Machine Learning Capstone Project - v5.0
Nanodegree Engenheiro de Machine Learning

Marcus Vinicius de Oliveira Cruz
21 de Março de 2018

Shanghai license plate buildding price prediction
Regressão Linear vs Time Series Forecasting

I - Definição

1. Definições do projeto

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 série 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/

Contexto

Xangai usa um sistema de leilões para vender um número limitado de placas de licença 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. Muitas vezes, os problemas das séries temporais são em tempo real, fornecendo continuamente novas oportunidades de previsão. Isso acrescenta uma honestidade às séries temporais que predizem rapidamente a descarga de maus pressupostos, erros na modelagem e todas as outras maneiras pelas quais podemos nos enganar. [17]

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. A partir da análise de dois modelos de machine learning vou poder entender melhor sobre os dados em questão para poder construir modelos mais robustos. 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, porém para esse trabalho vou seguir com a comparação entre regressão linear e time series forecasting.

Para criar avaliações do modelo de regressão linear utilizarei as métricas que vou descrever e explicar a seguir. Nas conclusões do projeto vou incluir os resultados. Utilizei essas métricas pois em minha pesquisa foram as melhores métricas para se trabalhar com modelos de regressão linear, como o modelo de time series forecasting é auto explicativo com as visualizações no final vou analisar os resultado e mostrar os resultados. Seguem as métricas utilizadas:

- **Scoring [9]**

Retorna o coeficiente de determinação R ^ 2 da predição. O coeficiente R ^ 2 é definido como (1 - u / v), onde u é a soma residual de quadrados ((y_true - y_pred) ** 2) .sum () e v é a soma total de quadrados ((y_true - y_true.mean ()) ** 2) .sum (). O melhor resultado possível é 1,0 e pode ser negativo (porque o modelo pode ser arbitrariamente pior). Um modelo constante que sempre prediz o valor esperado de y, desconsiderando os recursos de entrada, obteria um resultado R ^ 2 de 0,0.

- **Coefficient [11]**

Coeficientes estimados para o problema de regressão linear. Se vários destinos foram passados durante o ajuste (e 2D), esta é uma matriz 2D de forma (n_targets, n_features), enquanto que se apenas um alvo for passado, esta é uma matriz 1D de comprimento n_features.

- **Intercept [11]**

Termo independente no modelo linear.

- **Root Mean Square Error [12]**

Perda de regressão de erro quadrático médio. Retorna um valor de ponto flutuante não negativo (o melhor valor é 0,0), ou uma matriz de valores de ponto flutuante, um para cada alvo individual.

- **Mean Absolut Error [13]**

Perda de regressão de erro absoluto médio. Se multioutput for 'raw_values', então o erro absoluto médio é retornado para cada saída separadamente. Se multioutput for 'uniform_average' ou um ndarray de pesos, então a média ponderada de todos os erros de saída é retornada. A saída MAE é um ponto flutuante não negativo. O melhor valor é 0,0.

- **Mean Square Error [14]**

Perda de regressão de erro quadrático médio. Retorna um valor de ponto flutuante não negativo (o melhor valor é 0,0), ou uma matriz de valores de ponto flutuante, um para cada alvo individual.

- **R2 Score [10]**

R ^ 2 (coeficiente de determinação) função de pontuação de regressão. A melhor pontuação possível é 1,0 e pode ser negativa (porque o modelo pode ser arbitrariamente pior). Um modelo constante que sempre prediz o valor esperado de y, desconsiderando os recursos de entrada, obteria um resultado R ^ 2 de 0,0.

- **KFold Cross Validator [15]**

Fornece índices de treino / teste para dividir dados em conjuntos de trem / teste. Divida o conjunto de dados em dobras consecutivas (sem baralhar por padrão). Cada dobra é então usada uma vez como uma validação enquanto as k-1 dobras restantes formam o conjunto de treinamento.

II - 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 que vou disponibilizar abaixo, algumas são auto explicativos, caso não forem vou incluir uma descrição.

Tipos dos dados das colunas:

```
Data                 float64
Preço Médio           int64
Preço Mínimo          int64
Numero Cidadões       int64
Número Placas         int64
dtype: object
```

Resumo conciso do data frame juntamente com contagem de dados não nulos:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 189 entries, 0 to 188
Data columns (total 5 columns):
Data                 189 non-null float64
Preço Medio          189 non-null int64
```

Preço Minimo 189 non-null int64
Numero Cidadões 189 non-null int64
Numero Placas 189 non-null int64
dtypes: float64(1), int64(4)

Impressão de todo o data frame:

	Data	Preço Médio	Preço Mínimo	Numero Cidadãos	Numero Placas
0	2002.10	14735	13600	3718	1400
1	2002.20	14057	13100	4590	1800
2	2002.30	14662	14300	5190	2000
3	2002.40	16334	16000	4806	2300
4	2002.50	18357	17800	4665	2350
5	2002.60	20178	19600	4502	2800
6	2002.70	20904	19800	3774	3000
7	2002.80	21601	21000	4640	3000
8	2002.90	24040	23600	4393	3200
9	2002.10	27040	26400	4661	3200
10	2002.11	31721	30800	4021	3200
11	2002.12	27848	100	3525	3600
12	2003.10	24267	18800	9442	3000
13	2003.20	25254	23800	12030	3000
14	2003.30	29551	28800	11219	3000
15	2003.40	34845	34100	8794	3300
16	2003.50	36903	35000	14634	3800
17	2003.60	37667	36100	15507	5500
18	2003.70	38269	36900	11929	6000
19	2003.80	39369	38500	9315	4500
20	2003.90	38728	28800	8532	6650
21	2003.10	34842	32800	9383	4500
22	2003.11	34284	33100	9849	5042
23	2003.12	38054	37100	10491	4776
24	2004.10	39516	38000	8663	5000
25	2004.20	40053	39600	10156	4800
26	2004.30	43333	43000	9950	4800
27	2004.40	45492	44200	8150	5500
28	2004.50	34226	10800	8114	6527
29	2004.60	21001	17800	19233	6233
..	...	...	...	...	...
159	2015.50	79099	79000	156007	7482
160	2015.60	80020	80000	172205	7441
161	2015.70	83171	83100	166302	7531
162	2015.80	82642	82600	166939	7454
163	2015.90	82172	82100	165765	8727
164	2015.10	85424	85300	170995	7763
165	2015.11	84703	84600	169159	7514

166	2015.12	84572	84500	179133	7698
167	2016.10	82352	82200	187533	9409
168	2016.20	83244	83200	196470	8363
169	2016.30	83148	83100	221109	8310
170	2016.40	85127	85100	256897	11829
171	2016.50	85058	85000	277889	11598
172	2016.60	84483	84400	275438	11546
173	2016.70	87235	87200	240750	11475
174	2016.80	86946	86900	251188	11549
175	2016.90	86523	86500	229544	12889
176	2016.10	88359	88300	213212	11621
177	2016.11	88665	88600	215424	11549
178	2016.12	88412	88300	219882	12261
179	2017.10	87685	87600	232101	12215
180	2017.20	88240	88200	251717	10157
181	2017.30	87916	87800	262010	10356
182	2017.40	89850	89800	252273	12196
183	2017.50	90209	90100	270197	10316
184	2017.60	89532	89400	244349	10312
185	2017.70	92250	92200	269189	10325
186	2017.80	91629	91600	256083	10558
187	2017.90	91415	91300	250566	12413
188	2017.10	93540	93500	244868	11388

[189 rows x 5 columns]

Resumo estatístico do data frame:

In [21]:
```
#Resumo estatístico do DataFrame, com quartis, mediana, etc.
data_test.describe()
```

Out[21]:

	Data	Preço Medio	Preço Minimo	Numero Cidadões	Numero Placas
count	189.000000	189.000000	189.000000	189.000000	189.000000
mean	2009.814233	51632.867725	50431.216931	56142.423280	7336.111111
std	4.612254	22548.372228	23574.450403	78988.734651	2468.294304
min	2002.100000	14057.000000	100.000000	3525.000000	1400.000000
25%	2005.900000	34684.000000	33100.000000	10170.000000	5690.000000
50%	2009.900000	42262.000000	41900.000000	18575.000000	7500.000000
75%	2013.800000	74113.000000	74000.000000	41946.000000	9000.000000
max	2017.900000	93540.000000	93500.000000	277889.000000	16000.000000

Checando quais valores faltam no dataset:

2002 : (12, 4)
2003 : (12, 4)
2004 : (12, 4)
2005 : (12, 4)
2006 : (12, 4)
2007 : (12, 4)
2008 : (11, 4)
2009 : (12, 4)
2010 : (12, 4)
2011 : (12, 4)
2012 : (12, 4)
2013 : (12, 4)
2014 : (12, 4)
2015 : (12, 4)
2016 : (12, 4)
2017 : (10, 4)

Vendo os dados acima foi possível perceber que faltam dados em 2008 e 2017. As informações do dataset informam que em 2008 falta um mês de dados, e em 2017 os dados vão até outubro. Será necessário completar os dados faltantes para 2008, acredito que completando os dados de 2017 me ajudará bastante. Para resolver esse problema vou criar uma função que vai trazer uma média entre os dados da série temporal e completar os dados faltantes.

5. Exploração das visualizações

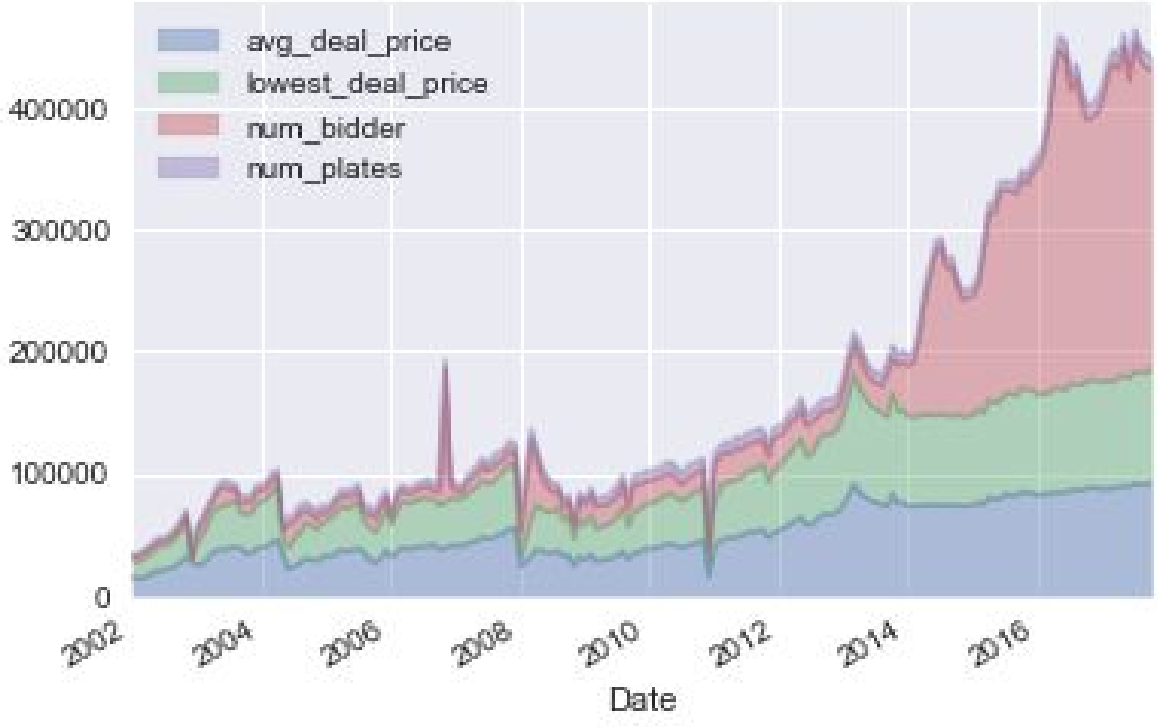
Em todo o Jupyter Notebook você vai poder observar visualizações claras sobre os dados e sobre os resultados na parte do time series forecasting. Os gráficos mostra, algumas questões que devem ser analisados nos dados:

Gŕafico 1:

1. A longo prazo fica claro aumento no volume de placas e cidadãos, até por volta de 2014 às 4 séries temporais apresentam um crescimento seguindo um padrão, após 2014 o número de cidadão e número de placas aumenta muito e tem um grande aumento mantendo um padrão, a curva de preço mínimo e preço médio também seguem o mesmo padrão.

2. Entre 2006 e 2008 aconteceu alguma anomalia, por volta de 2007 tem um pico de número de cidadão e número de placas, isso deve ser devido há alguma variação sazonal do mercado de Shangai. Em 2008 tem uma baixa abrupta, aqui consigo explicar pois no Kaggle é informado que faltam dados em 2008, comprovamos um pouco acima na análise exploratória dos dados.

3. Em 2012 acontece uma queda drástica referente a 2008, aqui não conseguimos explicar com a falta de dados, e deve ser alguma anomalia sazonal referente ao mercado de Shangai.

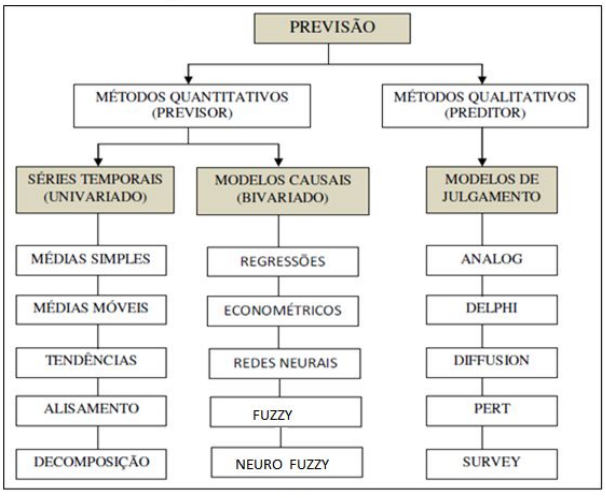


Visualização de todas as séries temporais do dataset em um mesmo plot

6. Técnicas e Algoritmos

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)

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.

- **Regressão Linear [16];**

É um algoritmo estatístico e paramétrico no qual, resumidamente, consiste em expressar a saída desejada na forma de uma função linear, onde cada instância é relacionada com um peso. Sua principal vantagem é a sua simplicidade, sendo reconhecidamente o algoritmo mais simples e utilizado para a criação de modelos de regressão. Já suas desvantagens consistem em trabalhar apenas com dados numéricos, e sua eficiência em problemas não-lineares é baixíssima, devido a sua limitação de tentar transformar uma função não-linear em um modelo linear simples. Ou seja, sua utilização é muito limitada. Para maiores detalhes, ver (WITTEN; FRANK, 2000).

A regressão linear é um tipo básico e comum de análise preditiva. A idéia geral de regressão é examinar duas coisas: (1) um conjunto de variáveis preditoras faz um bom trabalho na previsão de uma variável de resultado (dependente)? (2) Quais as variáveis em particular são preditores significativos do resultado variável e de que forma indicam a variável variável? Essas estimativas de regressão são usadas para explicar a relação entre uma variável dependente e uma ou mais variáveis independentes. A forma mais simples da equação de regressão com uma dependente e uma variável independente é definida pela fórmula y = c + b * x, onde y = escore variável dependente estimado, c = constante, b = coeficiente de regressão e x = pontuação no variável independente.

- **Time Series Analysis [17]**

Ao usar estatísticas clássicas, a principal preocupação é a análise de séries temporais. A análise de séries temporais envolve o desenvolvimento de modelos que melhor capturam ou descrevam séries temporais observadas para entender as causas subjacentes. Este campo de estudo busca o "porquê" por trás de um conjunto de dados da série temporal. Isso geralmente envolve fazer suposições sobre a forma de dados e decompor as séries temporais em componentes de constituição. A qualidade de um modelo descritivo é determinada pela forma como descreve os dados disponíveis e a interpretação que fornece para melhor informar o domínio do problema O objetivo principal da análise de séries temporais é desenvolver modelos matemáticos que fornecem descrições plausíveis a partir de dados de amostra. A análise de séries temporais fornece um conjunto de técnicas para entender melhor um conjunto de dados. Talvez a mais útil seja a decomposição de uma série temporal em 4 partes constituintes:

- Nível. O valor da linha de base para a série se fosse uma linha reta.
- Tendência. O comportamento crescente ou decrescente, opcional e muitas vezes linear, da série ao longo do tempo.

- Sazonalidade. Os padrões de repetição opcionais ou ciclos de comportamento ao longo do tempo.
- Ruído A variabilidade opcional nas observações que não pode ser explicada pelo modelo.

    - **Time Series Forecasting [17]**

Fazer previsões sobre o futuro é chamado de extrapolação no tratamento estatístico clássico de dados da série temporal. Os campos mais modernos se concentram no tópico e se referem a ele como previsão de séries temporais. A previsão envolve a inclusão de modelos em dados históricos e a sua utilização para prever futuras observações. Os modelos descritivos podem emprestar para o futuro (por exemplo, para suavizar ou remover o ruído), eles só procuram melhor descrever os dados. Uma distinção importante na previsão é que o futuro está completamente indisponível e só deve ser estimado a partir do que já aconteceu. A habilidade de um modelo de previsão de séries temporais é determinada pelo seu desempenho na previsão do futuro. Isto é muitas vezes à custa de poder explicar por que uma previsão específica foi feita, intervalos de confiança e até mesmo uma melhor compreensão das causas subjacentes ao problema.

Os dados da série temporária geralmente requerem limpeza, dimensionamento e transformação uniforme. Por exemplo:

- Frequência: Talvez os dados sejam fornecidos em uma freqüência que seja muito alta para modelar ou seja espaçada de forma desigual ao longo do tempo que requer uma nova implementação para uso em alguns modelos.
- Outliers Talvez haja valores ultrapassados corruptos ou extremos que precisam ser identificados e tratados.
- Faltando. Talvez haja lacunas ou dados faltantes que precisam ser interpolados ou imputados.

7. Benchmark

Benchmark 1: https://www.kaggle.com/danmanlott/shanghai-licenses

O modelo que foi criado em R gerá alguns insights importantes, onde o criador do Kernel utilizou das visualizações de dados e das correlações do modelo para fazer a análise dos dados, busquei usar visualizações de dados assertivas que citei acima. além de gerar os resultados das métricas de coeficientes que mostrarei logo a frente.

Benchmark 2: https://www.kaggle.com/fangya/price-analysis-on-shanghai-car-license-plate

O modelo que foi criado em R mostra que a autora abusou das visualizações de dados realizando a análise exploratória dos dados, usei como base no momento de definir a metodologia desse projeto e também no momento de registrar insights sobre o modelo, onde a autora utilizou de um modelo de regressão linear como base.

Benchmark 3: https://www.kaggle.com/bazingasu/data-exploration

O modelo de exploração dos dados é bastante interessante, mesmo sendo um modelo mais básico consegui insights relevantes sobre a limpeza e análise dos dados, no meu benchmark de alguma forma evolui o modelo criado pela autor o que me ajudou bastante no processo.

Benchmark 4: http://www.51chepai.com.cn/paizhaojiage/

Com base de validação do modelo utilizarei a comparação das previsões do modelo do dataset onde os dados vão até outubro e comparar com os dados do final de 2017 e começo de 2018, dados postados no site de onde os dados do modelo foram minerados.

III - Metodologia

8. Pré-processamento dos Dados

Resultado da função para completar os dados de fevereiro de 2008 a partir da média de valores:

Out[134]:

	avg_deal_price	lowest_deal_price	num_bidder	num_plates
2008-01-01	23370.000000	8100.000000	20539.000000	16000.000000
2008-02-01	32379.181818	29945.454545	22616.636364	7681.818182
2008-03-01	32169.000000	31300.000000	63534.000000	9300.000000
2008-04-01	37659.000000	37300.000000	37072.000000	9000.000000
2008-05-01	36047.000000	34400.000000	26341.000000	8200.000000
2008-06-01	34947.000000	33900.000000	21208.000000	7700.000000
2008-07-01	34491.000000	33800.000000	16783.000000	6800.000000
2008-08-01	36460.000000	35900.000000	13451.000000	6000.000000
2008-09-01	31788.000000	29300.000000	11002.000000	6500.000000
2008-10-01	33224.000000	32600.000000	11882.000000	5000.000000
2008-11-01	24351.000000	21800.000000	10170.000000	5500.000000
2008-12-01	31665.000000	31000.000000	16801.000000	4500.000000

Resultados da função para transformar a coluna date em uma coluna datetime:

In [160]:
```
# converter coluna date para uma coluna datetime, caso essa conversão não seja feita o modelo vai apresentar problemas
shangaipred["Date"] = pd.to_datetime(shangaipred["Date"])
shangaipred.set_index("Date",inplace=True)
shangaipred.head()
```

Out[160]:

Date	avg_deal_price	lowest_deal_price	num_bidder	num_plates
2002-01-01	14735.0	13600.0	3718.0	1400.0
2002-02-01	14057.0	13100.0	4590.0	1800.0
2002-03-01	14662.0	14300.0	5190.0	2000.0
2002-04-01	16334.0	16000.0	4806.0	2300.0
2002-05-01	18357.0	17800.0	4665.0	2350.0

9. Implementação

Apliquei um padrão na análise de dados que consiste em um processo aplicado através do time series forecasting. .A seguir farei uma breve descrição sobre cada parte da implementação, processo que foi descrito logo acima:

01. Reunir os dados

Realizei uma análise durante alguns meses em busca de qual base de dados utilizar para esse projeto, no começo queria uma solução complexa, acabei escolhendo uma base de dados em que pouco conhecia como desafio. No início quis usar dados referentes ao meu trabalho, mas acreditei que fosse anti ético.

02. Avaliar os dados

No momento em que fiz o download das bases de dados quiz logo de cara implementar algum modelo, acabei perdendo algum tempo e vi que teria que iniciar do zero. Dessa forma resolvi fazer a análise exploratória completa dos dados.

03. Limpar os dados.

Não tive muitas dificuldades na parte de limpar os dados, pois após a análise dos dados percebi que essa seria mais fácil de ser trabalhada que outras que já encontrei, porém foi necessário fazer várias análises antes de chegar a essa conclusão. Foi necessário criar a função para completar os dados faltantes do mês de fevereiro.

04. Analisar e criar visualizações sobre a série temporal;

No início tentei aplicar o máximo de visualizações que consegui e vi que foi perda de tempo, pois tinha um monte de gráficos bonitos que nada significavam. Após isso busquei fazer análises significativas sobre o modelo , que é o meu maior interesse ao se trabalhar com séries temporais.

05. Criação do modelo;

Implementações do modelo de regressão linear para entender melhor sobre os dados e conseguir entender os padrões, aqui o machine learning foi utilizado para dividir os dados em treinamento e teste e aplicação de função do Scikit Learn e de Regressão Linear.

06. Realizar predições

O objetivo da regressão linear foi criar um modelo determinação futura a partir da análise e da regressão dos dados da série temporal.

07. Métricas para validar os modelos;

As métricas foram implementadas conforme os estudos da regressão linear, com base nas melhores métricas para se medir modelos de regressão, e os resultados são explicados em outros itens nesse documento.

08. Aplicar as funções que realizar o estacionamento das séries temporais.

Aqui apliquei as funções clássicas da literatura científica e de blogs internacionais para a resolução de problemas de séries temporais, assim como o processo de descobrir se a série temporal é estacionária ou não.

09. Criar plotagens e criação de métricas para validar os modelos;

Durante os processos de se trabalhar com time series forecasting o notebook apresenta as visualizações do processo de se estacionar o modelo para se trabalhar com a série temporal univariável.

10. Construir e aplicar um modelo ARIMA;

Conforme a literatura que analisei no projeto implementei o modelo clássico ARIMA para revelar estatísticas relevantes sobre o modelo e para prepará-lo para realizar previsões.

11. Previsões sobre o modelo;

Após o processo clássico de time series forecasting são criadas previsões sobre o modelo e seus resultados podem ser observado nas visualizações.

Após fazer o pré processamento dos dados obtive o dataset completo de 2002 até 2017. Tive uma grande dificuldade nessa parte por falta de experiência ao trabalhar com diferentes tipos de datasets, após consultar as documentações das bibliotecas e buscas nos locais adequados consegui resolver os problemas referentes a limpeza dos dados. Precisei resolver algumas dívidas técnicas de erros nos meus códigos onde tive que pesquisar soluções e realizar tentativa e erro para resolvê-los.

De forma geral tive algumas dificuldades pela minha pouca experiência com Python, cerca de 1 ano e 7 meses, somado a isso minha vontade de trabalhar com séries temporais gerou uma demora na finalização do projeto, pois precisei aprender bastante sobre trabalhar com a regressão linear e o time series forecasting.

10. Refinamento

De forma geral utilizei a aplicação de um modelo de regressão linear como forma de se obter predições e análises exploratórias do modelo, onde conseguimos os resultados obtidos e as métricas para se avaliar o modelo criado com base na

regressão. Também utilizei um modelo de time series forecasting para analisar essa base de dados como um problema de série temporal, também consegui criar o modelo e realizar previsões sobre os dados do modelo. Ambos os modelos criados geram como saída estatísticas descritivas que serão apresentadas junto aos resultados. A seguir vou explicar a aplicação dos algoritmos e suas métricas, os resultados serão apresentados na próxima sessão.

Regressão Linear:

Apliquei os algoritmos de regressão linear obtendo resultados razoáveis de predições, obtive diversas métricas a partir do treinamento do modelo e pude entender melhor sobre os dados.

Time Series Forecasting:

Trabalhei com as séries temporais assim como aprendi em minha pesquisa, primeiramente fiz análises descritivas do modelo e após isso fiz aplicações do ARIMA e de time series forecasting que expliquei posteriormente, o modelo conseguiu resultados significativos que se assemelham aos dados reais.

IV - Resultados

11. Evolução do modelo de validação

Como a base desse projeto consiste em processos estatísticos sendo na regressão linear ou no time series forecasting ficou claro para mim que o time series forecasting foi o vencedor, pois é possível se criar previsões mais assertivas, além de a análise de séries temporais por si só ser extremamente produtiva gerando insights relevantes. Esse modelo criado está de acordo com as expectativas que gerei na introdução do problema, acredito que os parâmetros finais são bem utilizados. De forma geral acredito que os resultados são confiáveis. Na próxima sessão explicarei melhor a confiabilidade e validação dos resultados do modelo, aqui vou apresentar os resultados das métricas analisando cada modelo.

Regressão Linear:

Scoring	=	0.980134810973
Coefficient	=	[1.05493475e+00 5.12246910e-04 -2.46541042e-01]
Intercept	=	-2252.66585117
Root Mean Square Error	=	299.516433859
Mean Absolute Error	=	262.569366425
Mean Square Error	=	89710.0941515
R^2	=	0.975975325936

Mean Absolute Error	=	-1257.618 (684.052)
Mean Squared Error	=	-9227395.015 (12880831.606)
Neg Mean Squared Log Error	=	-0.177 (0.488)

Explained Variance = 0.875 (0.158)
R^2 = 0.844 (0.161)
Normalized Mutual Info Score = 0.987 (0.009)

Time Series Forecasting :

Gráfico 1: Aqui podemos ver a aplicação do forecasting que seguiu um padrão baseado nos dados, observando os resultados podemos ver que a previsão seguiu próxima aos dados originais.

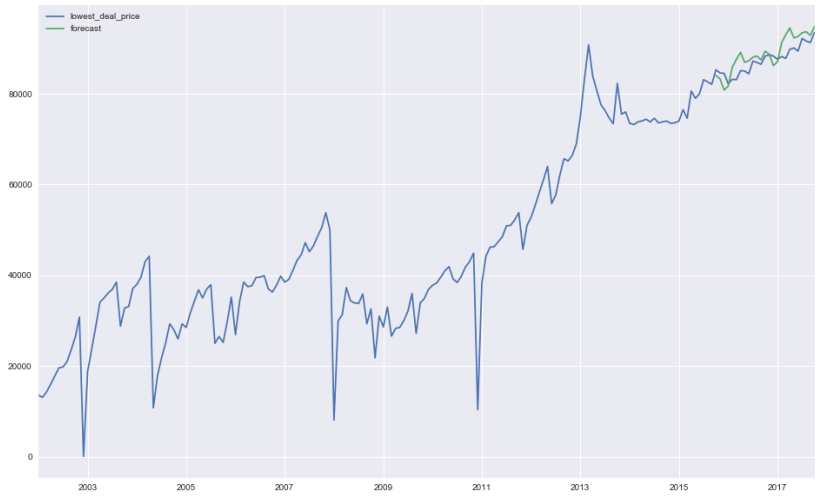
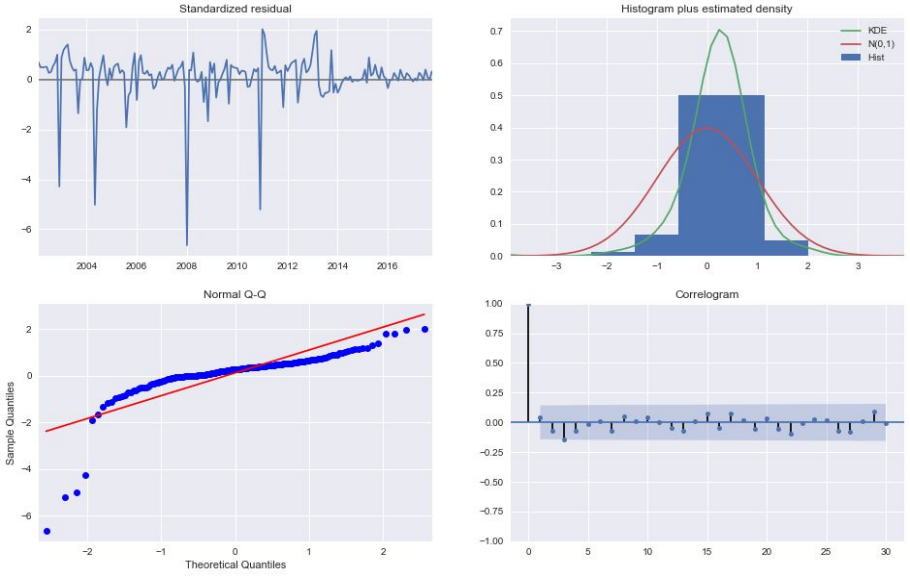


Gráfico 2: Visualização do plot_diagnostics para ver os melhores plots que representam as melhores visualizações para essa base de dados.



12. Justificativa

Para ter referenciais conclusivos sobre a eficácia desse modelo, criei uma função para completar o mês de novembro e poder comparar com dados atualizados que achei. Benchmark com dados reais: http://www.51chepai.com.cn/paizhaojiage/. De forma geral acredito que o modelo atingiu resultados interessantes e com um pouco mais de trabalho pode ser usado em produção.

Dados Forecasting

	avg_deal_price	lowest_deal_price	num_bidder	num_plates	forecast
2017-01-01	87685.0	87600.0	232101.0	12215.0	87052.451704
2017-02-01	88240.0	88200.0	251717.0	10157.0	91255.006393
2017-03-01	87916.0	87800.0	262010.0	10356.0	92981.148757
2017-04-01	89850.0	89800.0	252273.0	12196.0	94523.382474
2017-05-01	90209.0	90100.0	270197.0	10316.0	92306.959886
2017-06-01	89532.0	89400.0	244349.0	10312.0	92633.954192
2017-07-01	92250.0	92200.0	269189.0	10325.0	93441.837445
2017-08-01	91629.0	91600.0	256083.0	10558.0	93682.208018
2017-09-01	91415.0	91300.0	250566.0	12413.0	92854.367603
2017-10-01	93540.0	93500.0	244868.0	11388.0	94771.744437
2017-11-01	90226.6	90150.0	253335.3	11023.6	92550.306091

Após isso consultei o local de onde os dados iniciais foram retirados e percebi que havia uma atualização com os meses de novembro, dezembro de 2017 e janeiro de fevereiro de 2018, podemos ver que as previsões chegam mais perto do valor original que a função de criar o resultado com base na média dos dados.

Dados Originais

Date	avg_deal_price	lowest_deal_price	num_bidder	num_plates
2017-01-01	87685.0	87600.0	232101.0	12215.0
2017-02-01	88240.0	88200.0	251717.0	10157.0
2017-03-01	87916.0	87800.0	262010.0	10356.0
2017-04-01	89850.0	89800.0	252273.0	12196.0
2017-05-01	90209.0	90100.0	270197.0	10316.0
2017-06-01	89532.0	89400.0	244349.0	10312.0
2017-07-01	92250.0	92200.0	269189.0	10325.0
2017-08-01	91629.0	91600.0	256083.0	10558.0
2017-09-01	91415.0	91300.0	250566.0	12413.0
2017-10-01	93540.0	93500.0	244868.0	11388.0
2017-11-01	93130.0	93100.0	226911.0	11002.0

Também percebi que tanto a função que criei com a média dos valores, quanto às previsões apontam uma queda no valor mínimo a partir de outubro, analisando os dados originais podemos ver que a tendência após outubro segue tanto o padrão da função da média quanto do forecasting seguem o padrão dos dados originais que vou citar abaixo.

Abaixo podemos ver uma plotagem com os dados originais e fica claro de perceber que o modelo criado está assertivo.

V. Conclusão

13. Reflexão

Durante todo o processo desse projeto procurei documentar da melhor forma possível, seja nos notebooks ou aqui nesse relatório. Basicamente esse trabalho consistiu em uma comparação entre aplicações de modelos de regressão linear e também com modelos de time series forecasting. Houveram diversos aspectos interessantes no projeto, afinal, trabalhar com séries temporais é muito interessante. O problema principal consistiu em criar predições e previsões sobre a base de dados de licença de automóveis de Shangai com objetivo de se entender quais serão os valores mínimos futuros para se obter uma placa de automóvel no leilão, como quem dá o valor mínimo é o vencedor esse projeto pode ser usado claramente para a vida real. 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. Os resultados apresentados no final do time series forecasting são claros e explicativos por si só. Após a análise do item Justificativa fica claro que o modelo atingiu um resultado interessante e pode ser utilizado na prática. Com mais estudos e refinamento e melhoria no modelo é claramente possível ter bons resultados ao se prever o preço mínimo das placas de licença de automóvel.

14. 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. Outra forma de se melhorar os modelo seria evoluindo através das estatísticas que são geradas pelo modelo, tanto na regressão linear como no time series forecasting. Basicamente trabalhamos com uma única série temporal em um modelo univariado, acredito que conseguimos melhorias significativas ao se trabalhar com modelos multivariados.

Referências

[1] 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

[8] https://www.kaggle.com/arpitharavi/shanghai-notebook

[9] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[10] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

[11] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

[12] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

[13] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html

[14] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

[15] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

[16] http://www.statisticssolutions.com/what-is-linear-regression/

[17] https://machinelearningmastery.com/time-series-forecasting/

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