Elementos de Inteligência Artificial e Ciência de Dados — Assignment 2

Matheus Teixeira - 202110548

Proposta de trabalho

- Dataset de video-games
- Avaliação e exploração dos atributos do dataset
- Limpeza do dataset
  - Lidar com valores nulos
  - Exclusão de atributos irrelevantes
- Criação de novos atributos a partir de já existentes
- Normalização de atributos numéricos
- Avaliação de modelos através de k-fold cross validation
  - Decision Trees
  - k-nearest neighbors
- Análise da matriz de confusão

Bibliografia utilizada

- One hot encoding
  - https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/
- Feature engineering
  - https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10
- Cross validation
  - https://machinelearningmastery.com/k-fold-cross-validation/

Ferramentas e algoritmos

- Manipulação de dados
  - Pandas
  - NumPy
- Gráficos e visualizações
  - Plotly
  - Seaborn
- Modelos de aprendizado de máquina
  - Sci-kit Learn

Pré-processamento dos dados

- Remoção das linhas que contém algum valor faltando
  - Aproximadamente 1.2% do dataset original
- Remoção dos jogos não ‘main_game’
  - Aproximadamente 14% do dataset original
- Seleção de atributos claramente irrelevantes para serem removidos
  - ‘id’
  - ‘name’
  - ‘year’
  - ‘sumarry’

Exploratory data analysis

- Matriz de correlação das variáveis numéricas
  - ‘follows’ e ‘n_user_reviews’ consideravelmente relacionadas com ‘user_score’
- Distribuição de ‘user_score’
  - Apresenta um desbalanceamento grave
- Análise da quantidade de aparições e distribuição das variáveis categóricas ‘genre’, ‘companies’ e ‘platforms’
  - Os atributos apresentam distintas distribuições de ‘user_score’ dentre eles, o que os tornam atributos interessantes de estarem representados no dataset final

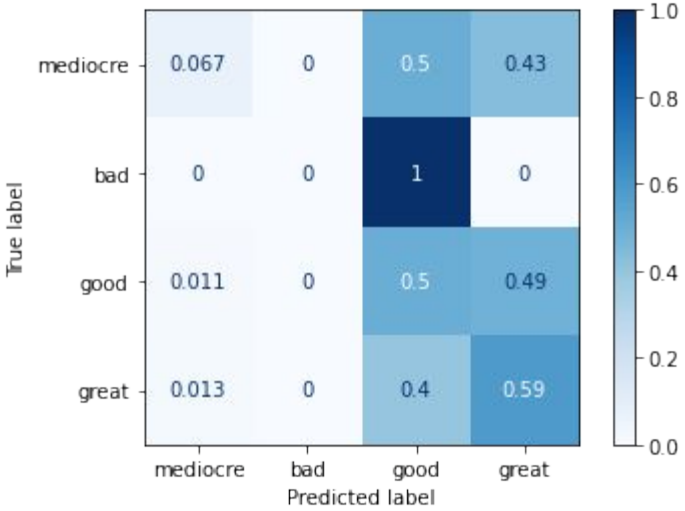
Feature engineering

- Foi realizado o one-hot-encoding das variáveis ‘genre’, ‘companies’ e ‘platforms’
- Possuem uma gama grande de valores distintos
- Foram então selecionados somente os valores mais presentes
- Para ‘platforms’, foram somente considerados as dos principais consoles da Microsoft, Sony e Nintendo a partir da era do Nintendo 64, além da plataforma PC

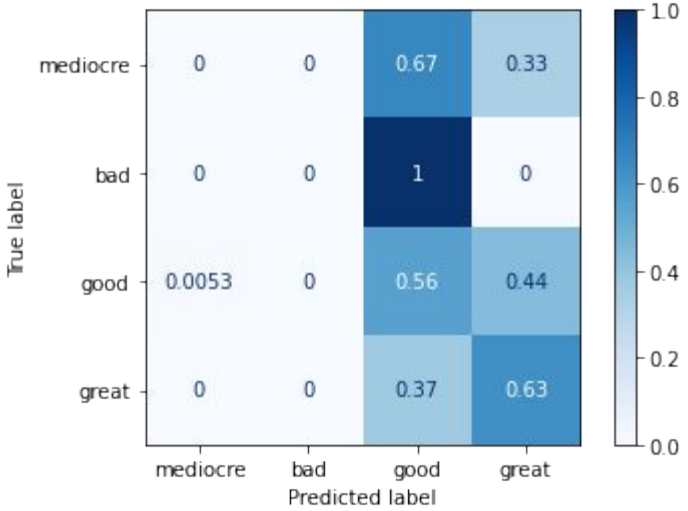
Treino e avaliação dos modelos

- Foi utilizado do k-fold cross validation para a seleção de hiperparâmetros
  - kNN:
    - k = 23
    - score: ~0.56
  - Decision Tree:
    - Max depth = 6
    - score: ~0.6
- Resultados não muito satisfatórios
  - Será explorado as razões no próximo slide

Matriz de confusão



kNN



Decision Tree

Conclusões

- O desbalanceamento de classes se mostrou um problema incontornável
- Problemas como este eram esperados na Decision Tree, mas foram uma surpresa no kNN
- Novas abordagens de feature engineering poderiam ser exploradas
  - Porém, o problema do desbalanceamento de classes ainda parece ser o principal
- Uma nova discretização de ‘user_rating’ para ‘user_score’ poderia ser realizada
- A conversão do problema para de regresão, usando ‘user_rating’ como label, se provou uma boa solução em alguns testes rápidos