Netflix Recommendation

Método utilizado: Ultra Learning - Scott H. Young e CRISP-DS.

Por que esse projeto?

Esse vai ser meu primeiro projeto de machine learning. Vou desenvolver autonomia e uma carência que eu tenho com Machine Learning.

- Cinco coisas que desejo trabalhar nesse projeto:
  - Analise de Dados.
  - Machine Learning.
  - Entender sobre o negócio.
  - Autonomia.
  - Storytelling.

O que é necessário para fazer o projeto?

Conceitos:

Fatos:

Pratica:

Como vou buscar o conhecimento que não tenho?

1. Comunidade DS (tutores, perguntas e etc)
2. Livros.
3. Artigos.
4. Sites confiáveis.

Ideias de Tarefas Interessantes:

1. Compreender qual conteúdo está disponível em diferentes países
2. A Netflix tem mais foco em programas de TV do que filmes nos últimos anos.

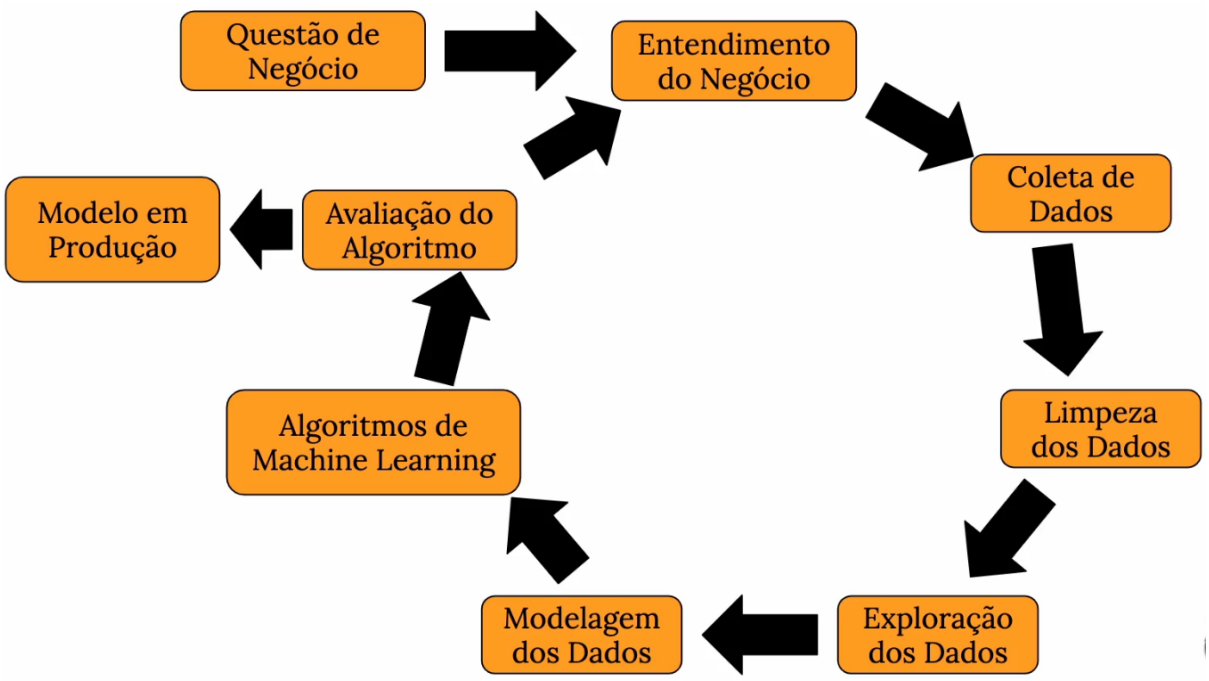
Questões:

Perguntas:

1. Com que tipo de dado eu estou lidando?
2. Que forma eu posso tratar os dados?
3. Qual é o tipo de analise que eu estou lidando?
4. Quais são os melhores gráficos para esse tipo de analise?
5. Como eu posso selecioar as melhores variaveis com algoritmo?
6. Como eu posso modelar os dados para o algoritmo aprender?
7. Quais os 3 principais modelos de machine learning para recomendação?
8. Como eu posso avaliar meu modelo?
9. Como eu explicaria para uma criança?
10. Como eu posso fazer o deploy?

Respostas:

1. Categorico.
2. Apenas zerei os NaN.
3. Analise com variaveis categoricas
4. BarPlot, CatPlot, graficos para mostrar as categorias.



Machine Learning Algoritmos de Recomendação:

TF-IDF e Similaridade Cosseno

pelo que entendi da sua pergunta vc deve estar fazendo algum projeto de NLP, certo? deve estar usando bag of words para montar seu dataframe.... Assumindo como verdade minha suposição estas duas métricas podem ajudar na hora de montar a matriz bag of words.... em principio usamos a contagem dos tokens (palavras) como entrada da matriz bag of words, entretanto não fica bom para fazer o treinamento do algoritmo de ML com tais entradas, pois a contagem leva a numeiros inteiros grandes. O que se faz é normalizar estes numeros inteiros grandes (as contagens dos tokens) ... uma forma de fazer isso é essa tal de TF-IDF que é nada mais que uma noramlização desses inteiros... Agora cossine similarity é usado para comparar o quantos dois vetores são similares. No contexto de um projeto de NLP pode ser usado para comparar o quanto duas linhas do data frame são similares, ou seja, cada linha pode ser entendida como um vetor e aí aplica-se o cossine similarity para ver o quanto as duas linhas são similares.