dnc

data® expert.

Data Science & Machine Learning

SUPER CHALLENGE

POWERED BY UELLO

>

DESAFIO

Entender Possíveis atrasos de pedidos

Em prol de melhorar a experiência e fidelizar seus clientes a UELLO coletou em sua base de dados os pedidos já entregues e solicitou ao time de ciências de dados para construir um modelo que os auxiliem em entender os atrasos e criar uma hipótese para evitá-los no futuro.

Dentro dessa solicitação a empresa também solicitou que fosse gerada uma coluna binária com as colunas que contém a data de prazo de entrega e a data real da entrega.

O gestor da área encontrou dois possíveis caminhos para gerar um modelo de ML que irá auxiliar na análise:

- Criar um previsor de atraso;
- Criar um classificador de pares (origem e destino);

Agora o seu time deve executar um dos caminhos propostos e criar uma explicação do raciocínio utilizado em formato de uma apresentação.

dnc

DESAFIO

Entrega e correção

No final devem ser entregues :

- Uma apresentação contendo o raciocínio utilizado para o entendimento dos dados, seleção do modelo e definição de hipótese para auxiliar na redução dos atrasos.
- Um colab contendo a aplicação de um modelo de ML.
- O colab deve conter o calculo do f1 score do modelo utilizado.

Correção:

A correção será feita pelo time de ciências de dados da Uello e os resultados serão divulgados no Slack da turma DEx02 assim que finalizadas.

dnc

DICA

Escolhendo seu Modelo

A escolha do modelo varia conforme o problema de negócio a ser resolvido, as características do processo e a disponibilidade da base de dados. Primeiramente é importante classificar o problema em questão em Supervisionado e Não-supervisionado.

Quando um determinado problema dispõe de resultados pré-definidos e uma base histórica que correlaciona as variáveis de entrada e saída, usamos **modelos de aprendizagem supervisionada**. Quando a criação do modelo parte de uma base que não possui uma referência para balizar e avaliar os testes, ou seja, temos inputs bem definidos, entretanto não temos outputs correlacionáveis com o processo, usamos **modelos de aprendizagem não-supervisionada**.

A partir da escolha do tipo de aprendizagem, deve-se definir qual tipo de modelo será usado. Dentre os modelos supervisionados temos dois tipos:

1. **Regressão:** modelo utilizado em processos que possuem uma saída contínua, ou seja, estes modelos visam encontrar um output numérico. Um sistema preditivo de aluguéis utiliza um sistema supervisionado de regressão, por exemplo.
2. **Classificação:** modelo utilizado através da ordenação de itens por um identificador, ou seja, através de características similares segmentam-se as variáveis de saída em classes. Um exemplo pode ser um modelo capaz de determinar qual a doença uma pessoa possui a partir de informações como idade, sexo, pressão arterial, oxigenação, entre outras. Neste exemplo, o modelo classifica os dados apresentados em alguma doença, comparando o histórico.

DICA

Escolhendo seu Modelo

Considerando os modelos não-supervisionados (modelos que não possuem referência estabelecida) temos outras duas possibilidades:

1. **Recomendação:** neste caso, o processo visa prever o interesse em determinado item. Isso é feito a partir de uma matriz de correlação, que pode ser criada pela similaridade do item, similaridade do usuário (aquele que escolhe o item) ou pelas ações executadas pelo usuário. O sistema de recomendação de filmes da Netflix é um exemplo de sistema.
2. **Clusterização:** na clusterização ocorre um agrupamento de dados através da similaridade. Essa classificação pode ser hierárquica, com mais de um nível de segmentação ou não hierárquica, em que a partição inicial é única. Um exemplo é a clusterização de consumidores com comportamentos similares de consumo a partir de uma base de dados.

DATASET

Base fornecida

A base contém algumas entregas realizadas pelo UELLO. Nesses dados conseguimos encontrar a data do pedido, a previsão de entrega e a entrega real.

ordens_case_DNC.csv: Entregas realizadas pela Uello

Grande parte do trabalho de um cientista de dados é compreender a base fornecida e linkar isso com o objetivo daquela análise. Portanto, use e abuse de **visualizações gráficas**, **análises de correlação**, **teste de hipótese** e não se esqueçam de guardar esses passos, para depois criarem um bom **storytelling**!

dnc

DATASET

Base fornecida - Dados

Id - Primary key - Int

Price - Valor da entrega - Float

Order_date - Data do pedido - Data

Due_date - Data de entrega Esperada - Data

Opt_date - Data em que o pedido sai para o destino - Data

Delivery_date - Data de entrega - Data

Cidade_origem_id - Cidade origem do pedido - int

Regiao_origem_id - Região origem do pedido - int

Cidade_destino_id - Cidade destino do pedido - int

Regiao_destino_id - Região destino do pedido - int

Peso - Peso do pacote - float

dnc

CRONOGRAMA

Recomendamos que iniciem o desenvolvimento do modelo com antecedência.

Evento	
Durante a semana	**Tempo**
Exploração de Dados (*Quarta*)	1:00:00
Definição do Tipo de Modelo (*Quinta*)	1:00:00
Tratamento dos dados (*Sexta*)	1:00:00

Sábado	**Horário**
Abertura	9:00
Palestra - Compreensão do Negócio	9:15
Maratona	10:15
Almoço	12:15
Maratona	13:45
Criar documentação e apresentação	17:25
Entrega final	18:15

dnc

dnc

data® expert.

Data Science & Machine Learning

#HARDWORK

Esperamos que o material tenha sido útil para você!
Foco nos estudos!