**PREVISÃO DO PREÇO DE VENDA DO ÓLEO DIESEL NO ESTADO DE SÃO PAULO UTILIZANDO REGRESSÃO LINEAR**

**Marco Aurélio Daniel**

**RESUMO**

Neste estudo será abordado um estudo onde será construído um modelo para treinamento de regressão linear múltipla, onde o objetivo é prever o preço de venda médio mensal do combustível Diesel, baseando-se nos preços de vendas de outros combustíveis como Gasolina, Etanol e GNV, e também na cotação do dólar. A base de dados mostrou-se promissora em questão de correlação e o modelo construído teve ótimos resultados estatísticos, como por exemplo um R-squared de 0.984 e um percentual médio de erro de 0.02%. Os valores previstos foram satisfatórios.

**INTRODUÇÃO**

Em 21 de maio de 2018, o Brasil foi marcado por uma crise que fez todo o país parar. Cidades decretando estado de emergência e calamidade pública , devida a falta de abastecimento de alimentos e remédios em todo o país. Escolas suspenderam aulas e provas, transportes públicos foram reduzidos e os combustíveis ficaram cada vez mais escassos, tendo assim um elevado aumento de preço devido a grande procura.

Essa crise aconteceu devido à paralisação dos caminhoneiros autônomos em todo o país, onde os mesmos manifestaram contra os reajustes frequentes e sem previsibilidade dos preços dos combustíveis, principalmente do óleo diesel. Os manifestantes bloquearam inúmeras rodovias do país, causando congestionamento e difícil deslocamento de cargas e pessoas.

A greve durou oficialmente dez dias dias e ocorreu durante o governo de Michel Temer, onde o preço final dos combustíveis eram compostos por 45% de tributação de impostos. Os preços vinham sendo aumentados desde 2017, e um dos motivos dos quais os economistas os explicam é que neste ano a Petrobras passou a acompanhar as oscilações internacionais e a variação do dólar para os preços do petróleo.

Desde então, os preços dos combustíveis veem cada vez sendo mais discutidos e analisados, E como foi visto nesta greve dos caminhoneiros, as oscilações destes preços, principalmente do diesel, podem ter um impacto gigantesco em toda economia do país.

Sendo assim, uma ferramenta que pudesse realizar meras previsões dos preços dos combustíveis seria interessante para tomada antecipada de decisões econômicas que possam prevenir tais crises.

A regressão linear estuda a relação entre uma ou mais variáveis preditoras e uma variável resposta. Esta relação é representada por um modelo matemático que encontra uma reta que melhor representa as variáveis de entrada com a variavel de saida. Deste modo é possível realizar previsões de futuras informações baseando-se nos cálculos de dados anteriores.

Neste trabalho será abordado o desenvolvimento de um protótipo a fim de prever o preço de venda o diesel para o Estado de São Paulo, com base no histórico de preços de outros combustíveis e da cotação do dólar utilizando regressão linear múltipla.

**METODOLOGIA**

Para desenvolvimento e modelagem dos dados foi utilizado a linguagem de programação Python em sua versão 3, juntamente com o ambiente de desenvolvimento de código aberto Jupyter Notebook.

Também foram utilizadas algumas bibliotecas externas disponíveis para a linguagem Python que auxiliam e facilitam o desenvolvimento. São elas:

* pandas - Para utilização de DataFrames
* numpy - Para manipulação de arrays e matrizes
* seaborn - Para plots detalhados de gráficos
* matplotlib - Para plots simples de gráficos
* statsmodels - Para análises estatísticas
* scipy - Para auxílio de cálculos matemáticos
* sklearn - Para auxílio de códigos de aprendizado de máquina

Para este projeto, o desenvolvimento foi dividido em algumas etapas, onde cada trecho de código foi separado por responsabilidade. Essas etapas são:

* Pré-processamento das bases de dados,
* Análise de correlação entre os dados,
* Análise estatística dos dados,
* Treinamento de regressão linear,
* Análise do modelo treinado e testes com previsões.

**BASES DE DADOS E PRÉ-PROCESSAMENTOS**

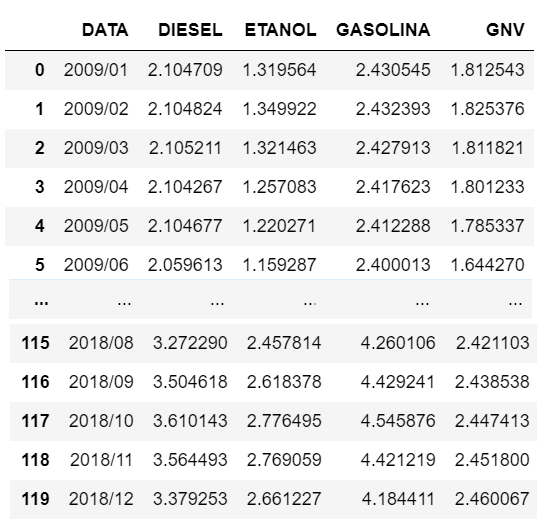
Foram utilizadas duas bases de dados diferentes para compor todas as informações necessárias para o estudo, ambos contendo informações a nível de série histórica, abrangendo o período de 2009 até 2018.

A primeira base de dados foi obtida do site da ANP (Agência Nacional de Petróleo, Gás Natural e Biocombustíveis), onde é possível acessar os dados históricos de compra e venda dos combustíveis produzidos pelo Brasil. Essas informaçoes sao de nivel de distribuidor final, ou seja, cada linha do arquivo CSV representa um alto posto de combustíveis do Brasil. As principais informações que foram utilizadas desta base de dados foram:

* Estado
* Cidade
* Produto
* Data da compra (Ano/Mês/Dia)
* Preço de venda p/ litro (R$)

O pré-processamento destes dados tiveram o propósito de primeiramente filtrar os dados somente dos estabelecimentos do Estado de São Paulo. Em seguida os dados foram separados por produto, assim o dividindo em quatro partes referente aos combustíveis: Diesel, Gasolina, GNV (Gás Natural Veicular) e Etanol. Para cada parte foi realizado um agrupamento a fim de se obter uma média mensal do preço de venda. É possível visualizar na Figura 1, uma demonstração do resultado pós processamento das informações.

Figura 1 - Dataset processado dos combustíveis.



Fonte: Plot de DataFrame da biblioteca pandas

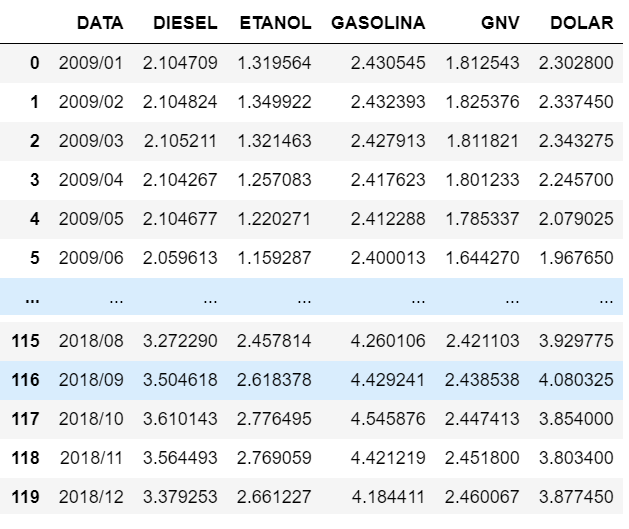
A segunda base de dados é referente às cotações do dólar, e foram obtidas no site do Yahoo Finanças, onde também é possível consultar e fazer download de várias informações sobre economia, cotações e bolsa de valores. O arquivo CSV das cotações do dólar já foram baixadas a nível mensal, e tais informações são compostas pelos seguintes dados:

* Data (Ano/Mês)
* Preço de abertura (R$)
* Preço de fechamento (R$)
* Preço mínimo (R$)
* Preço máximos (R$)

Para o pré-processamento, as informações de preço foram reduzidas para somente um valor por mês, e este valor foi obtido a partir da média das quatro informações citadas anteriormente (abertura, fechamento, mínimo, máximo).

A partir deste ponto já foi possível unir as duas bases de dados, pois ambas foram separadas a nível mensal. Essa composição final possui um total de 119 registros, referentes a médias mensais de preços em Real(R$) dos combustíveis e das cotações do dólar. Este dataset final é demonstrado na Figura 2.

Figura 2 - Dataset final dos dados utilizados no estudo.

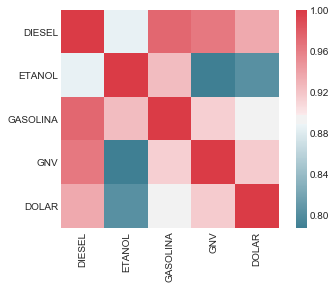


Fonte: Plot de DataFrame da biblioteca pandas

**CORRELAÇÃO ENTRE OS DADOS**

Para análise de correlação dos dados devemos levar em conta a base teórica dos dados para se obter uma conclusão, pois correlação não é causalidade. Sabe-se que em 2017 a Petrobras começou a acompanhar a cotação do dólar em seus preços, porém na base de dados que está sendo abordada aqui tem-se apenas dois anos a partir deste acontecimento (2017 e 2018), ou seja, os outros oito anos anteriores (2009 a 2016) podem ou não ser fato de causalidade. Na Figura 3 é demonstrada a matriz de correlação dos dados.

Figura 3 - Matriz de correlação dos dados.

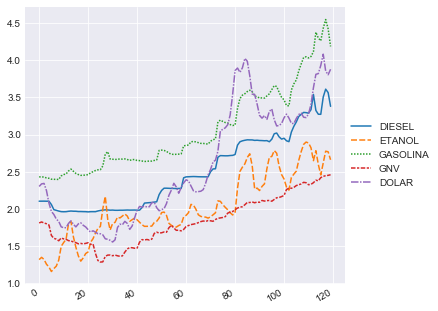


Fonte: Plot da biblioteca seaborn

Pode-se observar um alto índice de correlação dos dados do período de dez anos. O menor índice de correlação obtida não é menor que 80% e é referente ao combustível Etanol com o Dólar e o GNV.

Para comparativo de visualização histórica, é possível também analisar na Figura 4, considerando o eixo X como o período de 2009 até 2018, que todos os dados possuem uma semelhança na progressão temporal.

Figura 4 - Projeção histórica dos dados.

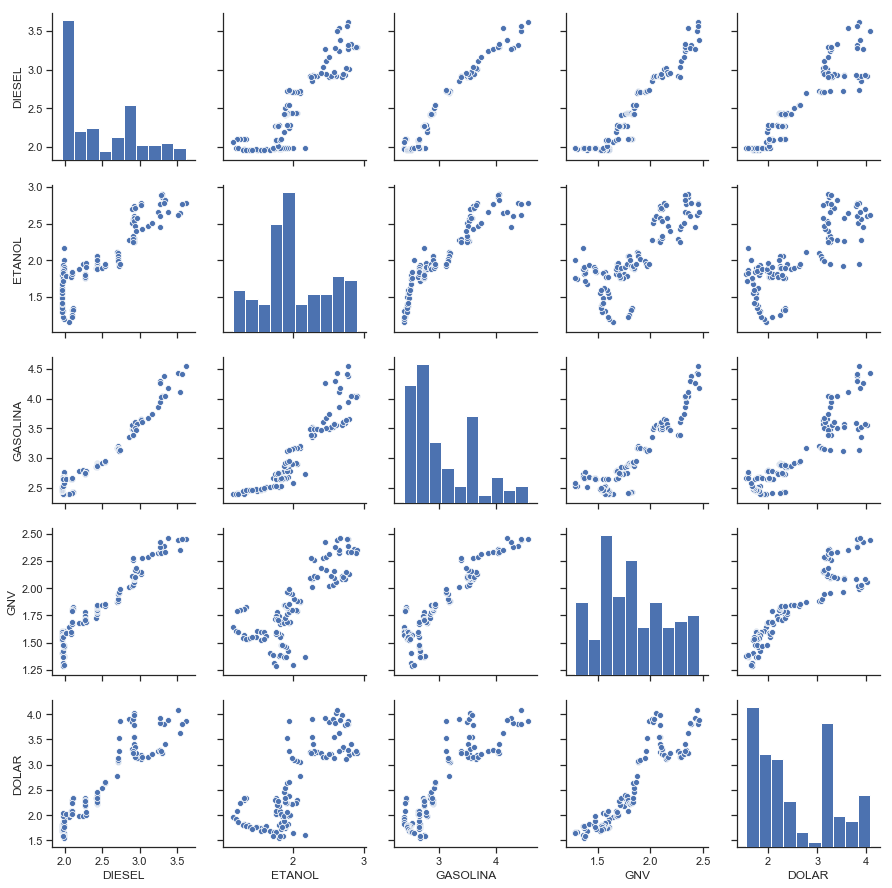


Fonte: Plot da biblioteca seaborn

Também é possível observar na Figura 4 que o modelo em questão possui certa linearidade em suas variáveis, ou seja, os dados parecem estar todos crescendo na mesma direção.

Na Figura 5 é possível observar que todas as correlações possuem uma visível linearidade das informações, garantindo assim certa coerência no ajuste da reta da regressão linear.

Figura 5 - Plots de dispersão em matriz



Fonte: Plot da biblioteca seaborn

Estes bons resultados significam que é possível sim analisar estatisticamente estes dados, pois já tem-se uma breve noção de que os mesmos não são completamente aleatórios e sem relação nenhuma.

**ANÁLISE ESTATÍSTICA DOS DADOS**

Para melhor entendimento da dimensão dos dados em questão, é necessário analisar algumas informações básicas como valores mínimos e máximos, média, desvio padrão e variância. Essas informações são demonstradas na Figura 6.

Figura 6 - Índices estatísticos dos dados



Fonte: Plot de DataFrame da biblioteca pandas

O Método dos Mínimos Quadrados é uma técnica matemática para minimizar as somas dos quadrados das diferenças entre os valores estimados e os valores observados. Essas diferenças são chamadas de resíduos, e para que a regressão linear tenha um bom ajuste da reta, é necessário que estes resíduos sejam distribuídos aleatoriamente e que essa distribuição seja normal.

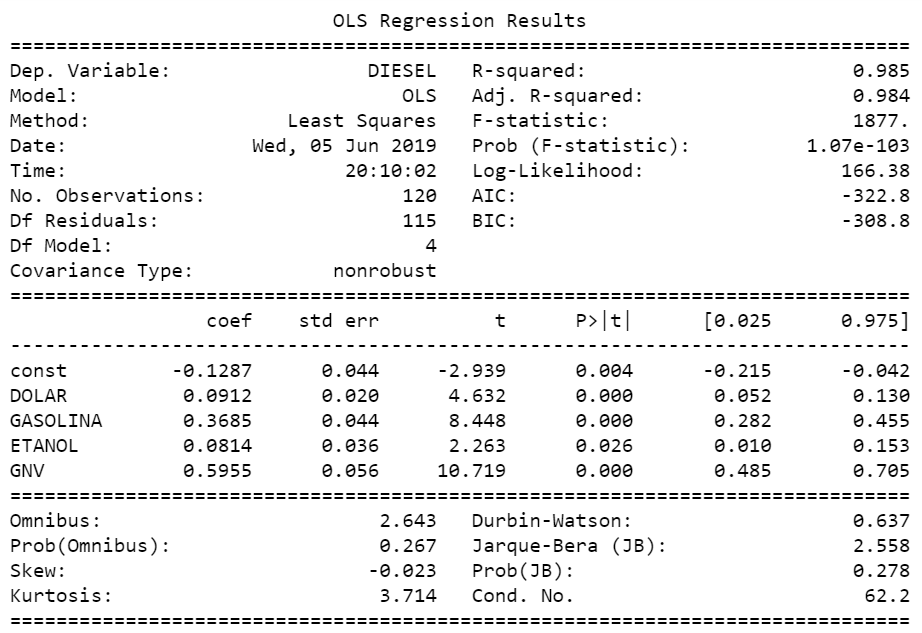
Utilizando a biblioteca statsmodels, tem-se acesso à função de análise OLS (Ordinary least squares), que retorna um relatório detalhado com todas as informações matemáticas para análise antes de um treinamento de regressão linear.

Antes é preciso definir quais variáveis compõem o modelo a ser analisado. Como o objetivo deste estudo é realizar uma previsão do preço do Diesel, esta será a nossa variável dependente. As informações da Gasolina, Etanol, GNV e Dólar são as nossas variáveis independentes, ou seja, baseando-se nos valores das mesmas, tem-se como objetivo prever a variável dependente Diesel. Deste modo, temos assim o modelo de uma Regressão Linear Múltipla, onde:

* Y = Diesel
* X = [Gasolina, Etanol, GNV, Dólar]

Com base nessas informações é possível testar nosso modelo e analisar os resultados do relatório OLS na Figura 7.

Figura 7 - Relatório OLS com informações matemáticas.



Fonte: Relatório da biblioteca statsmodels

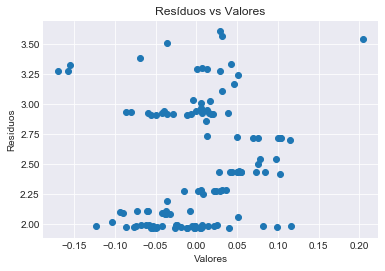
A primeira informação a se verificar neste relatório é o valor de Prob (F-statistic). Este valor tem que ser bem próximo de zero, para assim poder-se rejeitar a hipótese nula onde indica que o modelo não faz sentido algum. No nosso caso este valor deu bem próximo de zero, então não houve problemas quanto a este fator.

A próxima importante informação é Adj. R-squared que representa o quanto o nosso modelo explica dos dados observados. No modelo em questão obteve-se um valor de 0.984, ou seja, o mesmo explica 98.4% dos dados analisados.

Na parte do relatório que mostra as informações das variáveis, encontra-se a coluna P > | t |, que representa o p-valor de cada variável em relação à explicação dos dados do modelo. E para um fator de aceitação de 5%, conclui-se que todas as variáveis são coerentes com o modelo e fazem algum sentido na explicação dos dados da variável dependente.

Na Figura 8 pode-se analisar o gráfico que mostra os resíduos do modelo testado com os valores observados. E visualmente podemos concluir que os mesmos estão dispersos e não apresentam nenhum tipo de padrão aparente.

Figura 8 - Gráfico de Resíduos vs Valores observados

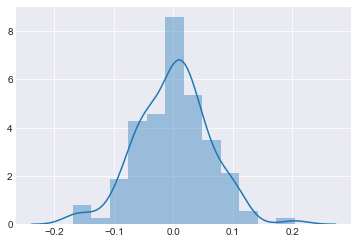


Fonte: Plot da biblioteca matplotlib

Para teste de normalidade dos resíduos, utilizou-se a função de shapiro da biblioteca statsmodels. Essa função retorna o p-valor referente a normalidade dos dados e deve-se comparar com um indicador de aceitação mínimo. Para o modelo em questão, obteve-se o valor de 0.3063, e conclui-se que para um fator de 5% de aceitação, estes resíduos são distribuídos normalmente.

Também é possível observar visualmente a distribuição normal dos resíduos, utilizando um plot simples de distribuição, conforme é demonstrado na Figura 9.

Figura 9 - Curva normal da distribuição dos resíduos.



Fonte: Plot da biblioteca seaborn

**TREINAMENTO DA REGRESSÃO LINEAR**

Com um alto valor de R-squared e sabendo que os resíduos do modelo estão distribuídos aleatoriamente e normalmente, pode-se assim testar um método de treinamento para a regressão linear.

A biblioteca utilizada para o treinamento foi a sklearn, que já possui todas as ferramentas necessárias para auxiliar com a configuração da regressão linear. Os dados foram divididos em duas partes, sendo elas para treinamento e teste. A proporção escolhida foi de 90% para treinamento e 10% para testes. Esses valores foram definidos com base na baixa quantidade de amostras. Para a fácil separação dos dados utilizou-se a função train\_test\_split da biblioteca citada.

O método utilizado para separação do dados foi o HoldOut, onde ele separa os dados de forma sequencial, ou seja, os primeiros 90% registros serão de teste e os 10% restantes como teste. Esse método foi escolhido por se tratar de uma série temporal onde cada registro representa um mês.

Os dados de do eixo X, que representam as variáveis independentes, foram todos transformados para uma distribuição normal. Com essa prática garantimos um melhor desempenho para o treinamento da regressão linear.

Utilizando a função StandardScaler.fit da biblioteca em questão, o cálculo das informações necessárias para se aplicar a normalização (média e desvio padrão), foram feitos com base nos 90% dos dados que são referentes ao treinamento. Em seguida com o método scaler.transform é possível transformar efetivamente a parte de treinamento e teste para a distribuição normal.

Com os dados separados e normalizados, podemos assim utilizar o módulo de LinearRegression, também contido na biblioteca citada. Esse módulo possui várias funções e informações que podemos utilizar para análise e predição.

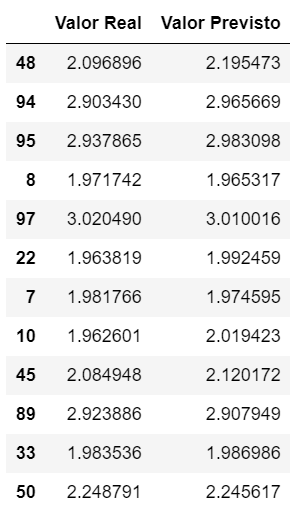
**RESULTADOS E DISCUSSÃO**

Os resultados do treinamento foram satisfatórios, e temos como resultado do treinamento realizado anteriormente da regressão linear múltipla, as seguintes informações:

* Erro Médio Absoluto: **0.0311**
* Erro Quadrático Médio: **0.0018**
* Raiz Quadrada do Erro Quadrático Médio: **0.0422**
* Percentual médio do Erro: **0.02%**

E na Figura 10, também podemos analisar um comparativo entre valor de teste e valor previsto.

Figura 10 - Comparação entre resultados reais e previstos.



Fonte: Plot de DataFrame da biblioteca pandas

Os valores previstos chegaram muito próximos dos valores de treino, ou seja, o modelo realmente conseguir prever valores dos quais não tinha como conhecimento.

**CONCLUSÃO**

É claro que não podemos considerar estes resultados como causalidade, pois como foi visto anteriormente, a quantidade de amostras foi reduzida drasticamente a nível mensal, e os dados são referente somente ao Estado de São Paulo.

Este experimento teve como objetivo tentar provocar um pensamento de que com a tecnologia, muitas vezes podemos prever crises econômicas que afetam inúmeras aéreas econômicas do país.

Em um modelo muito mais complexo deste assunto, seria possível tomar decisões antecipadas para prevenir que este mal acontecimento econômica aconteça. Os resultados foram promissores, portanto não descarta a possibilidade de um estudo muito mais aprofundado no assunto.

**REFERÊNCIAS**

Site da Petrobras - Produtos

<http://www.petrobras.com.br/pt/produtos-e-servicos/produtos/>

Site da ANP (Agencia Nacional de Petróleo, Gás Natural e Biocombustíveis)

<http://www.anp.gov.br/dados-abertos-anp>

Site do Yashoo Finanças

<https://br.financas.yahoo.com/>

Noticia Site UOL - Economia

<https://economia.uol.com.br/noticias/reuters/2017/06/30/petrobras-podera-reajustar-precos-de-gasolina-e-diesel-ate-diariamente.htm>

Noticia Site UOL - Economia

<https://economia.uol.com.br/noticias/bbc/2018/05/23/6-perguntas-para-entender-a-alta-nos-precos-da-gasolina-e-do-diesel.htm>

Notícia Site Em.com.br - Economia

<https://www.em.com.br/app/noticia/politica/2018/06/03/interna_politica,964040/dependencia-rodovias-crise-dos-combustiveis-colocam-pais-sob-ameaca.shtml>