Análise de casas de Nova York – Ciência dos Dados 2B

Regressão Linear Simples

Antonio Andraues

Lucca Constabile

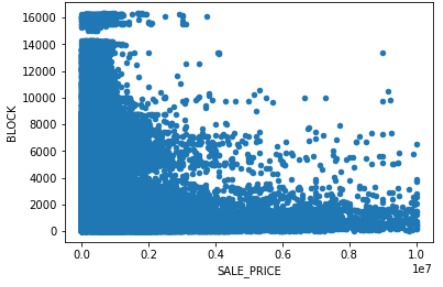
João Gabriel Rodrigues

21/11/2017

Introdução: Definição do tema e ferramental

A equipe à primeira vista optou por trabalhar neste projeto usando cluster. Com o dataset já localizado, "nyc-rolling-sales.csv", a equipe percebeu que há questões de sensibilidades em relação aos algoritmos disponíveis. Partindo do princípio de que o dataset possui informações de boa qualidade, achou-se melhor optar dentre DBSCAN, K-Means e Mean Shift, o algoritmo Mean Shift por ter mais afinidade as nossas váriaveis. Numa rápida lida no database, viu-se a necessidade de definir quais regiões possuem high budget e low budget. Para isso, um elemento necessário e identificar para qual centróide cada ponto está convergindo, dando uma ideia de agrupamento e assim definindo cada bairro com sua income. Entretanto para esse cluster ficar apresentável, sabe-se que um processo de smoothing será necessário, pois mesmo dentro de divisões via renda, ainda assim dentro de uma classe, possui-se diferenças grandes nas rendas e nas despesas. Por isso a escolha do mean shift, que vai pegar cada ponto convergindo e dar uma suavizada nesses resultados, mostrando mais definido cada região

No decorrer da atividade, percebeu-se que os dados não estavam completos, com muitas lacunas com valores incongruentes. O resultado disso foi um scatter com esse padrão:

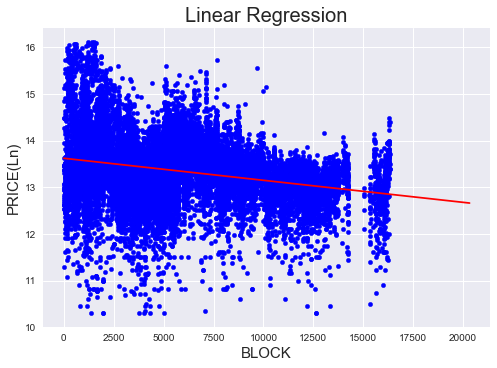


Mesmo trocando para um cluster do tipo gaussian mixture e até pelo próprio k-means não adiantou, uma vez que no gráfico de scatter ficou bem visível a não presença de centros bem definidos nem zonas em comum, portanto foi aconselhado a equipe trabalhar com uma regressão linear.

Regressão Linear: Montando as previsões e a base de treinamento

Para poder analisar o comportamento dos preços (price LN) por quarteirão (block), foi-se necessário dividir 20% dos dados em uma base não treinada para comparar o comportamento, e os 80% dos dados treinados, assim não enviesando as previsões.

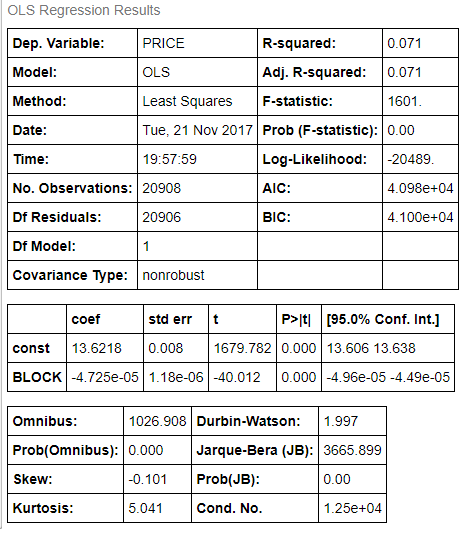
Para a primeira bateria de testes, os 80% comportaram-se da seguinte forma,



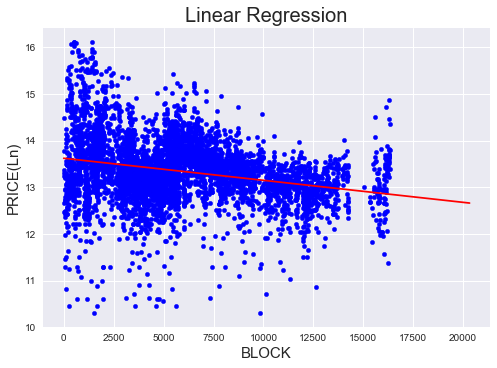
aonde pode-se ver claramente uma correlação, quanto mais longe do quarteirão inicial(0), menor o preço. Os coeficientes deram -14085604.354080722(Sxy), 298137063769.5566(Sxx), 13.621763268939793(Beta0) e -4.724539839490776e-05(Beta1).

Tabela OLS:

A partir da tabela, pode-se tirar várias conclusões. O modelo basicamente explica 7,1% dados, uma vez que o dataset não está bem formulado e não atende todos os requerimentos para um fit de bom nível. Além disso, um coef negativo (-4,725e^5) significa uma correlação inversa. Isto bate com a forma do gráfico, onde fica visível que quanto maior o block, menor o preço das casas e vice e versa. O valor p possuir atributo 0 significa que a chance de o coeficiente ser 0 é nula, o que dá suporte para a validação do modelo, defendendo uma correlação verdadeira entre as variáveis.

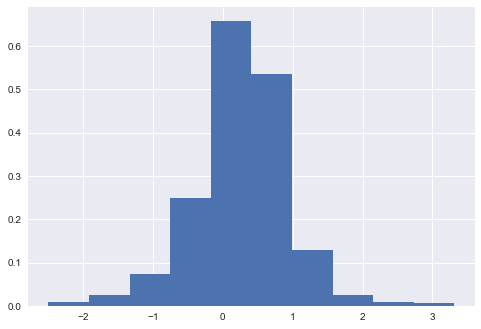


Regressão com os 20% não treinados:



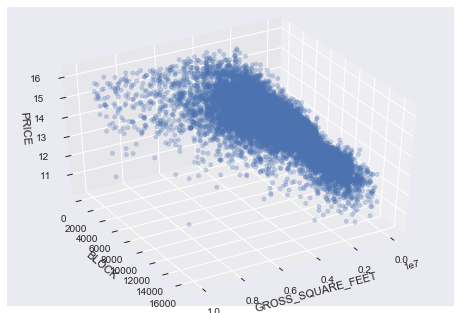
O comportamento dos 20% não treinados está quase igual com o dos 80%, indicando assim uma boa previsão acerca dos dados sem possuir um viés.

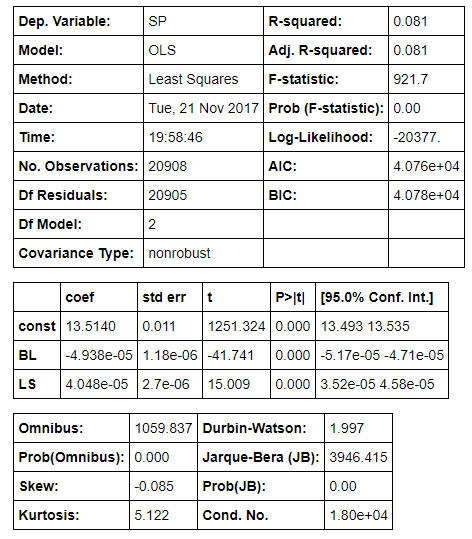
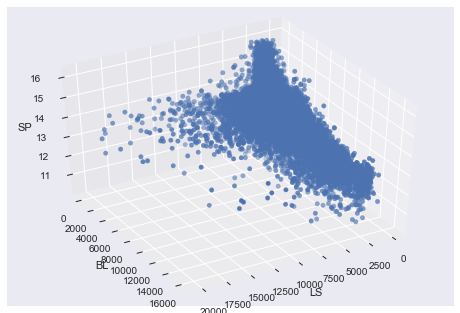
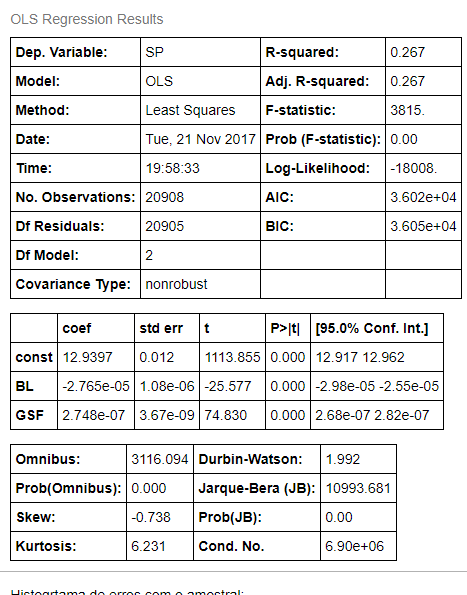
Histograma do Ruído:



O histograma do ruído indica uma distribuição normal com media zero e uma variância constate. Isso indica que a distribuição do modelo e boa e funciona bem.

Gráfico 3D: Gross Square x Land Square





Comparando os dois gráficos, pode-se perceber uma mudança quando ao comportamento do preço em relação a block/square. No gross, que considera o total do terreno, há uma “suavização”, tendo um comportamento mais constante. Já no landsquarre, que considera terreno habitável, já predomina um pico perto do block (0). Isso fica visível no OLS dos dois gráficos, aonde na parte do coef do block para os dois, a correlação de price block e square feet para o gross fica mais acentuada (-4,938e^-5). Já no land, esse valor é -2,75e^-5.

Conclusão:

Durante o desenvolvimento do trabalho, partiu-se de uma ideia que era identificar grupos de casas com preços e assim usar o cluster parar responder a nossa pergunta, que era como era a distribuição do preço das casas em nova York. Porém, chegou-se a uma outra resposta, que foi como estavam agrupadas as casas em torno dos quarteirões e possivelmente como se comporta essa distribuição. Pela regressão e pelos gráficos gerados, ficou claro que as casas mais caras estão em torno dos primeiros quarteirões(block), e que a medida que se afasta dele, mais barato é o valor da casa.