UDACITY

PROJETO

Creating Customer Segments

Uma parte do Machine Learning Engineer Nanodegree Program

REVISÃO DO PROJETO

REVISÃO DE CÓDIGO

COMENTÁRIOS

COMPARTILHE SUA REALIZAÇÃO!

Meets Specifications

Parabéns por ter atendido a todos os requisitos deste projeto! Acrescentei alguns comentários e sugestões abaixo, espero que sejam úteis. Bom trabalho na continuação do curso!

Exploração dos Dados

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

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

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

Excelente

Bom trabalho descrevendo a distribuição dos atributos, que de fato apresentam todos uma pronunciada assimetria positiva - indicando que alguns clientes gastam muito mais do que a média para cada tipo de produto.

Pré-processamento dos Dados

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

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

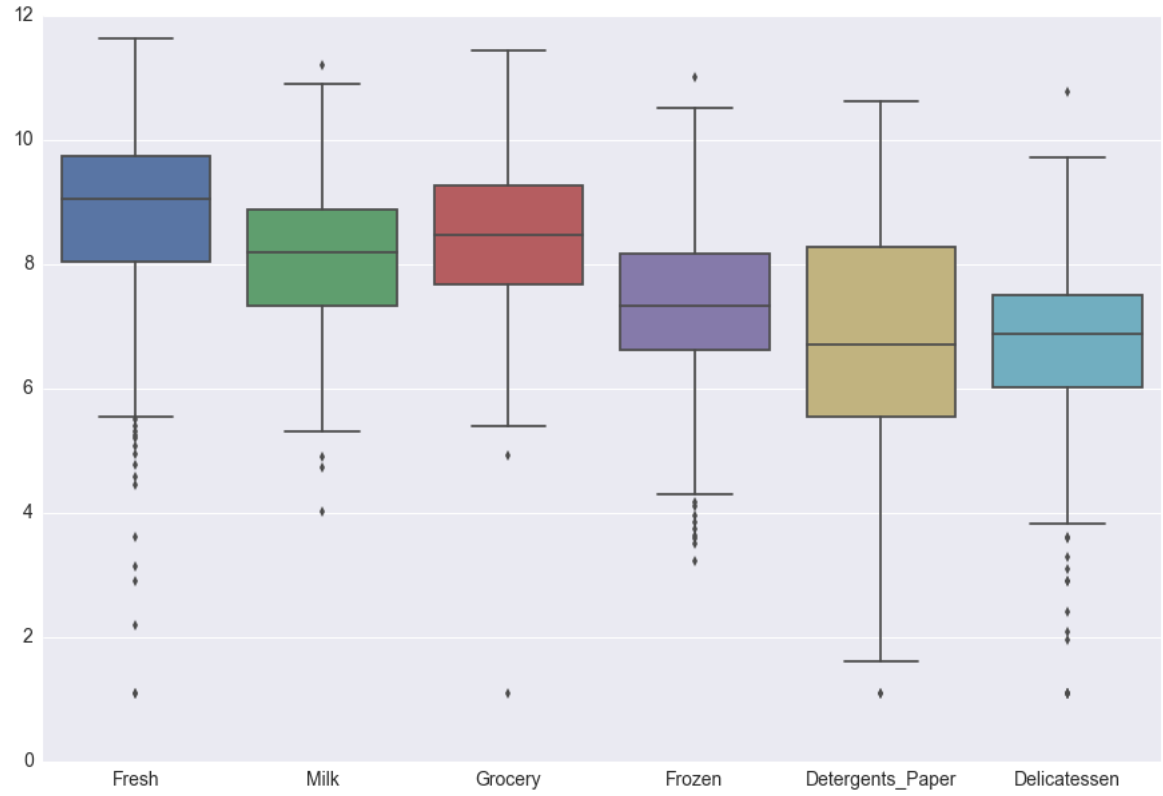
Sugestão

> Poderia me sugerir mais opções de visualizações de dados para esse modelo?

Não sei bem a que modelo você está se referindo, mas uma forma de analisar *outliers* é usando boxplots:

```
import seaborn as sns
sns.boxplot(data=log_data)
```

O código acima gera a seguinte figura:



Olhando para esse gráfico, quais *outliers* você consideraria remover do conjunto de dados, e por quê?

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.

Excelente

Ótimo trabalho analisando cada um dos quatro primeiros componentes principais e indicando qual tipo de cliente deve apresentar valores significativos para cada um deles. Parabéns!

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

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

> Atravez das probabilidades e distribuições gausianas os pontos de dados não precisam de pertencer necessariamente a algum dos grupos de clusters, os pontos de dados podem ser atribuidos a diversos clusters ao mesmo tempo. Esse tipo de algoritimo permite a previsão de probabilidade de eventos.

Bom acréscimo à sua resposta, embora seja mais apropriado dizer que um GMM fornece a *likelihood*, ou "verossimilhança", que determinada observação pertença a cada *cluster*, dando-nos uma medida da **incerteza** relativa à classificação de cada observação. Veja uma discussão sobre isso aqui, por exemplo.

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

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.

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

Conclusão

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.

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

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

BAIXAR PROJETO

RETORNAR

Avalie esta revisão

FAQ do Estudante