Creating Customer Segments

REVISÃO

REVISÃO DE CÓDIGO

HISTORY

Meets Specifications

O seu trabalho está muito bom. Parabéns!
Continue assim para manter sua trajetória excepcional.
😄😄😄

Boa sorte em seus próximos projetos!

Se quiser me adicionar no Linkedin (Rafael Buck) fique à vontade.

Exploração dos dados

✓

Três amostras diferentes dos dados são escolhidas, e o que elas representam é proposto com base na descrição estatística dos dados.

Excelente descrição dos exemplos.

✓

A pontuação do atributo removido foi corretamente calculada. A resposta justifica se o atributo removido é relevante.

conclusões estão ótimas. Realmente os atributos de menor pontuação são mais relevantes já que não pode ser previstos. Os atributos com maiores pontuações são facilmente previstos pelos outros atributos,

portanto não estão trazendo nenhuma informação nova para a análise.
Esse artigo em inglês cita a alta correlação para remover parâmetros antes de realizar as predições.

✓

Atributos correlacionados são corretamente identificados e comparados com o atributo previsto. A distribuição dos dados para esses atributos é discutida.

Aqui a análise foi perfeita. Consegue ver que isso reforça o mencionado no exercício anterior? O `Freso` não tem relação com nada, enquanto `Grocery` e `Detergents_Paper` são praticamente redundantes.
Importante notar que os dados estão distorcidos positivamente, mais concentrados na origem, por isso o pré-processamento (a seguir) é feito no dado. A imagem abaixo mostra exemplos de distorção:

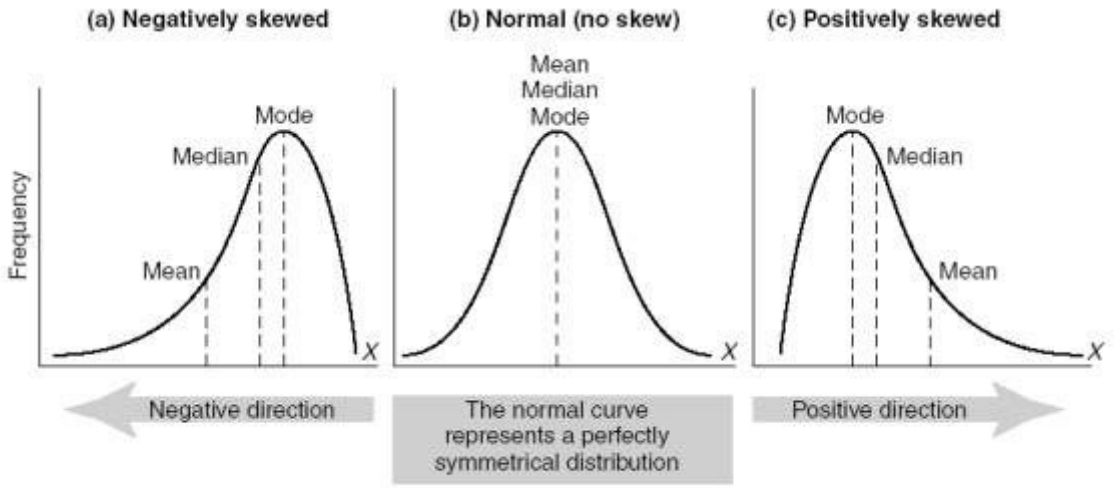

FIGURE 15.6 Examples of normal and skewed distributions

Referência: https://www.quora.com/How-do-outliers-affect-normal-distribution-in-statistics

Pré-processamento dos dados

✓

Os valores aberrantes extremos são identificados, e discute-se se eles deveriam ser removidos. A decisão de remover quaisquer dados é corretamente justificada.

Análise muito boa, verificando quais são os outliers. A explicação sobre remoção dos valores está ótima, não existe uma verdade universal aqui.
Acho esse artigo sobre o tópico interessante. E esse artigo discute sobre remover ou não outliers.

✓

O código de dimensionamento de atributos, tanto para os dados como para as amostras, foi corretamente implementado.

Excelente! Uma alternativa seria o uso do Box-Cox ou até do `preprocessing.scale` do Sklearn 😄

Transformação de atributos

✓

A variância explicada total para duas e quatro dimensões dos dados do PCA é corretamente relatada. As primeiras quatro dimensões são interpretadas como uma representação dos gastos do cliente com justificativa.

✓

O código do PCA foi corretamente implementado e aplicado, tanto para os dados dimensionados como para as amostras dimensionadas, no caso bidimensional.

Sua conclusão está ótima. Vale notar a relevância tanto do valor absoluto de cada categoria quanto a oposição de atributos (um atributo negativo, outro positivo).
Esse artigo em inglês tem um ótimo exemplo visual do PCA.
Esse artigo em inglês apresenta uma discussão interessante de como analisar cada dimensão.

Clustering

✓

Os algoritmos GMM e k-means são comparados em detalhes. A escolha do aluno é justificada com base nas características do algoritmo e dos dados.

Excelente explicação de ambos os algoritmos. Existe um post em inglês que trata disso. Ou essa apresentação pode te ajudar.
A justificativa de escolha também está boa.

✓

Amostras dos dados são corretamente relacionadas aos segmentos da clientela, e o grupo a que pertence cada ponto da amostra é discutido.

✓

Diversas pontuações são corretamente relatadas, e o número ótimo de grupos é escolhido com base na melhor. A visualização escolhida mostra o número ótimo de grupos baseado no algoritmo de clustering escolhido.

Os grupos estão propostos de forma correta com os dados justificando a escolha.

✓

Os grupos representados por cada segmento da clientela são propostos com base na descrição estatística do conjunto de dados. Os códigos de transformação e dimensionamento inversos foi corretamente implementado e aplicado para o centro dos grupos.

A análise dos clientes está boa. Essa é a ideia mesmo.

Conclusão

✓

O aluno identifica corretamente como um teste A/B pode ser feito com a clientela após uma mudança no serviço de distribuição.

Análise excelente.

✓

O aluno discute e justifica como os dados de clustering podem ser usados em um modelo de aprendizagem supervisionada para fazer novas estimativas.

A discussão de como usar os dados para aprendizagem supervisionada está ótima. Com essa segmentação ganhamos um dado para entender os clientes e consequentemente mais informações.

✓

Os segmentos da clientela e os dados em `Channel` são comparados. Os segmentos identificados pelos dados de `Channel` são discutidos, inclusive se essa representação é consistente com resultados anteriores.

Perfeita a análise de como tratar um teste A/B sabendo de ambos os segmentos de clientes. Essa análise é bem importante, podendo ser crítica para o negócio.
O primordial aqui é notar que se você testar somente um cluster, pode obter um resultado totalmente diferente no outro. Em um teste A/B seria importante entender cada cluster como um tipo de cliente distinto, sem misturar, para assim ter um conhecimento maior dos seus clientes.
Esse artigo do Netflix é ótimo sobre testes A/B e esse discussão do Quora apresenta os problemas que você pode encontrar nesses testes.

BAIXAR PROJETO

RETORNAR

Avalie esta revisão