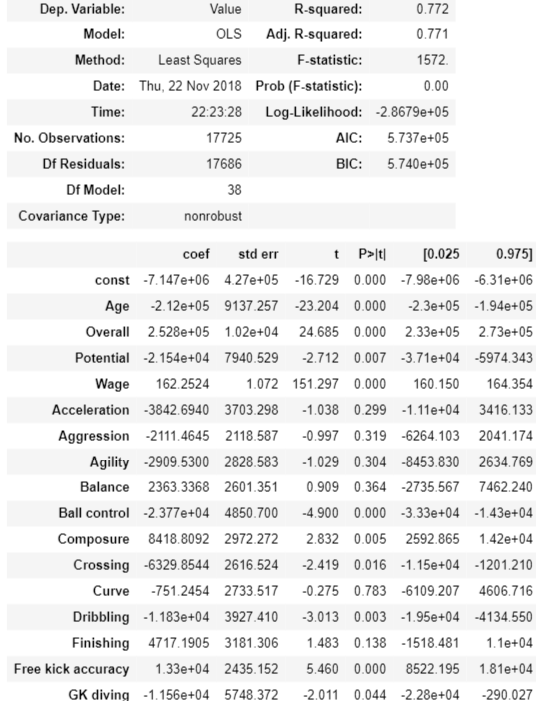


Além dessas, mais 4 features foram utilizadas para uma melhor acurácia do resultado:

* Age (Idade)
* Potential (Potencial)
* Wage (Salário)
* Overall (Nível geral)

**- Resultados**

* 1. Todas as Features



As variáveis:

- Overall

- Potential

representam cerca de **92.00%** na determinação do valor de mercado do jogador

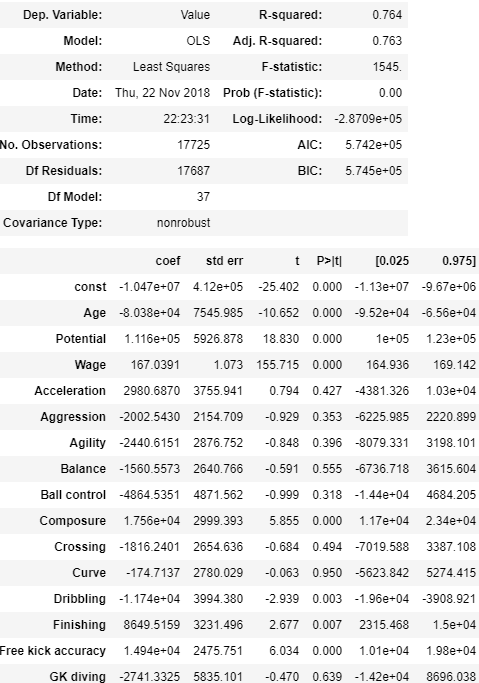
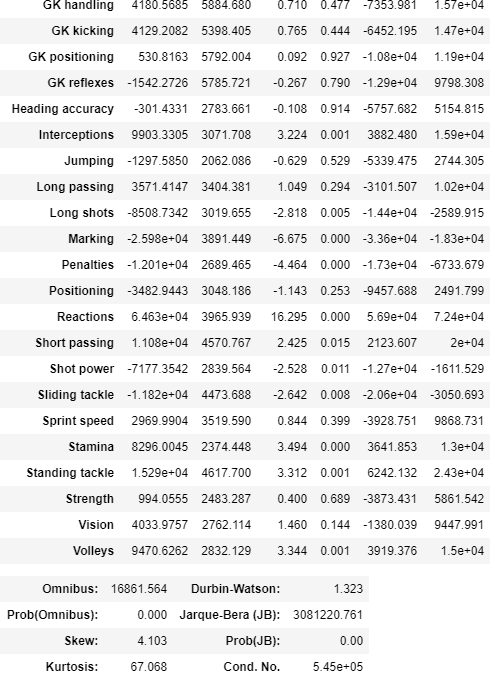
**Linear Regression**

R-squared: 0,75

**Random Forest Regression**

R-squared: 0,95

* 1. Sem Overall



As variáveis:

- Wage

- Potential

representam cerca de **80,00%** na determinação do valor de mercado do jogador

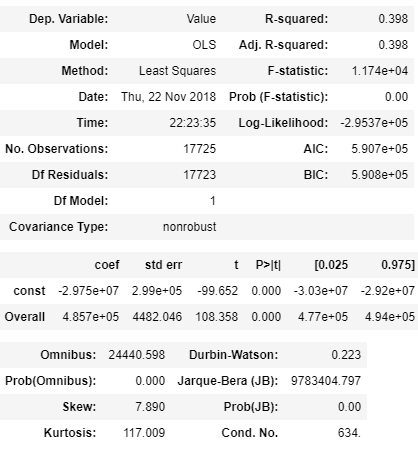
**Linear Regression**

R-squared: 0,74

**Random Forest Regression**

R-squared: 0,93

* 1. Só com Overall



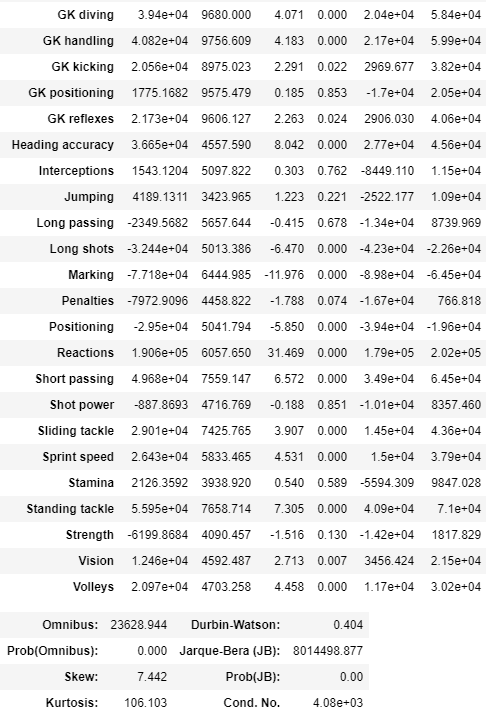
**Linear Regression**

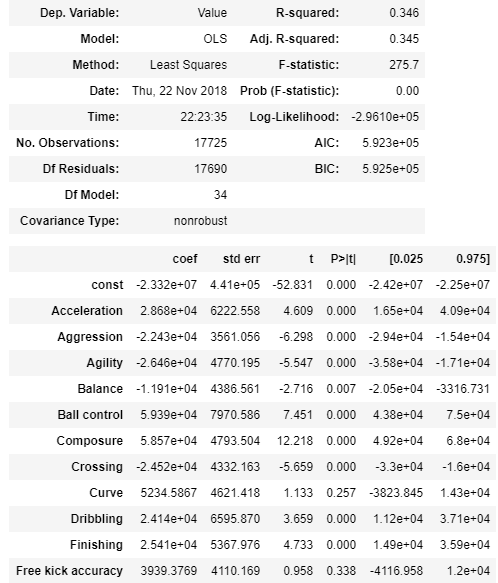
R-squared: 0,40

**Random Forest Regression**

R-squared: 0,85

* 1. Apenas com as Skills





As variáveis:

- Reactions

- Ball Control

representam cerca de **70,00%** na determinação do valor de mercado do jogador

**Linear Regression**

R-squared: 0,35

**Random Forest Regression**

R-squared: 0,83

**- Conclusão**

Concluímos, portanto, que a regressão linear múltipla não promovia a melhor análise para o nosso problema, e sim, a Random Forest Regression, pois essa se adapta melhor às diversas variações do grande dataset em estudo. O modelo foi subdivido para avaliar quatro datasets com a ausência de algumas categorias, usando primeiro a regressão linear e posteriormente a Random Forest Regression em cada uma das avaliações. Primeiramente, foram avaliadas todas as features (Salário, Potencial, Idade, Nível e Habilidades), obtendo um percentual de 75,66% de acurácia no primeiro método, e 96,36% com o segundo, deixando evidenciada a superioridade de análise pelo Random Forest Regression com o dataset em questão. Além disso percebe-se que o resultado obtido deste modelo completo, restringe-se grande parte à feature overall que apresenta uma influência de 86,54% no comportamento da variável Value, ou seja, mostra o quanto essa característica influencia na determinação do valor do jogador, o que dá suporte à hipótese de que este seja um número gerado a partir de uma regressão feita anteriormente pelo jogo, ou um valor já feito de forma a ter grande correlação com o Value, como foi mencionado antes no gráfico mostrado em Organização do Notebook (Value X Overall).

A segunda iteração foi feita sem Overall e obteve quase os mesmos resultados, 74,69% e 94,19%, no entanto, agora a influência no resultado foi de 48,24% do Wage e 31,81% de Potencial o que mostra uma maior distribuição entre as features, algo que já era esperado, pois salários maiores e um potencial de crescimento maior mostram que o jogador vale mais e que a relação entre Value X Wage e Value X Potential são muito mais correlacionadas que as demais.

A terceira iteração foi feita apenas com o Overall, no intuito de analisar as questões levantadas em momentos anteriores quanto a correlação intrínseca do Value com o Overall. Assim, os resultados obtidos foram 40,3% com Regressão Linear Múltipla e 86,63% usando Random Forest Regression, o que é menos do que o obtido nas últimas iterações. No entanto, é válido ressaltar que esta é uma categoria sozinha sendo analisada com todo o dataset, e assim, os resultados obtidos não foram ruins, concluindo que existe uma correlação muito alta entre as duas features.

A quarta e última iteração foi feita apenas com as Skills, que contém 34 categorias diferentes, obtendo 34,75%, resultado que já era esperado, pois os números são muito dispersos entre si e não são bem analisados por uma Regressão Linear Múltipla. Entretanto, atingiu-se 82,94% usando Random Forest Regression, o que mostra, como visto nos gráficos, exceto os 5 de goleiros, que há uma grande correlação entre as Skills do jogador e o seu valor de mercado. Analisando quais skills individuais são mais levadas em conta, chega-se à conclusão com 52,41% de relevância que as “Reactions” do jogador influenciam mais da metade de seu valor de mercado, seguido por “Ball Control” com 15,04% de influência.

**- Referências Bibliográficas**

<https://www.statisticshowto.datasciencecentral.com/bootstrap-sample/>

<https://en.wikipedia.org/wiki/Random_forest>

<https://books.google.com.br/books?hl=en&lr=&id=0yR4KUL4VDkC&oi=fnd&pg=PR13&dq=linear+regression&ots=p5oqFmlQvl&sig=DjPRZEuSoUUQPimGqBRkBnv2upA#v=onepage&q=linear%20regression&f=false>

<https://www.researchgate.net/profile/Andy_Liaw/publication/228451484_Classification_and_Regression_by_RandomForest/links/53fb24cc0cf20a45497047ab/Classification-and-Regression-by-RandomForest.pdf>

<https://pandas.pydata.org/>

<http://www.numpy.org/>

<http://hedibert.org/wp-content/uploads/2014/02/Econometria201401-Aula04-ARLM-I-Estimacao.pdf>