

1. Mapa de metodologia

Essa estrutura analítica de resolução de problemas ajuda a resolver, de forma sistemática, um problema de negócio. No entanto, isso não nos ajuda a determinar a melhor metodologia a ser utilizada.

É o mapa metodológico que nos ajudará a decidir a metodologia utilizada para resolver um problema de negócios, e continuará sendo utilizado conforme avançarmos na estrutura de resolução de problemas.

Business Problem					
Predict Outcome				Data Analysis	
Data Rich			Data Poor	Geospatial	
Numeric		Classification		A/B Testing	Segmentation
Continuous	Time Based	Binary	Non Binary	Aggregation	
Linear Regression Decision Tree Forest Model Boosted Model	ARIMA ETS	Logistic Regression Decision Tree	Forest Model Boosted Model	Descriptive	

2. Análise não preditiva

Nós dividimos a análise não preditiva em quatro categorias: Geoespacial, Segmentação, Agregação e Descritivo. Vamos falar sobre cada um.

Análise geoespacial

Este tipo de análise usa dados baseados em localização para ajudar a conduzir suas conclusões. Os exemplos incluem identificar clientes por uma região geográfica,

calcular a distância de locais de armazenamento e criar uma área de comércio com base nas localizações dos clientes.

Análise de segmentação

Segmentação é o processo de agrupar dados. Os grupos podem ser simples como clientes que compraram itens diferentes, ou podemos ter técnicas de segmentação mais complexas, nas quais você identifica lojas que são semelhantes com base na demografia de seus clientes.

Análise de agregação

Esta metodologia é muito usada na análise de dados e calcula um valor através de um grupo ou dimensão. Por exemplo, você pode agregar dados de vendas para um vendedor específico em um determinado intervalo de meses, somando todas as vendas fechadas em cada mês. Em seguida, você pode agregar dimensões como vendas mensais por região. A agregação é feita através de relatórios que esmiuçam os dados e ajudam os gestores a tomarem decisões baseados em análise de desempenho.

Análise descritiva

A estatística descritiva fornece resumos simples de uma amostra de dados, como: calcular a média das notas dos alunos que estejam se candidatando a um determinado curso ou a média de rebatidas de um jogador de beisebol profissional. Em nosso cenário de fornecimento de energia elétrica, podemos usar estatísticas descritivas para calcular a temperatura média por hora, por dia ou por data. Algumas das estatísticas descritivas comumente utilizadas são média, mediana, modo, desvio padrão e intervalo interquartil.

Problemas comerciais preditivos

Rico em dados x pobre em dados

Você tem dados sobre o que está tentando prever? Em caso positivo, considere o projeto rico em dados. Caso contrário, só lhe restará o caminho de um projeto pobre em dados. Veja abaixo um exemplo pobre em dados.

Testes A/B

Se não houver dados suficientes utilizáveis para resolver o problema, então precisamos configurar uma experiência para nos ajudar a obter os dados de que precisamos. Um experimento em um contexto de negócios é geralmente referido como um teste A/B.

Análises preditivas numéricas e não numéricas

Assumindo que temos dados suficientes para prosseguir com a análise, a nossa próxima decisão é olhar para o resultado que estamos tentando prever e determinar se é uma saída numérica ou não-numérica.

Modelos de regressão

Os resultados numéricos são aqueles em que o resultado é simplesmente um número. Prever a demanda por eletricidade ou a temperatura por hora são resultados numéricos. Modelos que predizem dados numéricos são chamados de modelos de regressão.

Modelos de classificação

Os resultados não-numéricos são aqueles em que estamos tentando prever a categoria na qual um caso (por exemplo, cliente) cai, como se um cliente pagará em tempo, pagará tarde ou o padrão em um pagamento. Outro exemplo é se um dispositivo eletrônico falhará antes de 1000 horas ou não. Modelos que predizem dados não numéricos são chamados de modelos de classificação.

Exemplos de resultados numéricos e não numéricos

Usaremos estes exemplos ao longo do resto da lição.

Departamento de produção do fabricante de triciclos

Para o nosso primeiro exemplo, imagine que um fabricante quer usar dados históricos de produção para saber quantos triciclos precisará produzir nos próximos seis meses para atender a demanda esperada. Uma vez que o resultado que o fabricante quer

prever é um número, então a variável de destino é numérica. Portanto, eles usariam um modelo numérico ou de regressão para resolver este problema.

Departamento de Marketing da Pizza Hot & Fresh

Para nosso segundo exemplo, a Hot & Fresh Pizza quer usar dados de vendas de suas lojas existentes e respectivos dados demográficos em torno dessas lojas para prever quantas pizzas venderão em sua nova loja. Uma vez que o resultado que Hot & Fresh Pizza está tentando prever é o número de pizzas, então a variável de destino é numérica e eles usariam um modelo numérico ou de regressão para resolver este problema.

Departamento de Gestão de risco de um banco

E para o nosso terceiro exemplo, um banco quer usar dados históricos de seus clientes para prever se um novo cliente irá negligenciar um empréstimo, sempre pagar a tempo, ou às vezes pagar. Uma vez que o resultado que o banco está tentando prever é uma categoria que o novo cliente vai cair, eles usariam não-numérico ou modelo de classificação para resolver este problema.

Variáveis resultado/alvo

As variáveis alvo representam o resultado que estamos tentando prever. Para selecionar o modelo preditivo correto, primeiro determinamos se a variável alvo é numérica ou não-numérica. O tipo de variáveis de destino numéricas ou não-numéricas nos ajudará a selecionar qual modelo é apropriado. Vamos começar com variáveis numéricas.

Tipos de variáveis numéricas

Os três tipos mais comuns de variáveis numéricas são contínuas, baseados no tempo e contagem.

Contínua

Uma variável contínua é aquela que pode assumir todos os valores em um intervalo. Por exemplo, sua altura pode ser medida até muitas casas decimais. Nós não crescemos em intervalos de centímetros.

Baseada em tempo

Uma variável numérica baseada no tempo é aquela em que você está tentando prever o que vai acontecer ao longo do tempo. Isso geralmente está relacionado à previsão.

Contagem

As variáveis de contagem são números que são discretos, inteiros positivos. Eles são chamados de números de contagem, porque eles são usados para analisar variáveis que você pode contar. Como a modelagem desses tipos de variáveis não é comum nos negócios, não estaremos cobrindo esse tópico neste curso.

Variáveis não numéricas

Uma variável não numérica é muitas vezes chamada categórica porque os valores da variável assumem um número discreto de possíveis valores ou categorias. Exemplos incluem se um dispositivo eletrônico falhará antes de 1000 horas ou não; se um cliente pagará em tempo, pagará tarde, ou haverá um padrão em um pagamento, ou se uma loja é classificada como grande, média ou pequena.

Classificação de modelos: binário e não-binário

Quando modelamos variáveis categóricas, o número de resultados possíveis é um fator importante. Se houver apenas dois resultados categóricos possíveis, como sim ou não, ou verdadeiro ou falso, então a variável pode ser descrita como binária.

Se houver mais de dois resultados categóricos possíveis, como pequenos, médios ou grandes, ou pagar no prazo, pagar tarde ou padrão em um pagamento, a variável pode ser descrita como não-binária. A importante retirada desta lição é a capacidade de determinar se você deve usar um modelo de classificação e se deve ser um modelo binário ou um modelo não-binário. Ben Burkholder vai liderar um curso focado em modelos de classificação e vai entrar em detalhes sobre esses tipos de modelos.