Machine Learning Capstone Project

Nanodegree Engenheiro de Machine Learning

Marcus Vinicius de Oliveira Cruz
25 de Novembro de 2017

Definition - Time Series Forescasting

1. Project Overview

O aluno fornece uma visão geral de alto nível do projeto em termos leigos. As informações básicas, como o domínio do problema, a origem do projeto e conjuntos de dados ou dados de entrada relacionados são fornecidos.

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.

Um modelo de série temporal basicamente consiste em um modelo estatístico que analisa uma variação temporal e consegue realizar previsões.

Para esse modelo vamos usar de três bases de dados, afim de obter maiores conclusões sobre os modelos de time series.

Os principais objetivos das séries temporais consistem em:

- Compreender o mecanismo gerador da série;
- Predizer o comportamento futuro da série.

Dessa forma ao entender um mecanismo de série temporal pode-se:

- Descrever e analisar de fato o comportamento da série temporal;
- Entender sobre as periodicidades presentes nas séries temporais;

- Entender o que o ocasiona o comportamento da série temporal;
- Controlar a trajetória da série temporal.

Dessa forma para explicar o modelo vamos utilizar 4 aplicações de funções de time series forecasting, três delas focadas em passagens de avião é um deles focado em venda de tratores. Dessa forma é possível observar o comportamento dos algoritmos de forecasting em dados de testes validados.

Porém para esse trabalho vou usar dados reais de um projeto de consultoria de machine learning que estou participando. Para preservar os dados do clientes vou alterar os nome das variáveis.

2. Problem Statement (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 valores de passagens de avião.

Teoricamente um modelo de forecasting consiste na visualização de dados a cerca 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.

Implementar um modelo de forecasting é basicamente um problema de aprendizagem supervisiona, e pode ser entendimento como uma regressão linear. Visto que o algoritmo vai fazer vasculhar a base de dados em busca de padrões de irregularidades afim de implementar previsões assertivas sobre o comportamento futuro dos dados.

3. Metrics

As métricas usadas para medir o desempenho de um modelo ou resultado estão claramente definidas. As métricas são justificadas com base nas características do problema.

Como eu implementei quatros modelos de time series forecasting funcionais temos uma base de testes interessante. Teoricamente os modelos criados no dataset real deverão se comportar com resultados positivos assim como os modelos criados nos datasets de teste.

Pelo fato de o forecasting usar muito a visualização dos dados o forecasting poderá ser visto funcionando junto a essas visualizações. Dessa forma deverão sempre ser feitos testes baseados no comportamento real dos dados.

Analysis

4. Data Exploration

Se um conjunto de dados estiver presente, os recursos e as estatísticas calculadas relevantes para o problema foram relatados e discutidos, juntamente com uma amostragem dos dados. Em vez de um conjunto de dados, uma descrição completa do espaço de entrada ou dados de entrada foi feita. Anormalidades ou características sobre os dados ou a entrada que precisam ser endereçados foram identificada**.**

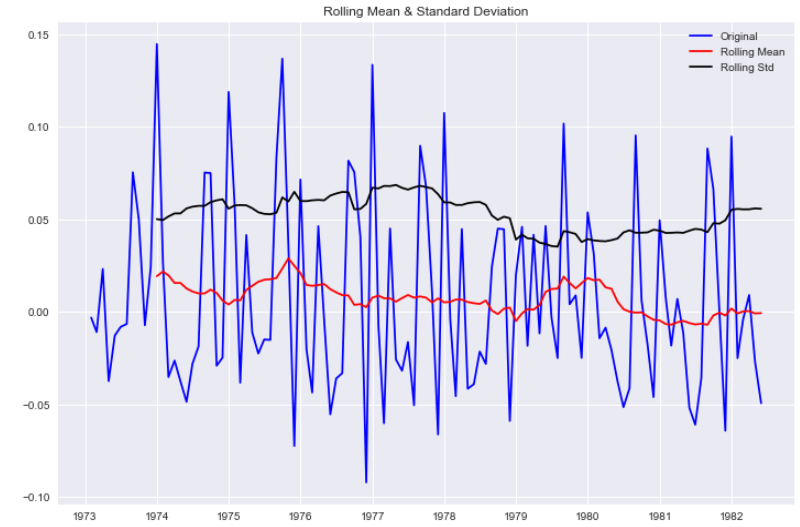
Em todo o modelo, partindo da premissa que implementei funções de forecasting em base de dados de teste para validar a solução do problema. meu objetivo foi validar o máximo possível de hipóteses antes de implementar de fato um algoritmo de machine learning propriamente dito.

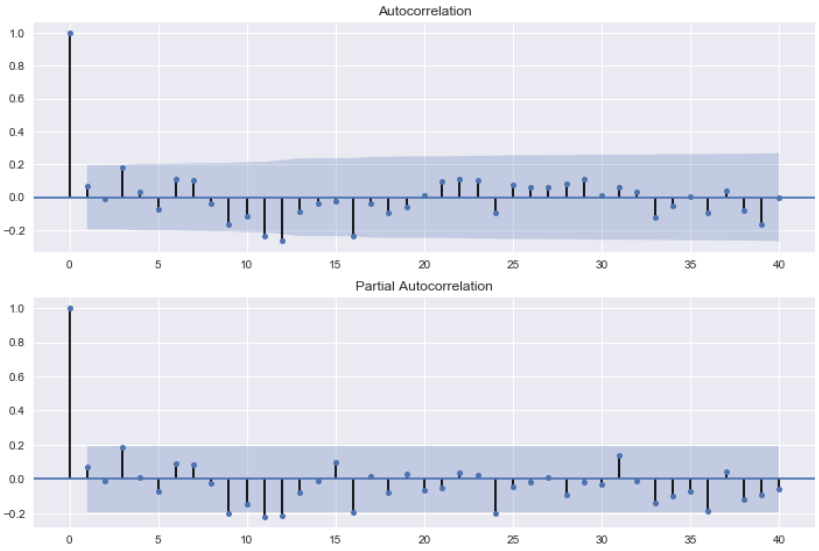
Os dados precisaram ser categorizados, onde conseguimos criar um padrão, no caso da base de teste real um crawler pegou o registro do valor atual da passagem em questão para registrar na base de dados, dessa forma, juntamente com o registro da data conseguimos criar um padrão, de forma que o mesmo algoritmo possa ser aplicado a diferentes base de dados conseguindo resultados positivos e genéricos em todas elas.

Para o futuro do modelo deverá ser criada uma função para trabalhar com diversas bases de dados de passagens diferentes aprendendo em tempo real com as mudanças e variações de preços.

5. Exploration Visualization

Foi fornecida uma visualização que resume ou extrai uma característica ou característica relevante sobre o conjunto de dados ou os dados de entrada com uma discussão aprofundada. As pistas visuais estão claramente definidas.





6. Algorithms and Techniques

Algoritmos e técnicas utilizadas no projeto são cuidadosamente discutidas e devidamente justificadas com base nas características do problema.

Uma série temporal basicamente consiste em 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.

Time series forecasting é um modelo de aprendizagem supervisionada, como tal precisa de exemplos claros e concisos para treinamento do modelo em questão.

Dessa forma em um primeiro momento esse problema tende a ser entendido como algo que possa ser solucionado com regressão linear, porém como minha pesquisa me mostra esse problema não deve ser resolvido com regressão linear por dois motivos.

Para implementar um modelo ARIMA temos os seguintes parâmetros:

- p - O número de observações de atraso incluídas no modelo, também é conhecido como ordem de atraso.
- d - O número de vezes que as observações em bruto são diferenciadas, também chamado de grau de diferenciação.
- q - Tamanho da janela da média móvel, também chamado de ordem da média móvel.

Um modelo ARIMA consiste em uma classe de modelo estatísticos para análise e previsão de dados de séries temporais. ARIMA significa AutoRegressive Integrated Moving Average

7. Benckmark

O aluno define claramente um resultado ou limite de referência para comparar os desempenhos das soluções obtidas.

1) ARIMA / SARIMA -

2) ET´s -

Methodology

8. Data Preprocessing

Todas as etapas de pré-processamento foram claramente documentadas. Anormalidades ou características sobre os dados ou entradas que precisavam ser endereçadas foram corrigidas. Se não for necessário um pré-processamento de dados, foi claramente justificado.

9. Implementation

O processo para o qual métricas, algoritmos e técnicas foram implementadas com os conjuntos de dados ou dados de entrada dados foi completamente documentado. As complicações ocorridas durante o processo de codificação são discutidas.

10. Refinement

O processo de melhoria nos algoritmos e técnicas utilizadas está claramente documentado. As soluções iniciais e finais são relatadas, juntamente com soluções intermediárias, se necessário.

Results

11. Model Evaluation and Validation

As qualidades do modelo final - como os parâmetros - são avaliadas em detalhes. Algum tipo de análise é usado para validar a robustez da solução do modelo.

12. Justification

Os resultados finais são comparados com o resultado ou o limite de referência com algum tipo de análise estatística. A justificação é feita para saber se o modelo final e a solução são suficientemente significativos para ter resolvido adequadamente o problema.

Conclusion

13. Free-Form Visualization

Foi fornecida uma visualização que enfatiza uma qualidade importante sobre o projeto com uma discussão aprofundada. As pistas visuais estão claramente definidas.

14. Reflection

O aluno resuma de forma adequada a solução de problemas de ponta a ponta e discute um ou dois aspectos específicos do projeto que eles acham interessante ou difícil.

15. Improvement

É feita uma discussão sobre como um aspecto da implementação poderia ser melhorado. As soluções potenciais resultantes dessas melhorias são consideradas e comparadas / contrastadas com a solução atual.

Quality

16. Presentation

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.

Dei a devida referência a todo material que consultei, consultei não só artigos de internet, mas também artigos acadêmicos. O referencial teórico está gigante pois procurei entender de fato os problemas e buscar a melhor ferramenta para resolver os problemas iniciais.

17. Functionality

Conforme orientação da Udacity o Jupyter Notebook está completamente explicativo. Procurei documento 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.

18. Referencias

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

https://pt.wikipedia.org/wiki/C%C3%B3digos_de_classes_da_IATA

https://skiplagged.com/flights/CNF/LAX/2017-10-25/2017-11-06

http://ucanalytics.com/blogs/python-code-time-series-forecasting-arima-models-manufacturing-case-study-example/

http://iot-ee.com/en/2017/08/07/analysing-iot-data-introduction-time-series-forecasting-python/

http://connor-johnson.com/2014/11/23/time-series-forecasting-in-python-and-r/

https://stackoverflow.com/questions/31379845/forecasting-with-time-series-in-python

https://fxdata.cloud/tutorials/a-guide-for-time-series-forecasting-with-arima-in-python-3

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

https://machinelearningmastery.com/time-series-forecast-study-python-monthly-sales-french-champagne/

https://machinelearningmastery.com/introduction-to-time-series-forecasting-with-python/

https://machinelearningmastery.com/introduction-to-time-series-forecasting-with-python/

https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-prophet-in-python-3

https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3

http://www.seanabu.com/2016/03/22/time-series-seasonal-ARIMA-model-in-python/

https://machinelearningmastery.com/convert-time-series-supervised-learning-problem-python/

https://machinelearningmastery.com/time-series-forecasting-supervised-learning/

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

http://iot-ee.com/en/2017/08/07/analysing-iot-data-introduction-time-series-forecasting-python/

http://www.dme.ufrj.br/dani/pdf/slidespartefrequentista.pdf