Machine Learning Capstone Project - v4.0
Nanodegree Engenheiro de Machine Learning

Marcus Vinicius de Oliveira Cruz
11 de Março de 2018

Time Series Forescasting - Shanghai license plate buildding price prediction

1. Definições do projeto

Introdução

Como conclusão do nanodegree Engenheiro de Machine Learning e por uma possibilidade de uma consultoria de machine learning resolvi ir a fundo em um modelo de time series forecasting. A definição de uma série temporal consiste basicamente em um modelo com base estatística que analisa uma variação de uma serie temporal e consegue realizar previsões.

Escolhi o dataset Shanghai license plate bidding price prediction para construir o meu modelo do capstone. Esse é um dataset que pertence ao Kaggle. Como base de pesquisa para o referencial teórico peguei os dados disponíveis nos links do Kaggle abaixo, assim como também nos Kernels disponíveis na plataforma.

Referencial Teórico

Os textos abaixo foram retirados do Kaggle, serão utilizados como base para explicar as minúcias desse dataset. Todos os textos abaixo dentro de Project Overview foram conseguidos através dessas pesquisas e foram traduzidos, em muitos momentos essa tradução é literal direto do Kaggle e eu apenas fiz adequações nos textos. Utilizei essa metodologia pois eu não tinha um conhecimento sobre o dataset e precisei pesquisar.

Os itens abaixo: definições principais, colunas e contexto são textos que consegui com essa pesquisa e não são de minha autoria.

Links principais com as informações e local para baixar os datasets:

https://www.kaggle.com/bogof666/shanghai-car-license-plate-auction-price
https://www.kaggle.com/bazingasu/shanghai-license-plate-bidding-price-prediction

Definições principais

O aumento da propriedade e uso de automóveis na China nas últimas duas décadas, aumentou o consumo de energia, piorou a poluição do ar e gerou um congestionamento exacerbado. O governo de Xangai adotou um sistema de leilão para limitar o número de placas emitidas para cada mês.

O conjunto de dados contém dados históricos de leilões de janeiro de 2002 a outubro de 2017. Como funciona o sistema de leilão: um preço inicial é dado no início do leilão, os licitantes só podem oferecer até 3 vezes por cada leilão e só podem marcar para cima ou para baixo dentro de 300 CNY (aproximadamente 46 USD) por cada lance. No final de cada leilão, apenas o número superior (número de placas que serão emitidas para o mês) receberá as placas de licença ao custo de suas propostas. A oferta n. ° será o preço mais baixo do mês. Por favor, note que os leilões são realizados on-line e cada licitante não poderá ver outros lances.

Colunas

Data: janeiro de 2002 a outubro de 2017 (observe que faltam em fevereiro de 2008)

- num_bidder *: número de cidadãos que participam do leilão para o mês
- num_plates *: número de placas que serão emitidas pelo governo para o mês
- low_deal_price *: preço mínimo do negócio, explicado acima, em CNY
- avg_deal_price *: preço médio do negócio, no CNY (observe que, como cada lance só pode ser marcado para cima ou para baixo no prazo de 300, não está se afastando muito do preço mais baixo)

O objetivo é prever o preço low *deal* para cada mês, o resultado real será atualizado no final de cada mês, o conjunto de dados é raspado de http://www.51chepai.com.cn/paizhaojiage/

Contato: ran_su147@hotmail.com

Contexto

Xangai usa um sistema de leilões para vender um número limitado de matadouros para compradores de automóveis com combustível fóssil todos os meses. O preço médio desta placa de licença é de cerca de US $ 13.000 e muitas vezes é referido como "a peça de metal mais cara do mundo". Então, nosso objetivo é prever o preço médio ou o preço mais baixo para o próximo mês.

2. Declaração do Problema

O maior problema em questão é conseguir quantificar uma série temporal e saber qual a melhor forma de tratar tais dados. Para esse projeto em específico o problema e implementar um modelo de forecasting em um dataset com registros de preço médio e preço mínimo das placas de licença para dirigir em Shangai. Teoricamente um modelo de forecasting consiste na visualização de dados acerca da variação da série temporal. Em um determinado momento, após o algoritmo ter entendido sobre a variação de preços e com as implementações de condições já pré estabelecidas e conhecidas acontece uma previsão, que basicamente consistem em regressões do modelo.

Outro grande problema ao trabalhar com séries temporais está no momento de fazer a limpeza dos dados, eu fiz um processo completo de pré processamento dos dados com base nas necessidades dos dados, e eles podem ser claramente observados tanto na parte de análise como no tratamento dos dados. O projeto está com diversas visualizações de dados. Os dados se mostraram “limpos” e pouco problemáticos, a falta de variáveis categóricas é um problema claro no modelo.

Implementar um modelo de forecasting é basicamente um problema de aprendizagem supervisionada, visto que o algoritmo vai vasculhar a base de dados em busca de padrões de irregularidades afim de implementar previsões assertivas sobre o comportamento futuro dos dados. Entre os passos mais importantes ao se trabalhar com séries temporais temos que ter uma compreensão sobre o mecanismo que gerou essa série para após isso isso conseguir gerar predições com base no comportamento da série e assim prevendo comportamentos futuros.

Após o momento que entendemos o mecanismo que gerou a série temporal fica mais fácil para conseguir realizar análises assertivas que descrevem melhor seus comportamentos. Com a análise exploratória dos dados e ajuda das visualizações podemos tirar conclusões melhores sobre os períodos presentes daquela série temporal, e assim é possível entender mais a fundo sobre as variações da série e o que representam e quando construir um bom modelo entender melhor sobre o as trajetórias futuras das séries temporais.

3. Métricas

Pelo fato de o forecasting usar muito a visualização dos dados poderá ser visto funcionando junto a essas visualizações. Dessa forma deverão sempre ser feitos testes baseados no comportamento real dos dados. Como métrica de avaliação principal, usei a comparação entre os algoritmos implementados. A partir da análise de diferentes modelos de machine learning vou poder entender melhor sobre os dados em questão para poder construir modelos mais robustos.

A principal métrica do projeto é comprovar que o treinamento do modelo como um problema de time series será o método mais efetivo e rápido para ser desenvolvido. Em estudos particulares fiz um levantamento e comprovei que uma melhor maneira para desenvolver esse modelo de previsões seria com base em uma rede neural artificial.

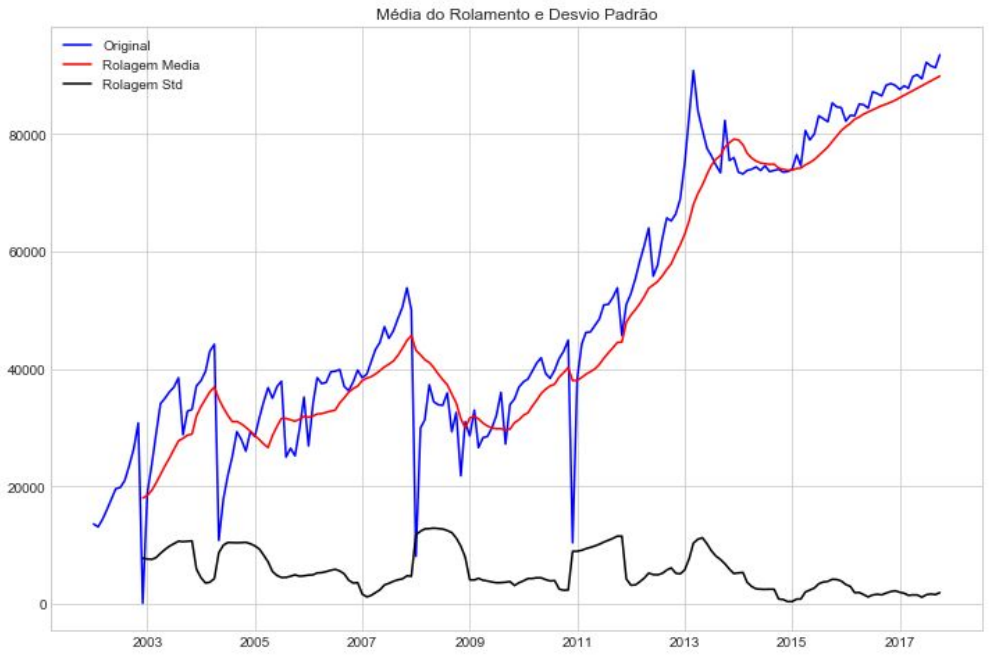
Durante o processo da implementação da Regressão Linear implementei algumas métricas específicas para modelos de regressão linear. Aqui consultei um artigo do Mário Filho que fala sobre as métricas mais populares para avaliar modelos de machine learning" [1]. Então apliquei as seguintes métricas ao modelo que geraram os seguintes resultados:

MAE (Erro Absoluto Médio) = 1350.45985006

MSE (Erro Quadrado Médio) = 3231963.93481

RMSE (Desvio Médio Quadrático) = 1797.76637381

Diretamente na parte de time series utilizei o teste de Dyckey Fuller para gerar estatísticas significativas sobre o modelo. Modelos ver abaixo os resultados do teste a partir da primeira visualização da aplicação do modelo de time series forecasting.



Resultados do teste Dickey-Fuller:
Test Estatistico -0.819643
p-value 0.813284
#Lags Usados 3.000000
Número de Observações Usadas 186.000000
Valores Criticos (1%) -3.466005
Valores Criticos (5%) -2.877208
Valores Criticos (10%) -2.575122
dtype: float64

Análises

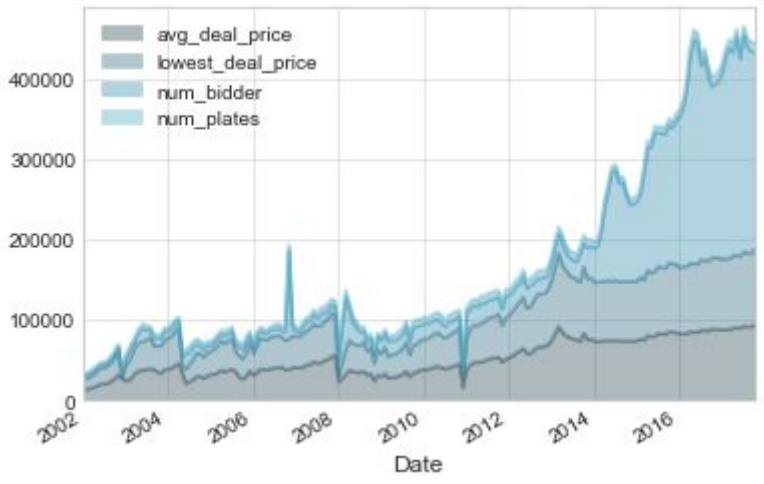
4. Exploração dos dados

Durante a análise exploratória dos dados várias questões foram sendo respondidas e consegui entender melhor sobre o dataset e sobre a série temporal com o valor mínimo do modelo. Pelas bibliotecas de time series forecasting usarem muito de visualizações para mostrar resultados ficam claras as características dos conjuntos de dados, as visualizações em si compõem os resultados dos cálculos e das previsões realizadas no projeto. Na parte da análise exploratória dos dados procurei criar diversas visualizações onde comparei as diversas colunas com os resultados do modelo.

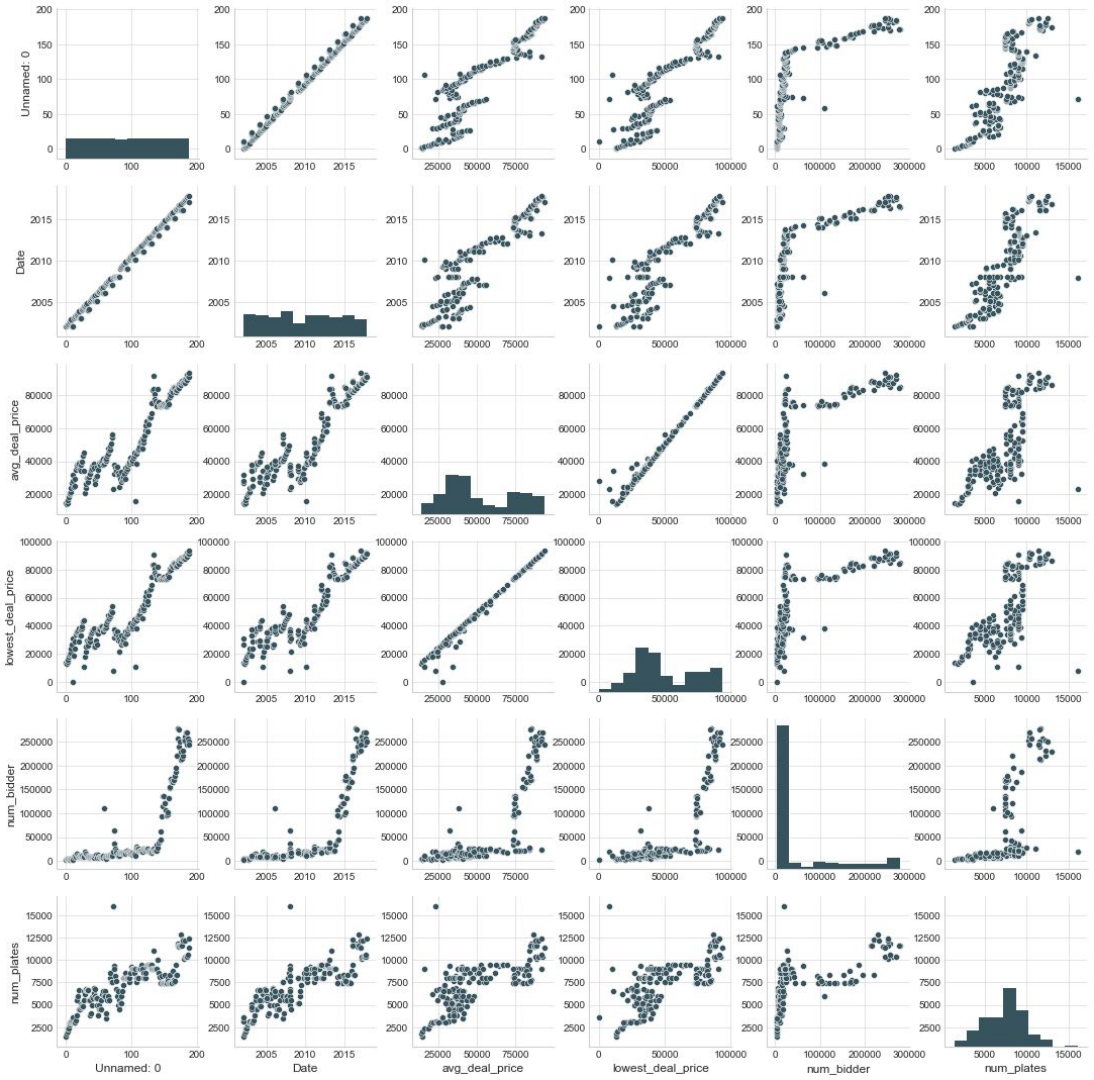
5. Exploração das visualizações

Em todo o Jupyter Notebook você vai poder observar visualizações claras sobre os dados e sobre os resultados na parte do time series forecasting.

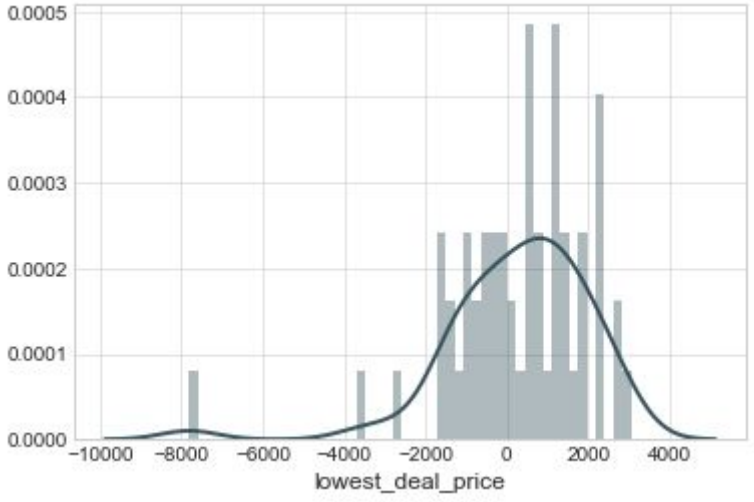
Seguem abaixo algumas visualizações significativas para o modelo:



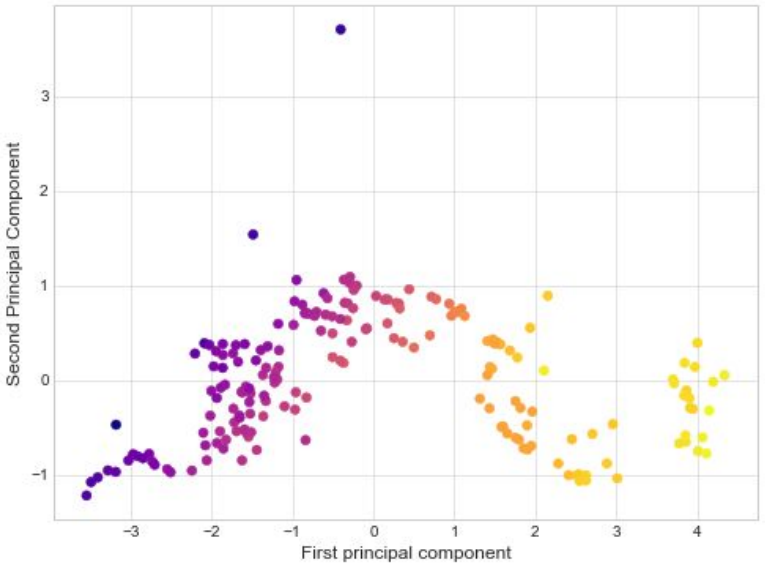
Visualização de todas as séries temporais do dataset em um mesmo plot



Matriz de visualização dos dados de todas as colunas do dataset



Visualização do resultado dos dados fazendo a diferença entre os dados de teste e os dados previstos



Visualização do resultado da aplicação do modelo PCA

6. Técnicas e Algoritmos

Uma série temporal basicamente consiste na coleta de dados em intervalos de tempos constante. De forma que sejam criadas visualizações de dados para apontar hipóteses mais conclusivas durante a análise de dados. Ao se trabalhar com previsões em séries temporais, basicamente encontramos um problema de aprendizagem supervisionada,dessa forma é necessário fornecer exemplos claros e concisos para treinamento do modelo em questão.

Abaixo citei uma lista com todas as aplicações presentes no Jupyter Notebook:

- Limpeza dos Dados;
- Análise Exploratória dos Dados;
- Criação dos modelos de aprendizado de máquina;
    - Regressão Linear;
    - Regressão Logística;
    - PCA - Análise do Componente Principal;
    - K-nearest Neighboors;
    - Árvores de Decisão e Florestas Aleatórias;
    - Time Series Forecasting;

7. Benchmark

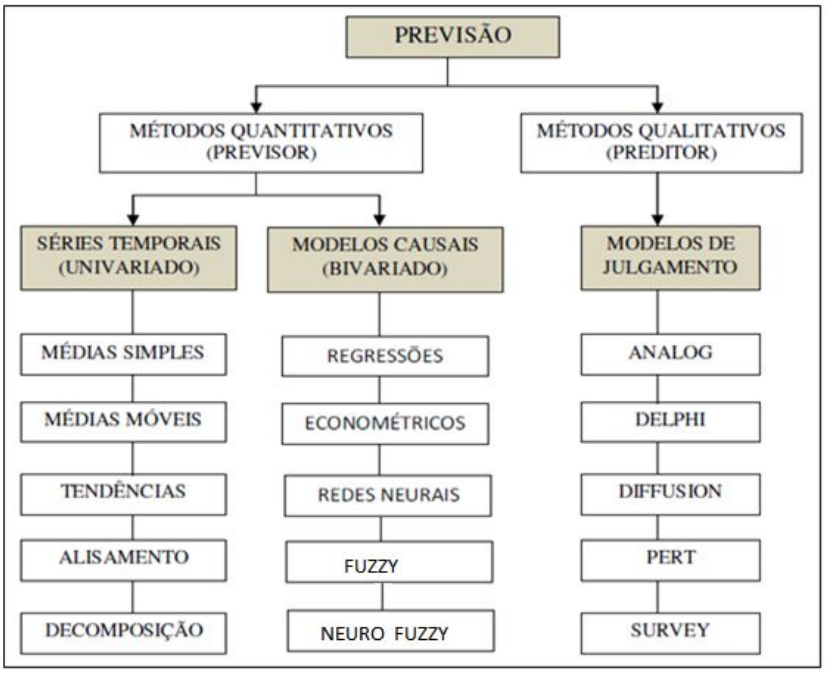
Os resultados apresentados no final do time series forecasting são claros e explicativos por si só. Por estar envolvido com um problema de time series no trabalho pude perceber que o R tem bibliotecas prontas e de maior facilidade de implementação, e também é mais fácil de trabalhar com as visualizações de dados. Enquanto o Python é mais robusto se pensando

em aplicações, o R ganha em facilidade de implementação, curva de aprendizagem rápida e por possuir pacotes focados por exemplo em “outliers”.

Com relação aos modelos de machine learning apliquei diversos algoritmos citados acima, fica claro que para esse modelo o time series forecasting é melhor para ser aplicado pois não depende de variáveis categóricas. Em um projeto de trabalho essa questão das time series ficou tão clara que estou fazendo feature engineering para conseguir ter variáveis categóricas para transformar um problema de time séries em um problema de aprendizagem supervisionada e aplicar um modelo de classificação

Na imagem abaixo, que consegui em minha pesquisa, a figura mostra diferentes formas para se trabalhar com algoritmos de previsão de demanda.

Figura 2 - Métodos de Previsão de Demanda



Fonte: Adaptado de SILVA (2003)

Metodologia

8. Pré-processamento dos Dados

Na etapa de pré processamento dos dados apliquei um padrão na análise de dados que consiste em:

1. Reunir os dados.
2. Avaliar os dados.
3. Limpar os dados.

Processo aplicado através do time series forecasting, após considerar já ter passado pelos passos anteriores:

1. Analisar e criar visualizações sobre a série temporal;
2. Aplicar as funções que realizar o estacionamento das séries temporais.
3. Criar plotagens e criação de métricas para validar os modelos;
4. Construir e aplicar um modelo ARIMA;
5. Realizar predições e previsões sobre o modelo;

9. Implementação

O Jupyter Notebook por ser uma ferramenta didática consegue transmitir muitas informações e explicar o que está acontecendo nos modelos, está claro e conciso com relação ao que está acontecendo no modelo. Durante o processo de desenvolvimento encontrei alguns problemas com dificuldades técnicas com relação ao Python e limpeza dos dados, todos esses processos estão documentados no Jupyter Notebook. Houveram mudanças no escopo do projeto, mudanças de dataset, pesquisa e análise das variáveis para chegar no modelo final. Acredito que todo o projeto e metodologia para a análise de dados merece ser documentado por isso o Jupyter Notebok é um guia para se trabalhar com séries temporais.

10. Refinamento

A regressão linear foi o primeiro algoritmo pensado para se resolver o problema, após a aplicação não fiquei com os resultados, dessa forma apliquei os próximos algoritmos de machine learning que também não me satisfizeram. Por estar atuando com pesquisa diretamente em time series estou mais confiante para trabalhar com as séries de tempo. Todos os algoritmos implementados no projeto são soluções alternativas para a resolução do problema principal de prever o preço mínimo das licenças.

Resultados

11. Evolução do modelo de validação

O modelo está bastante de acordo com as minhas expectativas, pensando nesse modelo ainda como um mvp com o objetivo de testar o modelo os parâmetros finais do modelo são bastante apropriados. A aplicação de time série forecasting está robusta para o modelo e pequenas perturbações sempre serão um problema para todos que trabalham com séries temporais. Os modelos mesmo com baixa precisão são confiáveis e comprovam que baseado na comparação da aplicação dos outros modelos de machine learning.

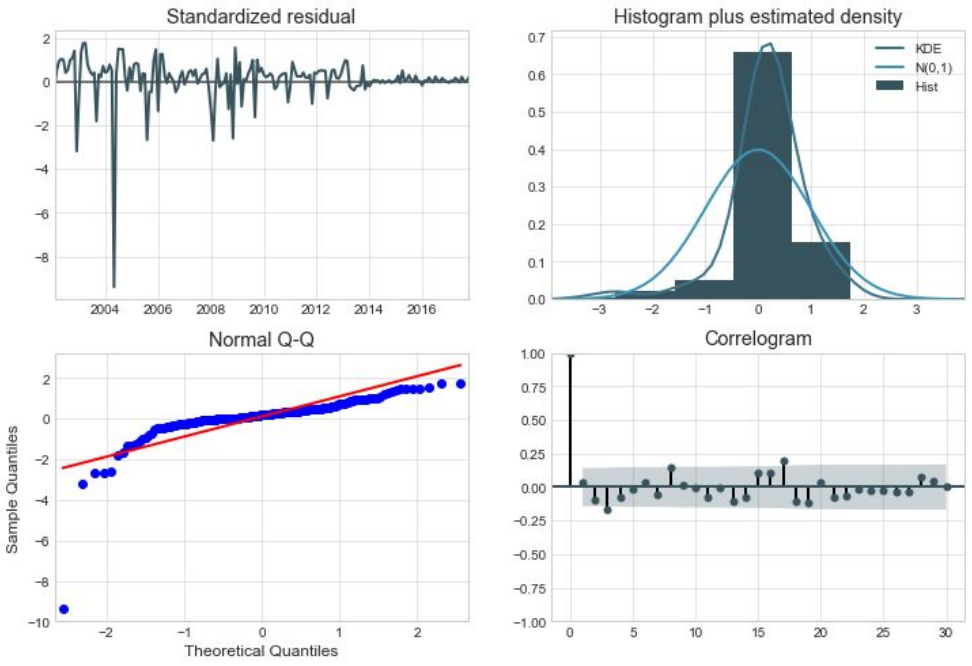
Conclusões

12. Melhorias

Uma melhoria clara para esse problema em especial serial trabalhar com uma rede neural artificial treinada para conseguir previsões mais refinadas e robustas do modelo, claramente se eu estivesse mais evoluído tecnicamente em implementar redes neurais artificiais eu teria feito esse projeto como uma comparação entre a resolução desse problema como time series ou com um rede neural artificial.

A solução final pode ser usada como referência porém ainda precisa ser melhor treinada com testes com os reais interessados na variação do preço das licenças para dirigir.

O plot abaixo pode ser utilizado para entender quais visualizações específicas são melhores representadas por esse modelo.



Visualização do plot_diagnostics para ver os melhores plots que representam o modelo

13. Conclusão / Justificativa

Em todo o projeto procurei deixar os textos o mais claros e explicativos possíveis. Revisei o projeto diversas vezes e acredito que os erros de português serão mínimos. Foi fornecida uma visualização que enfatiza uma qualidade importante sobre o projeto com uma discussão aprofundada. As pistas visuais estão claramente definidas. Dei a devida referência a todo material que consultei, consultei não só artigos de internet, mas também artigos acadêmicos, o projeto tem bastante embasamento. O referencial teórico está gigante pois procurei entender de fato os problemas e buscar a melhor ferramenta para resolver os problemas iniciais. Conforme orientação da Udacity o Jupyter Notebook está

completamente explicativo. Procurei documentar todo o processo de criação, assim como toda a validação das hipóteses do início ao final do projeto.

Esse modelo de time series forecasting está funcional e conseguiu atingir os resultados esperados quando ainda estava apenas especulando sobre os resultados do projeto. Como eu foquei em deixar um projeto interessante não somente para uma aplicação específica acredito que esse projeto está atendendo perfeitamente aos pré requisitos.

Séries temporais são extremamente interessantes, redes neurais artificiais e deep learning ainda mais. Com a realização do projeto acredito mil vezes mais no poder das visualizações de dados, fiz diversos cursos focados em R que me ajudaram a entender melhor os problemas e refletiram em mudanças nas implementações do Python.

Ao final do projeto foi possível criar previsões que em alguns pontos ficaram bastante assertivas com relação ao modelo de dados original. Nos últimos códigos de séries temporais foi realizada uma previsão do comportamento do preço mínimo para o leilão das placas de licitação

Referências

[1] - Site do Mário Filho. Acesso em: http://mariofilho.com/as-metricas-mais-populares-para-avaliar-modelos-de-machine-learning/

[2] https://www.kaggle.com/bazingasu/data-exploration

[3] https://github.com/seanabu/seanabu.github.io/blob/master/Seasonal_ARIMA_model_Portland_transit.ipynb

[4] https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/

[5] http://ucanalytics.com/blogs/wp-content/uploads/2017/08/ARIMA-TimeSeries-Analysis-of-Tractor-Sales.html

[6] https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/03.11-Working-with-Time-Series.ipynb

[7] https://www.kaggle.com/kentata/time-series-data-exploration

https://www.kaggle.com/bogof666/shanghai-car-license-plate-auction-price/data

https://www.kaggle.com/bazingasu/shanghai-license-plate-bidding-price-prediction/data

http://deeplearningbook.com.br/capitulo-9-a-arquitetura-das-redes-neurais/

https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/

https://codeburst.io/jupyter-notebook-tricks-for-data-science-that-enhance-your-efficiency-95f98d3adee4

https://towardsdatascience.com/why-you-should-forget-for-loop-for-data-science-code-and-embrace-vectorization-696632622d5f

https://towardsdatascience.com/10-machine-learning-algorithms-you-need-to-know-77fb0055fe0

http://www.cbcity.de/timeseries-decomposition-in-python-with-statsmodels-and-pandas

https://bibliotecadigital.ipb.pt/bitstream/10198/12709/1/Artur%20Jorge%20Ferreira%20da%20Costa%20Dias.pdf

http://www.redalyc.org/pdf/3291/329147536007.pdf

http://www.ceel.eletrica.ufu.br/artigos/ceel2016_artigo094_r01.pdf

https://repositorio.bc.ufg.br/tede/bitstream/tede/7563/5/Disserta%C3%A7%C3%A3o%20-%20Ricardo%20Henrique%20Fonseca%20Alves%20-%202017.pdf

http://ftp.cptec.inpe.br/labren/publ/teses/DISSERTACAO_RICARDO-GUARNIERI.pdf

http://www.confea.org.br/media/contecc2017/eletrica/1_audrnanpdrsg.pdf

http://www.inovarse.org/sites/default/files/T14_0291_5.pdf

http://www.redalyc.org/html/3291/329147536007/

https://repositorio.ufu.br/handle/123456789/14569

https://repositorio.ufsc.br/bitstream/handle/123456789/178026/TCC_Final_Jhuan_Souza.pdf?sequence=1&isAllowed=y

https://www.producaoonline.org.br/rpo/article/view/2542/1596

revistas.ufpr.br/rber/article/download/48431/pdf

https://pt.stackoverflow.com/questions/192098/como-funciona-uma-rede-neural-artificial

https://martin-thoma.com/classification-with-pybrain/

http://conteudo.icmc.usp.br/pessoas/andre/research/neural/

http://www.din.uem.br/ia/neurais/

http://www.cerebromente.org.br/n05/tecnologia/rna.htm

ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia006_03/topico5_03.pdf

http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0016231_04_cap_05.pdf

https://www.embarcados.com.br/redes-neurais-artificiais-introducao/

https://periodicos.utfpr.edu.br/recit/article/view/4330/Leandro

http://www2.ica.ele.puc-rio.br/Downloads/33/ICA-introdu%C3%A7%C3%A3o%20RNs.pdf

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1982-21702017000100150&lng=pt&tlng=pt

http://www.scielo.br/scielo.php?pid=S1678-86212017000300103&script=sci_arttext

https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/

https://www.monolitonimbus.com.br/processos-estacionarios/

https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/

http://www.portalaction.com.br/series-temporais/11-estacionariedade

http://www.icad.puc-rio.br/cfeijo/pdf/revis%C3%A3o%20b%C3%A1sica%20s%C3%A9ries%20temporais_material%20de%20apoio_curso%20teoria%20macroeconomica_PPGE%20UFF.pdf

https://www.ime.unicamp.br/~hlachos/MaterialSeries.pdf

http://www.inf.ufsc.br/~marcelo.menezes.reis/Cap4.pdf

https://www.researchgate.net/publication/229040330_JTIMESAT_uma_ferramenta_para_a_visualizacao_de_series_temporais_de_imagens_de_satelite

http://bdm.unb.br/bitstream/10483/7239/1/2013_JoseRobertoGoncalvesdeRezendeFilho.pdf

https://www.maxwell.vrac.puc-rio.br/16824/16824_4.PDF

https://www.maxwell.vrac.puc-rio.br/24787/24787_4.PDF

http://conteudo.icmc.usp.br/pessoas/ehlers/stemp/stemp.pdf

http://cdsid.org.br/sbpo2015/wp-content/uploads/2015/08/140250.pdf

https://www.marinha.mil.br/spolm/sites/www.marinha.mil.br.spolm/files/101711_0.pdf

https://www.lume.ufrgs.br/bitstream/handle/10183/31034/000782115.pdf?sequence=1

http://www2.ufersa.edu.br/portal/view/uploads/setores/232/TCC%20-%20VALCIANO%20CAMILO%20GURGEL.pdf

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1982-21702017000100150&lng=pt&tlng=pt

http://www.ime.eb.br/arquivos/teses/se4/mec2008/2008Bianca.pdf

http://repositorio.unicamp.br/bitstream/REPOSIP/267746/1/Conti_JoseCarlos_M.pdf

http://www.ctec.ufal.br/professor/cfs/Sul_Sud06%20-%20Series.pdf

http://pdf.blucher.com.br.s3-sa-east-1.amazonaws.com/marineengineeringproceedings/spolm2015/140011.pdf

http://www.portalaction.com.br/series-temporais/15-modelos-para-series-temporais

http://www.scielo.br/scielo.php?pid=S1678-86212017000300103&script=sci_arttext

https://www.researchgate.net/publication/289479535_Previsao_de_energia_eletrica_modelagem_e_uso_de_combinacoes_de_previsoes

https://www.ufrgs.br/sbai17/papers/paper_506.pdf

http://www.scielo.org.co/pdf/eia/n26/en_n26a09.pdf

http://www.sciencedirect.com/science/article/pii/S1877050915015641

http://www.uff.br/engevista/seer/index.php/engevista/article/viewFile/433/236

http://www.ufjf.br/pgmc/files/2011/05/Disserta%C3%A7%C3%A3o-Guilherme-G-Neto-18-08.pdf

http://www.exatas.ufpr.br/portal/degraf_paulo/wp-content/uploads/sites/4/2014/09/EE022-08-08.pdf

http://www.datascienceinstitute.com.br/forecast-de-consumo-de-energia-eletrica/

https://docs.microsoft.com/pt-br/azure/machine-learning/preview/scenario-time-series-forecasting

https://translate.google.com.br/translate?sl=en&tl=pt&js=y&prev=_t&hl=pt-BR&ie=UTF-8&u=http%3A%2F%2Fwww.scielo.br%2Fscielo.php%3Fscript%3Dsci_arttext%26pid%3DS1678-86212017000300103%26lng%3Dpt%26tlng%3Dpt&edit-text=

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1678-86212017000300103&lng=pt&tlng=pt

http://acervo.ufvjm.edu.br/jspui/bitstream/1/1327/1/rodrigo_magalhaes_mota_santos.pdf

http://tede2.pucgoias.edu.br:8080/bitstream/tede/2484/1/Paulo%20Henrique%20Borba%20Florencio.pdf

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.4455&rep=rep1&type=pdf

https://gab41.lab41.org/the-10-algorithms-machine-learning-engineers-need-to-know-f4bb63f5b2fa

http://minerandodados.com.br/index.php/2017/05/19/prevendo-precos-de-acoes-da-bolsa-de-valores-com-machine-learning/

https://fga.unb.br/articles/0000/5556/TCC_Hialo_Muniz.pdf

http://www.feis.unesp.br/Home/departamentos/engenhariaeletrica/pos-graduacao/327-dissertacao_ciceromarcelo.pdf

https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/

https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

http://www.semantix.com.br/10-algoritmos-de-machine-learning/

http://www.scielo.br/scielo.php?pid=S1678-86212017000300103&script=sci_arttext

https://fga.unb.br/articles/0000/7804/TCC_Hialo_Muniz.pdf

http://www.leec.eco.br/downloads/R-tutorial-de-bolso.pdf