Data Science
Pre Work

Saint Paul | LIT

Pre Work módulo de Data Science
Fabrício J. Barth

Remova a seguinte mensagem quando imprimir!
“Por favor, pense em sua responsabilidade ambiental antes de imprimir este documento. Caso precise imprimir este material, considere esta página como capa para sua impressão, minimizando o desperdício de tinta.”

Sumário

Introdução e objetivos desta atividade

Data Science é uma área interdisciplinar voltada para o estudo e a análise de dados, estruturados e não-estruturados, que visa a extração de conhecimento, detecção de padrões e/ou obtenção de *insights* para possíveis tomadas de decisão [1]. É uma área que une conceitos, técnicas e ferramentas da Estatística, Aprendizagem de Máquina, Banco de Dados e outras áreas. Inclusive, é uma área que está diretamente conectada com Inteligência Artificial no que tange projetos de mercado.

A prática de Data Science é baseada em um conjunto de metodologias, processos, conceitos e ferramentas. É objetivo deste curso discutir e exercitar tais metodologias, processos, conceitos e ferramentas para que os participantes do curso possam fazer uso da prática de Data Science no seu dia-a-dia. Ao compreender melhor a prática de Data Science, espera-se que os participantes do curso possam: (i) identificar potenciais projetos de Data Science; (ii) avaliar a complexidade de cada projeto; (iii) acompanhar a execução dos projetos; (iv) identificar skills relevantes para a formação de equipes de Data Science, e; (v) interagir com profissionais técnicos da área. Este curso não tem como objetivo formar engenheiros de machine learning, engenheiros de dados, cientista de dados ou qualquer outro profissional que compreenda em profundidade os algoritmos e ferramentas normalmente empregados no desenvolvimento de modelos preditivos ou descritivos. No entanto, é também objetivo deste curso discutir os conceitos e cenários de utilização dos principais algoritmos da prática de Data Science.

Para que possamos aproveitar ao máximo o nosso encontro, eu sugiro a leitura de alguns artigos e execução de alguns exercícios. Sugiro a leitura dos seguintes textos:

- "*From Data Mining to Knowledge Discovery in Databases*" do Fayyad [2]. Este texto descreve o processo conhecido como KDD (*Knowledge Discovery in Databases*). Apesar do termo utilizado no texto ser *Data Mining*, os conceitos e processos descritos no mesmo se aplicam perfeitamente na prática de *Data Science.*
- "*The Discipline of Machine Learning*" do Tom Mitchell [3]. O texto do Fayyad [2] fornece uma visão da área de pesquisa em Banco de Dados. O texto do Tom Mitchell [3] fornece uma visão da área de Inteligência Artificial, mais especificamente, Aprendizagem de Máquina. O texto do Tom Mitchell [3] é um texto introdutório sobre o assunto discutindo exemplos de aplicações, desafios da área e alguns outros aspectos.
- "*Tidy Data"* do Hadley Wickham [4]. É um texto um pouco mais técnico que descreve como os dados precisam estar estruturados para que possamos fazer qualquer análise preditiva ou descritiva.

- "*Foudational Methodology for Data Science*", um white paper da IBM [5]. Este texto apresenta uma versão mais moderna do processo de KDD. Ele está disponível como referência no curso de *Data Science Practitioner* do IBM Skills Academy [6].
- "*Data Science and Prediction"* do Vasant Dhar [7]. Apresenta alguns conceitos que vamos discutir em sala de aula, como por exemplo, a capacidade de generalização dos modelos.

Os exercícios que devem ser executados são apresentados nos próximos capítulos deste material.

Datas e prazos

O nosso encontro será no dia 30 de julho de 2020. Peço que cada participante envie este caderno preenchido **até o dia 27 de julho de 2020**. A atividade deverá ser entregue **via Blackboard** até às 23h59 dessa data.
No dia 30 de julho nós vamos discutir os casos de uso descritos neste documento. Por favor, mantenha este documento com você durante o dia 30 de julho pois a atividade final da disciplina irá utilizar como base este documento.

Materiais complementares

Para a realização do Pre Work, além da leitura dos artigos, os alunos poderão complementar os seus estudos com a execução do curso de *Data Science Practitioner*, do IBM Skills Academy. Caso você não tenha conhecimento de estatísticas descritivas, também recomendamos a execução do curso de Métodos Quantitativos Aplicados aos Negócios, do LIT, como o primeiro passo para os seus estudos.

A execução desses cursos não é pré-requisito para esta disciplina, mas você pode fazer se desejar. Seguem as indicações das seções mais relevantes desses cursos.

a) *Data Science Practitioner*, do IBM Skills Academy – assistir os vídeos:
   - Lecture 1 - Data Science Landscape
   - Lecture 1 - The Future of Cognitive Computing
   - Lecture 2 - Data Methodology
   - Lecture 3 - Data Science on the Cloud
   - Lecture 4 - Explore and Prepare Data

- Lecture 5 - Represent and Transform Data
- Lecture 6 - Data Visualization and Presentation
- Lecture 7 - Data Modeling

b) Métodos Quantitativos Aplicados aos Negócios, do LIT – assistir os vídeos:

- Visão geral da estatística descritiva e suas principais medidas
- Medidas de posição - Média aritmética, mediana e moda
- Medidas de dispersão
- Covariância
- Correlação

Atividades

Criação de um modelo que determina o preço de imóveis na cidade de São Paulo

Um imobiliária de grande porte com atuação na cidade de São Paulo deseja criar um modelo de preços de imóveis para a cidade de São Paulo. Este modelo deverá ser incorporado aos sistemas que são utilizados pelos corretores de imóveis da empresa. O objetivo é que o preço do imóvel seja automaticamente calculado por este modelo, servindo de base para as negociações. Deve-se levar em consideração informações como tamanho, quantidade de quartos, quantidade de suítes, quantidade de vagas de estacionamento, entre outras informações para o cálculo do preço.

Esta imobiliária possui uma base histórica de imóveis e seus respectivos preços e pretende utilizar esta base para desenvolver o modelo descrito acima. Parte desta base pode ser encontrada no dataset [Base de dados com imóveis da cidade de São Paulo]. Para que você tenha acesso a este dataset basta acessar o link descrito no mesmo. Faça o download do dataset e responda as perguntas abaixo:

1. O dataset disponível respeita as regras de “tidy data” descritos por [4]? Justifique a sua resposta.

2. Qual é a relação entre o conceito de “tidy data” descrito por [4] e as etapas de “data cleaning”, “data access” e “data reduction” descritos por [2]?

3. Faça uma análise exploratória dos dados. Você poderá fazer uso do próprio Excel ou de qualquer outra ferramenta que desejar. O importante nesta análise é entender um pouco sobre as características dos dados disponíveis. Quais são os tipos dos atributos (categórico, numérico)? Quantos exemplos temos disponíveis? Qual é a distribuição dos valores para cada atributo (mínimo, média, mediana, máximo)? Se você quiser, pode até calcular a correlação entre atributos, principalmente entre o atributo dependente (saída do modelo) e os independentes (entrada do modelo). Se você desejar, você também pode fazer uso de qualquer livro de estatística descritiva para te auxiliar nesta atividade.

4. Este é um problema de regressão, classificação ou clustering? Justifique a sua resposta.

5. Considere que um modelo foi criado. Este modelo recebe as informações sobre o bairro onde o imóvel está, a área do imóvel em metros quadrados, a quantidade de suítes, quantidade de dormitórios, quantidade de banheiros e quantidade de vagas, e retorna o preço do imóvel. Como sabemos se este modelo é bom? Como sabemos se este modelo pode ser utilizado pelos corretores de imóveis?

6. O texto [5] descreve uma metodologia para Data Science ligeiramente diferente do processo descrito por [2]. A principal diferença está depois de “modeling” e “evaluation”. Depois destas etapas o texto [5] descreve as etapas de deployment e feedback. O que são estas etapas e como podemos implementar neste caso de uso? Principalmente no que diz respeito a etapa de feedback?

Análise de Risco de Crédito

Neste dataset, cada linha representa uma pessoa que solicitou crédito para um banco. Cada pessoa é classificada como "good" ou "bad" em termos de risco de crédito. O dataset original tem 1000 exemplos e 20 atributos e está disponível em [8]. O dataset que temos aqui já é uma versão onde boa parte dos atributos numéricos foram transformados em categóricos e o número de atributos reduzidos de 20 para 10. Este dataset pode ser encontrado em [Base de dados sobre análise de risco de crédito]. Para que você tenha acesso a este dataset basta acessar o link descrito no mesmo. Além do atributo Risk, este dataset tem os seguintes atributos:

1. Age (numeric)
2. Sex (text: male, female)
3. Job (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
4. Housing (text: own, rent, or free)
5. Saving accounts (text - little, moderate, quite rich, rich)
6. Checking account (numeric, in DM - Deutsch Mark)
7. Credit amount (numeric, in DM)
8. Duration (numeric, in month)
9. Purpose (text: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)

Faça o download do dataset e responda as perguntas a seguir:

1. O dataset disponível respeita as regras de "tidy data" descritos por [4]? Justifique a sua resposta.

2. Faça uma análise exploratória dos dados. Você poderá fazer uso do próprio Excel ou de qualquer outra ferramenta que desejar. O importante nesta análise é entender um pouco sobre as características dos dados disponíveis. Quais são os tipos dos atributos (categórico, numérico)? Quantos exemplos temos disponíveis? **Existem valores faltantes (missing)?** Qual é a distribuição dos valores para cada atributo (mínimo, média, mediana, máximo)? Se você quiser, pode até calcular a correlação entre atributos, principalmente entre o atributo dependente (saída do modelo) e os independentes (entrada do modelo). Se você desejar, você também pode fazer uso de qualquer livro de estatística descritiva para te auxiliar nesta atividade.

3. Este é um problema de regressão, classificação ou clustering? Justifique a sua resposta.

4. Como sabemos se este modelo é bom? Como sabemos se este modelo pode ser utilizado por uma instituição financeira?

5. Considerando que este modelo é bom, como podemos fazer o “deploy” dele? Como podemos coletar “feedback” sobre o seu uso? Qual é a utilidade do “feedback” neste caso? Entendemos “deploy” de um modelo como a atividade de colocar este modelo em uso. Ou seja, os usuários ou a empresa irá de fato utilizar este modelo.

Descobrir segmentos de adolescentes em redes sociais

Interagir com amigos em redes sociais, tais como Facebook, tem se tornado um ritual entre os adolescentes do mundo todo. Estes adolescentes, ao mesmo tempo que interagem com os seus amigos, também estão sujeitos a propagandas de diversos produtos e empresas. É muito importante para as empresas, e até mesmo para os usuários de redes sociais, que tais propagandas sejam melhor direcionadas, ou seja, que sejam entregues para aqueles que realmente tem interesse naquele tipo de produto. Dado o texto produzido por adolescentes em Serviços de Redes Sociais, acredita-se que podemos identificar grupos de adolescentes que compartilham interesses em comum, tais como esporte, religião ou música. Algoritmos de clustering podem automizar este processo de descoberta de segmentos naturais desta população. No entanto, cabe a vocês decidirem se tais agrupamentos são interessantes ou não para o uso em propaganda.

Para esta análise, nós vamos utilizar um dataset que representa uma amostra aleatória de 30.000 estudantes de high school americana que possuem perfil em uma rede social em 2006. Para proteger o anonimato dos usuários da rede, o nome da rede social não é fornecido. O dataset pode ser encontrado em [Base de dados sobre comportamento de adolescentes em redes sociais]. Os atributos deste dataset são:

- **gradyear**: ano de graduação;
- **gender**: sexo, masculino ou feminino;
- **age**: idade representada por um número real;
- **friends**: quantidade de amigos na rede social, e;
- **demais atributos**: basketball, football, soccer, entre outros indicam classes onde os textos das pessoas foram classificados. Textos de usuários publicados na rede social foram processados para gerar esta tabela. Como esta tabela foi gerada é um problema de classificação de texto e deverá ser abstraído, por enquanto.

Normalmente, segue-se as seguintes etapas para a realização desta análise:

- Aquisição, análise exploratória e pré-processamento dos dados;
- Construção do modelo;
- Análise do modelo, e;
- Preparação do relatório e principais artefatos que devem ser entregues.

Faça o download do dataset e responda as perguntas a seguir:

1. Faça uma análise exploratória dos dados. Você poderá fazer uso do próprio Excel ou de qualquer outra ferramenta que desejar. O importante nesta análise é entender um pouco sobre as características dos dados disponíveis. Quais são os tipos dos atributos (categórico, numérico)? Quantos exemplos temos disponíveis? Existem valores faltantes (missing)? Qual é a distribuição dos valores para cada atributo (mínimo, média, mediana, máximo)? Se você desejar, você também pode fazer uso de qualquer livro de estatística descritiva para te auxiliar nesta atividade.

2. O que queremos aprender com este projeto? Existe algum atributo no dataset que será o output do modelo?

3. Este é um problema de classificação, regressão ou clustering? Justifique a sua resposta.

4. Como podemos usar o resultado desta modelagem? Qual é o melhor formato (relatório estático, relatório dinâmico, componente de software)?

Questões gerais

Responda as questões abaixo:

1. Considerando a sua área de atuação, liste três casos de uso que você considera como potencial projeto de Data Science. Para cada caso de uso descreva a pergunta que precisa ser respondida e qual é o valor de negócio ao respondermos esta pergunta.

2. Para cada caso de uso descreva quais seriam os atributos e quais são os exemplos do dataset.

3. Descreva também qual seria o processo para obter este dataset.

Bases de dados

Base de dados com imóveis da cidade de São Paulo

Link:https://github.com/fbarth/ds-saint-paul/raw/master/data/20140917_imoveis_filtrados_final_csv_shaped.xlsx

Base de dados sobre análise de risco de crédito

Link: https://github.com/fbarth/ds-saint-paul/raw/master/data/german_credit_data.xlsx

Base de dados sobre comportamento de adolescentes em redes sociais

Link: https://github.com/fbarth/ds-saint-paul/raw/master/data/snsdata.xlsx

Referências

[1] Verbete da Wikipedia sobre Ciência de Dados [https://pt.wikipedia.org/wiki/Ci%C3%AAncia_de_dados]. Acessado em Junho de 2020.

[2] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37. https://doi.org/10.1609/aimag.v17i3.1230 [https://github.com/fbarth/ds-saint-paul/blob/master/references/fay1996.pdf]

[3] Mitchell, T. The Discipline of Machine Learning. 2006. http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf [https://github.com/fbarth/ds-saint-paul/blob/master/references/MachineLearning.pdf]

[4] WICKHAM, Hadley. Tidy Data. **Journal of Statistical Software**, [S.l.], v. 59, Issue 10, p. 1 - 23, sep. 2014. ISSN 1548-7660. Available at: <https://www.jstatsoft.org/v059/i10>. doi:http://dx.doi.org/10.18637/jss.v059.i10. [https://github.com/fbarth/ds-saint-paul/blob/master/references/tidy-data.pdf]

[5] *Foudational Methodology for Data Science.* Material que faz parte do curso Data Science Practitioner da plataforma Skills Academy da IBM. https://skills-academy.mylearnerportal.com/ [https://github.com/fbarth/ds-saint-paul/blob/master/references/foundational_ds_ibm.PDF]

[6] Plataforma Skills Academy da IBM. https://skills-academy.mylearnerportal.com/

[7] Vasant Dhar. 2013. Data science and prediction. Commun. ACM 56, 12 (December 2013), 64–73. DOI:https://doi.org/10.1145/2500499 [https://github.com/fbarth/ds-saint-paul/blob/master/references/ds_prediction.pdf]

[8] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.