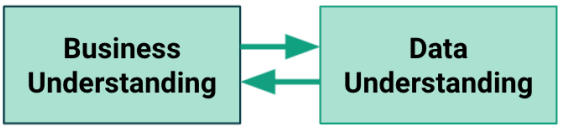
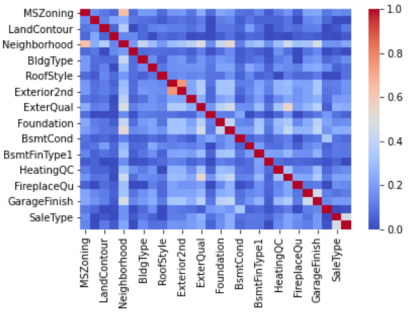
Roteiro de um projeto de Ciência de Dados em 7 passos!

Analisar o Contexto do Problema



- Qual o problema que deseja resolver?
- Como você espera usar e se beneficiar deste modelo?
- Qual a solução no momento (se houver)?
- Qual modelo você usará? supervisionado ou não-supervisionado? classificação ou regressão?
- Quais métricas você usará para medir o desempenho do modelo?
- Quais hipóteses foram feitas até agora (por você ou por outros)?

Obter os dados, descobrir e visualizar os dados para obter informações



- Download dos dados
- Verificar a estrutura dos dados
- Criar um conjunto de teste para evitar viéses pré-análises
- Visualizar os dados graficamente para identificar padrões
- Buscar correlações entre cada par de atributos
- Experimentando combinações de atributos

Preparar os dados para os algoritmos de Machine Learning

- Criar funções para limpar os dados para otimização das tarefas.

- Manipular textos e transformar textos em atributos categóricos.
- Customizar transformadores: combinação de variáveis.
- Escalonar características (normalizar os dados).

Selecionar e treinar um modelo

1a Etapa
Conjuntos de Treino do Modelo - Problema é supervisionado ou não supervisionado? De classificação, regressão?

2a Etapa
Conjunto de técnicas a serem avaliadas no ajuste do modelo, por exemplo, Regressão Logística, Random Forest, XGBoost, Redes Neurais.

3a Etapa
Definição das métricas de qualidade de ajuste do seu modelo, por exemplo, acurácia, sensibilidade, especificidade, RMSE?

4a Etapa
Tabela resumo com as medidas de acurácia de cada modelo.

Refinar seu modelo

Com os algoritmos escolhidos, é hora de refinar seus resultados considerando o conjunto de validação.

- Ajuste dos hiperparâmetros do modelo.
- Análise do melhor modelo e os erros no conjunto de validação.
- Investigar possíveis viéses do seu modelo, avaliando por exemplo, a acurácia por **variáveis-chave**
- **Interpretabilidade** de Modelos Opacos: Importância de variáveis, PDP, ICE, ALE, LIME, Shapley Values.
- Apuração dos resultados no **conjunto de teste**.

Apresentar sua solução

- Traduzir para o contexto do problema!
- Divida o seu processo:
    - O que você aprendeu?
    - O que funcionou? O que não funcionou?
    - Quais as limitações?
- Dividir os resultados de forma clara, objetiva e utilizar recursos de visualização!
- Qual vai ser a utilização do modelo?
- Deploy, Monitoramento do Modelo e suas variáveis, Manutenção

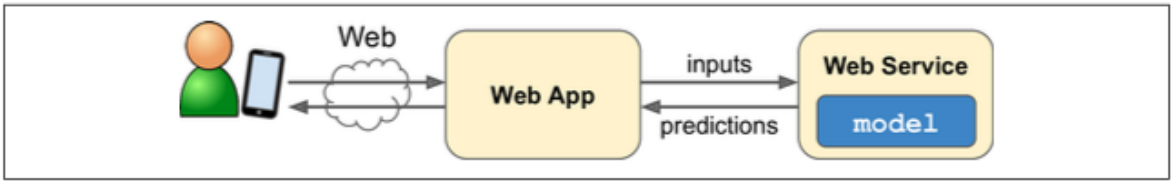


Imagem retirada do livro Hands On Machine Learning - Aurelién Géron

Referências

- Livro Hands On Machine Learning - Aurelién Géron. Repositório com os códigos disponível em https://github.com/ageron/handson-ml2
- Livro Aprendizado de Máquina, disponivel em http://www.rizbicki.ufscar.br/AME.pdf
- Algoritmos de Destruição em Massa - Cathy O'Neil

Material produzido pelo Grupo de Estudos Data Science - Pyladies São Paulo. Novembro de 2021.

Licença de uso