Insper Instituto de Ensino e Pesquisa

Fernando Cesar Furtado Ballesteros Fincatti

Gabriel Pascua de Freitas Moreira

Pedro Dinacci Célia

PROJETO FINAL: CIÊNCIA DOS DADOS

São Paulo

2019

Insper Instituto de Ensino e Pesquisa

Fernando Cesar Furtado Ballesteros Fincatti

Gabriel Pascua de Freitas Moreira

Pedro Dinacci Célia

PROJETO FINAL: CIÊNCIA DOS DADOS

Trabalho elaborado pela disciplina Ciência dos dados, sob supervisão dos Professores Orientadores:

Fábio Miranda
Bárbara Agena

São Paulo

2019

SUMÁRIO

1 INTRODUÇÃO

O projeto tem como objetivo realizar previsões de vendas de jogos e comparar os resultados obtidos, por meio da análise de dados de um dataset, na qual estão contidas informações diversas, como: gênero, desenvolvedora, plataforma, ano de lançamento ect; sobre inúmeros games.

Esta análise irá interessar a possíveis desenvolvedores e destribuidoras para aumentar os lucros de suas respectivas empresas.

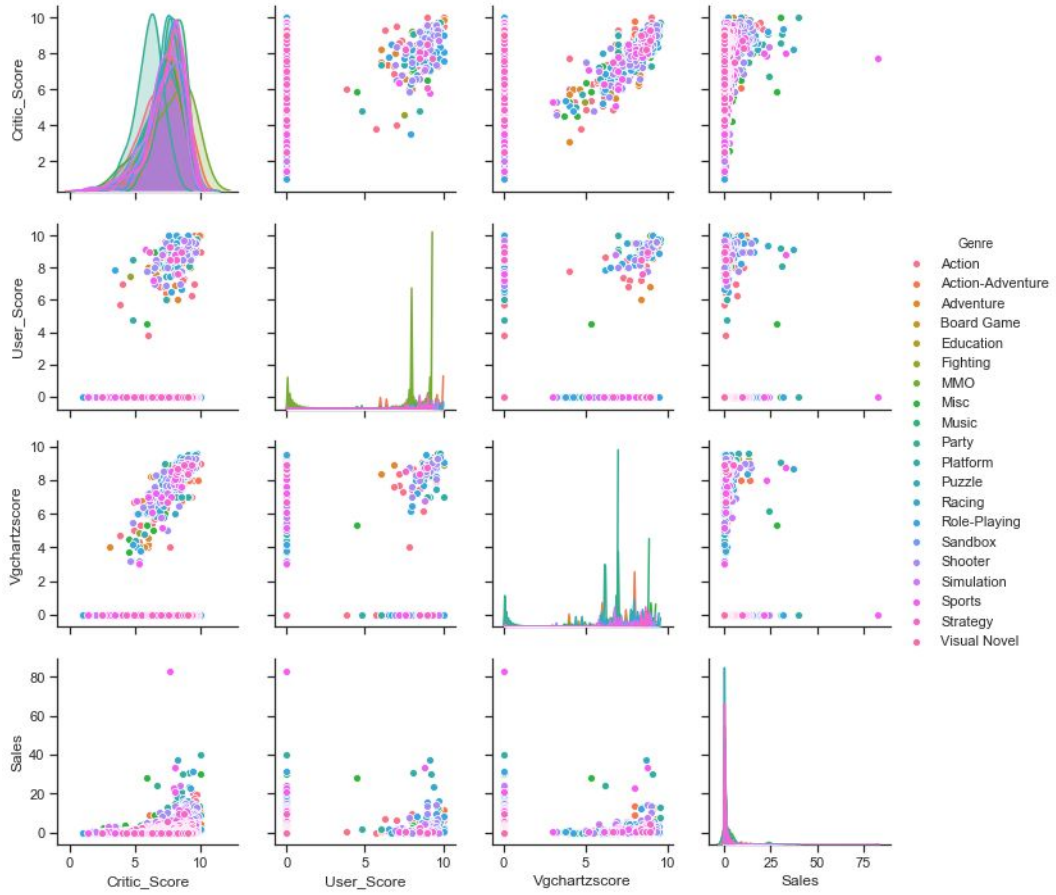
Para realizar as previsões foram utilizados três métodos de regressões diferentes: SGDRegressor, Random Forest Regressor e Linear Regression.

2 ANÁLISE EXPLORATÓRIA

A prática de análise exploratória faz parte das etapas imprescindíveis para uma melhor eficiência das análises estatísticas. Desta maneira, antes de aplicar os métodos de regressão, fez-se necessário compreender as informações disponíveis no dataset utilizado, bem como a influência e relação entre elas. (OESTATÍSTICO, 2015)

Para realizar a análise foram gerados inúmero gráficos que relacionam as diversas coluna da planilha excel, o resultado é demonstrado na Figura1.

Figura1 - Análise exploratória



Fonte: elaborado pelos autores

Tendo vista que um dos objetivos do projeto é a predição de vendas de jogos, através das conclusões tiradas dos resultados obtidos através da análise exploratória, a variável vendas não parece se relacionar de forma clara com apenas uma variável, ou seja, parte-se da hipótese de que existem outros fatores que influenciam no resultado das vendas.

Destarte para as regressões que foram utilizadas para construir os modelos preditivos foram combinadas mais de uma variável a fim de aumentar a taxa de acerto e diminuir os erros.

3 REGRESSÕES

Nessa seção serão apresentados cada um dos métodos de regressão utilizados, bem como o resultado de cada um. Para fins comparativos foi calculado o erro quadrático e a índice de acerto de cada modelo.

3.1 LINEAR REGRESSION

Para a realizar essa regressão inicialmente foi utilizado somente as variáveis quantitativas, em virtude de serem o único tipo de variável a ser usado nesse modelo de regressão.

Após isso, os valores-p das variáveis foram analisados, e aqueles que eram mais do que 10% foram excluídos da análise, dado que isso significa que não havia influência desse parâmetro.

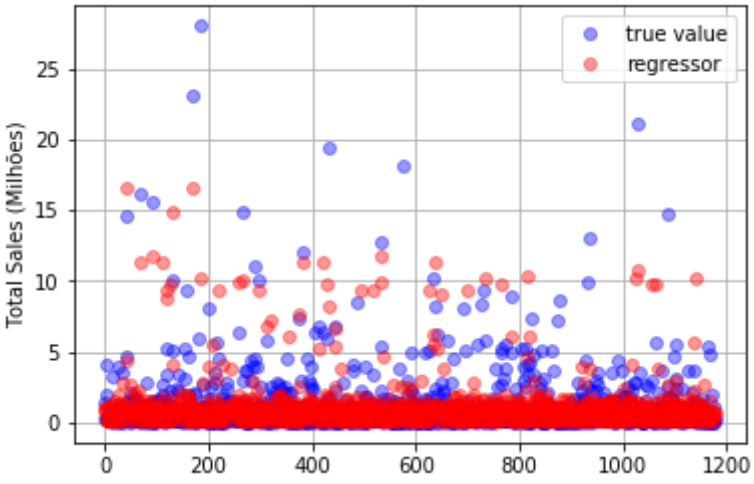
Com essa regressão foi obtido para a quantidade de cópias vendidas, um índice de acerto de 3,40% e um erro quadrático de 0,24, todavia para a nota dos críticos, o índice de acerto foi de 46,30% e o erro quadrático de 0,15.

Visando mais precisão no modelo, foi utilizado o método “Dummy”, que transforma todas as variáveis qualitativas em colunas com valores binários, com 1 sendo usado para definir que o item a presença dessa característica e 0 para a ausência.

Usando funções prontas no python a regressão foi feita novamente e os valores-p foram novamente revistos, ademais as variáveis foram filtradas até sobrarem as que possuíam menos de de 10% de valor-p.

Com a nova regressão foi obtido para a quantidade de cópias vendidas, um índice de acerto de 5,01% e um erro quadrático de 0,30, todavia para a nota dos críticos, o índice de acerto foi de 48,94% e o erro quadrático de 0,15.

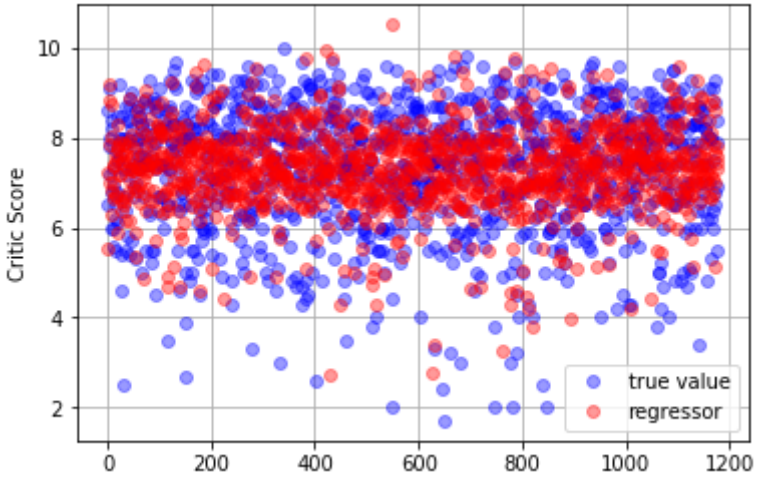
Figura2 - True values Vs Regression predict (Total Sales em milhões)



Fonte: elaborado pelos autores

A Figura2 apresenta o scatterplot dos pontos obtidos pela regressão comparando com os valores reais. O gráfico serve basicamente para ter uma mínima noção da diferença do modelo para os valores reais.

Figura3 - True values Vs Regression predict (Critic_Score)



Fonte: elaborado pelos autores

A Figura3 tem o mesmo objetivo da Figura2, comparar pontos obtidos pela regressão comparando com os valores reais.

3.2 SGD REGRESSOR

Na realização dessa regressão foi inicialmente utilizado apenas a variável de “Critic_Score”, contudo os valores do erro quadrático e índice de acerto foram extremamente insatisfatórios.

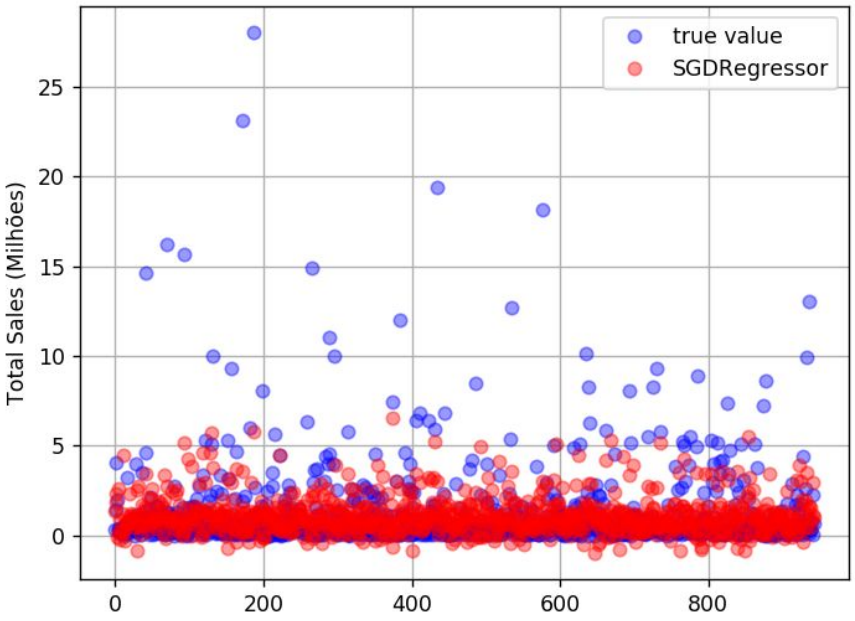
Objetivando uma melhoria da regressão, foi necessário trabalhar o dataset, ou seja, fez-se a remoção de dados nulos e de variáveis que não se demonstravam

relevantes, ademais foi criado uma tabela de junção das vendas dos jogos chamada de “Sales”.

Outrossim foi adicionado mais variáveis, contudo foi preciso utilizar o método “Dummify” visando melhorar a regressão, o valor do erro quadrático foi de 0,237 e o índice de acerto 5,31%.

Um recurso gráfico foi utilizado e está apresentado abaixo na Figura3.

Figura4 - True values Vs Regression predict



Fonte: elaborado pelos autores

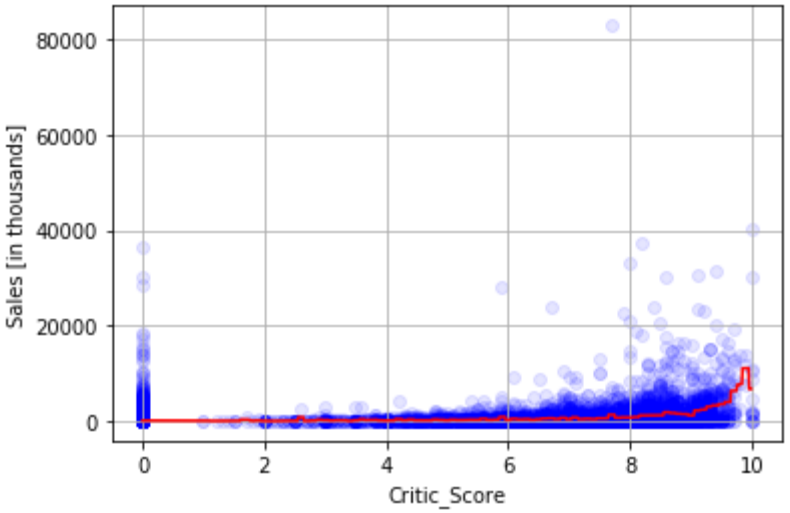
A Figura4 apresenta o scatterplot dos pontos obtidos pela regressão comparando com os valores reais. O gráfico serve basicamente para ter uma mínima noção da diferença do modelo para os valores reais.

3.3 RANDOM FOREST REGRESSION

Na primeira iteração do modelo foi utilizada apenas a variável “Critic_Score” para prever os resultados de vendas. O resultado obtido para este primeiro modelo foi ruim, atingindo um R quadrático de 0.162.

O resultado acima pode ser justificado pela hipótese citada na seção da “análise exploratória”. Em outra palavras para este primeiro modelo era esperado um R quadrado baixo, uma vez que seria difícil achar uma relação entre vendas e somente uma outra variável.

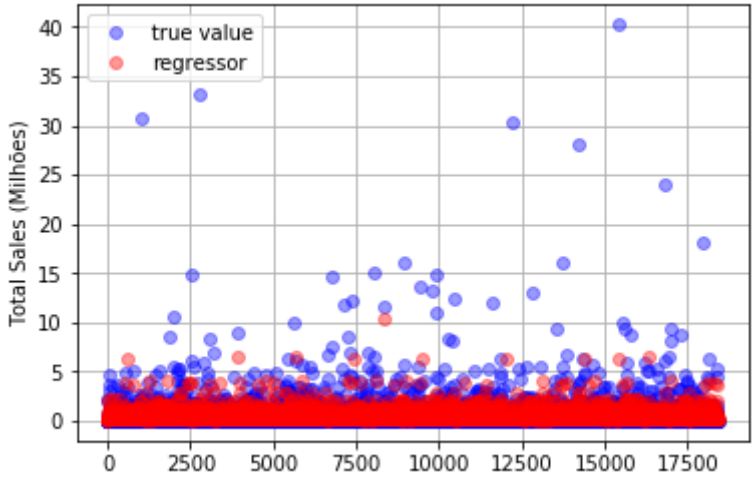
Figura5 - Gráfico da regressão e scatterplot (Critic_Score x Sales)



Fonte: elaborado pelos autores

A Figura5 apresenta o scatterplot (Critic_Score x Sales) dos valores reais retirados do dataset, bem como a linha de regressão construída pelo modelo, representada em vermelho.

Figura6 - True values Vs Regression predict



Fonte: elaborado pelos autores

A Figura6 apresenta o scatterplot dos pontos obtidos pela regressão comparando com os valores reais. O gráfico serve basicamente para ter uma mínima noção da diferença do modelo para os valores reais.

3.3.1 SEGUNDA ITERAÇÃO

A fim de melhorar o resultado do modelo, a segunda iteração foi pensada de modo a utilizar e combinar diversas variáveis para prever o resultado de vendas.

Foram utilizadas as seguintes variáveis: "Critic_Score", "User_Score", "Vgchartzscore", "Genre", "Sales", "Platform","Publisher","Developer" , "Genre", "Platform", "Publisher", "Developer".

Além das múltiplas variáveis o dataset também foi tratado devidamente, aplicando métodos de filtragem, tais como: remoção de colunas irrelevantes e valores nulos.

Utilizando mais variáveis o R quadrado foi melhorado, obtendo um valor de 0.49 e um índice de acerto de 0.083. Sendo assim, a hipótese inicial de que a combinação de mais variáveis ajudariam na previsão do resultado de vendas, foi confirmada.

Como foram utilizadas mais de 3 variáveis torna-se impossível construir gráficos, ou seja, não é possível a visualização gráfica dos resultados.

4 CONCLUSÃO

A partir dos resultados obtidos pelos modelos construídos, pode-se concluir que a base de dados utilizada não é a mais apropriada, pois a mesma não possui informações suficientes para prever o resultado de vendas.

Ademais os resultados obtidos através dos métodos aplicados, não é possível determinar com exatidão e precisão o estilo de jogo que venderá mais nos anos seguintes. Entretanto a regressão que obteve os melhores resultados para previsão de vendas foi a Random Forest Regression, obtendo o maior R quadrático e o melhor índice de acerto, 0.48 e 0.083 respectivamente.

A fim de melhorar as predições, poderiam ser levadas em conta a sazonalidade, investimento em marketing, pertencimento a alguma série e etc.

BIBLIOGRAFIA

https://oestatistico.com.br/analise-exploratoria-de-dados/

https://sdsawtelle.github.io/blog/output/week2-andrew-ng-machine-learning-with-python.html

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html#sklearn.metrics.mean_squared_error

https://www.kaggle.com/