DSRPT 21 - 2TBD

Tema: Estrelas do amanhã - Qual a probabilidade de um filme ser um sucesso?

Contextualização

O filme "Quebrando a Banca" (21) de 2008 (https://youtu.be/HBYtlT2PM50) mostra um grupo de alunos que, todo fim de semana, parte para Las Vegas com identidades falsas e com o objetivo de ganhar muito dinheiro. O grupo é liderado por um professor de matemática e gênio em estatística, com quem consegue montar um código infalível. Contando cartas e usando um complexo sistema de sinais, eles conseguem quebrar diversos cassinos.

E se usássemos essa mesma habilidade em estatística para trabalharmos com a indústria bilionária do cinema? Só em 2015, os EUA e o Canadá arrecadaram US$ 11,1 bilhões, em bilheteria.

O problema dessa indústria é que sempre existe um fator de incerteza quanto ao retorno financeiro de um filme. Sucessos e fracassos de bilheteria aparecem todos os anos. Por exemplo, o filme "Super Troopers" custou US$ 3 mi e teve bilheteria de US$ 18.5 mi ao passo que "Evan Almighty" custou US$ 175 mi e teve bilheteria de US$ 100 mi.

Segundo o IMDB (empresa da Amazon), entre os filmes produzidos entre 2000 e 2010 nos EUA, apenas 36% tiveram uma receita de bilheteria superior ao orçamento de produção e, do ponto de vista dos investidores, a lucratividade significa sucesso.

Vocês foram contratados por uma empresa de análise de dados para desenvolver uma solução que tente prever o sucesso de um filme de forma automatizada para melhor apoiar as decisões dos investidores dos filmes. Como vocês trabalham projetos com a metodologia ágil Scrum, o primeiro passo será vocês confirmarem os recursos que a solução vai oferecer (requisitos) e documentarem esse escopo em forma de Backlog de Produto.

Produza um vídeo de até **3 minutos** apresentando a proposta e solução desenvolvida em formato de Pitch.

Metas de entrega contratas

As entregas devem ser realizadas na área de Tarefas da sala do DSRPT21.

- Não serão aceitas entregas em outras áreas!
- Não serão aceitas entregas com atraso.

ENTREGA 1 - até 13/11

- Definição dos requisitos contemplados no projeto em forma de Histórias de Usuário no Azure Boards - deixe o seu projeto como PUBLICO e envie o link como resposta dessa atividade [Atende à disciplina de DB Projects & Operations]
- De posse das estruturas de dados e a partir das informações cadastradas, é possível definir perguntas executivas feitas pelo negócio. Descreva as mais importantes que

possam contribuir para a tomada de decisão e que atendam a necessidade dos gestores. [Atende à disciplina de Datawarehousing]**.**

- Obter os datasets que melhor podem se adequar a proposta. No site www.boxofficemojo.com (ver links úteis para alguns exemplos) temos alguns dados abertos e compilados contendo bilheteria, lucratividade e avaliação (score do IMDB, por exemplo). Busque combinar esses dados com o dataset de movie reviews do IMDB. [Atende à disciplina Enterprize Analytics e Big Data]. Explique qual será sua estratégia para carregar esse volume de dados. Definir quais dados devem ser criptografados/mascarados visando a privacidade dos dados. [Atende à disciplina de Backup e Segurança de Dados] e, se for o caso, como esses dados serão armazenados e administrados [Atende à disciplina de Administração de Banco de Dados e NoSQL].
- Descreva como dados textuais relacionados ao dataset selecionado poderiam auxiliar nessa tarefa de predição de sucesso de filmes. Não é obrigatório incluí-los e integrá-los ao dataset selecionado, mas se o fizer, certamente enriquecerá seu modelo com essa informação [Atende à disciplina de MPP e IA].
- Definir a infraestrutura (componentes) que será utilizada para o processamento (extração, limpeza, carregamento) e armazenamento dos dados (SQL, NOSQL). [Atende à disciplina de Administração de Banco de Dados e NoSQL].
- Após encontrar os dados que melhor possam estar adequados à proposta da equipe, realizar uma análise exploratória dos dados (a equipe pode utilizar tanto R ou Python para as análises), incluindo a parte textual, se existir. Entregar arquivo de código (.R ou .ypnb) e arquivo (.pptx com resumo das análises). [Atende às disciplinas Enterprize Analytics e Big Data e de MPP e IA].

ENTREGA 2 - até 20/11

- Aplicar algum modelo preditivo (Regressão linear, logística, árvore de decisão, random forest, xgboost ou redes neurais) para realizar previsões a respeito dos indicadores de sucesso do filme (lucratividade, bilheteria, etc...). Entregar arquivo de código (.R ou .ypnb) e arquivo (.pptx com resumo das análises). [Atende às disciplinas Enterprize Analytics e Big Data e à MPP e IA].
- Após a definição das perguntas executivas feitas pelo negócio, crie o modelo de dados dimensional ou outra estrutura de persistência de dados, que irá apoiar as respostas para essas perguntas. [Atende `a disciplina de Datawarehousing].
- Definir a necessidade de alta disponibilidade e escalabilidade prevendo o crescimento da aplicação. [Atende à disciplina de Administração de Banco de Dados e NoSQL].
- Prever uma arquitetura de contingenciamento caso ocorra, por exemplo, falha de hardware em um servidor. [Atende à disciplina de Backup e Segurança de Dados]
- Link de acesso ao vídeo de Pitch (de até 3 min) do seu projeto, publicado no Youtube.

Links relacionados

https://ai.stanford.edu/~amaas/data/sentiment/

https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/notebooks

https://www.cs.cornell.edu/people/pabo/movie-review-data/

https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

https://www.boxofficemojo.com/

https://data.world/eliasdabbas/boxofficemojo-alltime-domestic-data#

https://teses.usp.br/teses/disponiveis/12/12139/tde-13122017-153711/publico/CorrigidoRafaela.pdf

https://data.world/crowdflower/blockbuster-database#

https://data.world/promptcloud/imdb-data-from-2006-to-2016#