Ficha de Exercícios 01

Gonçalo Rodrigues Pinto - A83732
Universidade do Minho

(26 de Fevereiro de 2021)

Resumo

O presente relatório descreve o trabalho de introdução à metodologia CRISP-DM. O presente trabalho teve como objetivo a compreensão das diferentes etapas da metodologia CRISP-DM. De forma ao seu desenvolvimento colocou-se em prática as primeiras etapas da metodologia (*Business Evaluation*, *Data Evaluation*). Assim foi possível através de um problema que pudesse ser enquadrado dentro do processo de *Data Mining* perceber estas primeiras etapas da metodologia. Posto isto, foi possível retirar os diversos de tipos de benefícios que se esperou retirar da aplicação de *Data Mining*

1 Introdução

No 2º semestre do 1º ano do Mestrado em Engenharia Informática da Universidade do Minho, existe uma unidade curricular enquadrada no perfil de Descoberta do Conhecimento denominada por Engenharia do Conhecimento, que tem como objetivo a introdução ao conceito de descoberta do conhecimento.

A presente ficha enquadra-se nesta unidade curricular e pretende introduzir a metodologia CRISP-DM.

Nesta ficha pretendeu-se estudar as primeiras etapas da metodologia acima referida, mais propriamente *Business Understanding*, compreensão do negócio, que consiste em compreender os objetivos e requisitos do projeto além de determinar e consolidar qual o objetivo a atingir com o processo de Data Mining. Outras das etapas estudadas consistiu em *Data Understanding*, compreensão dos dados, para isso é necessário a recolha, exploração e familiarização com os dados bem como identificar problemas de qualidade nos dados.

2 Questões

1. ***Identifique um problema que possa ser enquadrado dentro do processo de Data Mining. Para esse problema descreva sucintamente as seguintes fases do processo CRISP-DM:***
   Utilizando, por exemplo, uma das técnicas de *data mining* que é a classificação, é possível construir modelos (funções) que descrevem e distinguem classes ou conceitos para previsão futura.

   Assim sendo, dispondo do conjunto de dados[4] que foi utilizado na competição aberta do Prêmio Netflix para o melhor algoritmo para prever classificações de usuários para filmes é possível definir modelos de recomendação de conteúdos presentes na Netflix, que permite aumentar o lucro por parte desta plataforma de *streaming* pois fideliza os utilizadores no modo que fornece sempre conteúdo interessante e que vai ao encontro dos gostos utilizadores. Por exemplo, quando a *La Casa de Papel* foi lançada nesta plataforma como tinha origem espanhola é normal que seja recomendada a utilizadores de Portugal dada a localização geográfica destes dois países. Ou, ainda, quando é tido em conta o dia da semana por exemplo a Netflix descobriu que pessoas assistem séries durante a semana e filmes durante o fim de semana.

   (a) ***Business Understanding***
   - O objetivo do negócio, neste caso, é aumentar o lucro, e como é uma plataforma de *streaming*, a fidelização dos clientes é bastante importante de modo a alcançar tal fim;
   - A situação atual desta plataforma é já bastante desenvolvida, sendo que já tem um excelente sistema de recomendação dependendo fortemente de várias técnicas de ciência de dados para fornecer recomendações ao utilizador, contudo cria competições para melhor a sua previsão;
   - O objetivo do *data mining* é, através de modelos como classificação aumentar o lucro da empresa;

   (b) ***Data Understanding***
   - Os dados estão presentes em *datasets*, referentes a ficheiros de dados de qualificação e predição;
   - Os dados recolhidos são o identificador do filme, identificadores dos cliente e datas de classificação;
   - A informação está acessível, é legível e sem erros, uma vez que foi um concurso de forma a melhorar o algoritmo de recomendação.

2. ***Que tipo de benefícios espera retirar da aplicação de Data Mining:***

   - Sistema de recomendação baseado no conteúdo:
     - Neste sistema de recomendação, o conhecimento prévio dos produtos e as informações do cliente são levados em consideração. Com base no conteúdo que os utilizadores visualizaram na Netflix, é fornecido sugestões semelhantes. Por exemplo, se foi assistido um filme que tem como género de ficção científica, com *data mining* é sugerido filmes semelhantes que tenham o mesmo género.
   - Sistema de recomendação de filtragem colaborativa:
     - Ao contrário do outro sistema, a filtragem colaborativa fornece recomendações com base nos perfis semelhantes dos utilizadores, partindo do pressuposto básico de que o que os usuários gostaram no passado também gostarão no futuro. Por exemplo, se uma pessoa A assiste aos géneros de crime, ficção científica e suspense e B assiste aos géneros de ficção científica, suspense e ação, A também gostará de ação e B gostará do género crime.
   - Decidir quais os filmes para adicionar à biblioteca:
     - Com um preço básico relativamente baixo por mês por membro, a Netflix não pode se dar ao luxo de adicionar todos os sucessos de bilheteira. Eles precisam ser inteligentes sobre suas decisões e tirar o máximo proveito de *data mining*. Ser económico e fidelizar os utilizadores é uma habilidade essencial para o sucesso da Netflix. Uma outra tática utilizada é estudar sites ilegais para ajudá-los a decidir qual conteúdo comprar.

3 Conclusão

O presente relatório descreveu, de forma sucinta, o trabalho de introdução à metodologia CRISP-DM.

Após a realização deste trabalho, compreendi as diferentes etapas da metodologia CRISP-DM bem como foi colocado em prática as primeiras etapas da metodologia (Business Evaluation, Data Evaluation).

Por fim, espero que os conhecimentos obtidos e consolidados sejam de enorme utilidade tendo uma perspetiva futura.

Referências

[1] How Netflix Uses Analytics To Select Movies, Create Content, and Make Multimillion Dollar Decisions. 2020. URL: https://neilpatel.com/blog/how-netflix-uses-analytics/ (acedido em 26 de Fevereiro de 2021).

[2] Netflix Business Model (2020) — How does Netflix make money. 2020. URL: https://bstrategyhub.com/netflix-business-model-how-does-netflix-make-money/ (acedido em 26 de Fevereiro de 2021).

[3] Data Science at Netflix – A Must Read Case Study for Aspiring Data Scientists. 2019. URL: https://data-flair.training/blogs/data-science-at-netflix/ (acedido em 26 de Fevereiro de 2021).

[4] Netflix Prize data : Dataset from Netflix's competition to improve their reccommendation algorithm. 2020. URL: https://www.kaggle.com/netflix-inc/netflix-prize-data (acedido em 26 de Fevereiro de 2021).