UNIVERSIDADE FEDERAL DE MINAS GERAIS - UFMG
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
CURSO DE SISTEMAS DE INFORMAÇÃO

GILLIARD GABRIEL RODRIGUES

APRENDIZADO DE MÁQUINA NÃO SUPERVISIONADO APLICADO À SEGMENTAÇÃO DE CLIENTES

Belo Horizonte
2022

GILLIARD GABRIEL RODRIGUES

APRENDIZADO DE MÁQUINA NÃO SUPERVISIONADO APLICADO À SEGMENTAÇÃO DE CLIENTES

Trabalho apresentado para a Disciplina Mineração de Dados, pelo Curso de Sistemas de Informação da Universidade Federal de Minas Gerais, ministrada pelo Prof. Wagner Meira Júnior.

Belo Horizonte

2022

SUMÁRIO

1 INTRODUÇÃO **3**
2 MOTIVAÇÃO **4**
3 OBJETIVO **5**
3.1 Geral **5**
3.2 Específicos 5
4 METODOLOGIA **6**
5 DESENVOLVIMENTO **7**
5.1 Entendimento do negócio **7**
5.2 Entendimento dos dados **7**
5.3 Preparação dos dados **8**
5.4 Modelagem **8**
6 RESULTADOS/CONCLUSÕES **9**
6.1 Avaliação e Implantação **9**
REFERÊNCIAS **10**

1 INTRODUÇÃO

A divulgação de produtos e o alcance de clientes são essenciais para que as empresas consigam aumentar seus números de vendas. No entanto, não se trata apenas de divulgar tudo para todos, conhecer quem são seus clientes e quais os seus gostos é fundamental para saber a quem direcionar cada tipo de anúncio e produto.

A segmentação de clientes se trata de uma subdivisão de um mercado em grupos distintos de clientes que compartilham características semelhantes e pode ser a chave para converter mais clientes e ter vantagem frente à concorrência.

Muitos produtos são lançados e não atingem suas metas de receita justamente por não serem direcionados ao público certo, uma vez que as empresas não conseguem entender quem é o seu público e criar campanhas personalizadas e boas estratégias de aquisição de clientes. Nesse contexto, o presente trabalho irá apresentar um processo moderno e valioso para descobrir grupos de clientes de um negócio.

2 MOTIVAÇÃO

A segmentação de clientes traz inúmeros benefícios para os negócios, dentre eles pode-se citar:

- Aprimorar o ad targeting;
- Aprimorar o desenvolvimento da solução;
- Atrair e converter leads mais qualificados;
- Criar mais afinidade com o consumidor;
- Desenvolver estratégias de retenção mais eficientes;
- Desenvolver melhor a comunicação da equipe de Marketing;
- Diferenciar sua marca da concorrência;
- Identificar oportunidades em mercados nichados;
- Identificar táticas de Marketing mais eficientes;
- Manter o foco no público-alvo correto;
- Oferecer uma melhor Experiência do Cliente;
- Tornar o orçamento mais eficiente.

Dada uma base de dados contendo informações de clientes, sejam elas de sociais-demográficas ou de compras feitas, um algoritmo de aprendizado de máquina não supervisionado pode ser aplicado a fim de descobrir grupos de clientes e tal contribuição seria valiosa para os donos de negócios.

3 OBJETIVO

3.1 Geral

O objetivo aqui é analisar as características dos clientes de um supermercado durante 2 anos e, baseado nessas características, segmentar esses clientes em grupos de acordo com suas semelhanças a fim de descobrir como esses consumidores se comportam e qual a melhor forma de alcançar cada segmento. O critério de sucesso será a capacidade de segmentar os clientes de forma suficiente para que insights novos sejam gerados.

3.2 Específicos

As tarefas específicas podem ser divididas em:

- Extração dos dados, que virão de uma base em formato *.csv*, retirada do *Kaggle*;
- Limpeza dos dados, pois a base pode conter informações faltantes ou dados desnecessários para o nosso objetivo;
- Análise exploratória dos dados, a fim de obter *insights* sobre as diversas informações disponíveis por cliente e decidir quais utilizar no modelo.
- Preparação dos dados, transformando e reescalando os dados para aplicar o algoritmo de aprendizado de máquina não supervisionado.
- Aplicação do algoritmo de agrupamento *K-Means*;
- Análise dos *clusters* gerados, relacionando-os às várias *features* da base de dados a fim de descobrir o perfil de cada segmento de clientes.

4 METODOLOGIA

A metodologia foi inspirada no CRISP-DM, ou seja, dividida em: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação.

Uma parte da aplicação dessa metodologia pode ser vista através do link para o *Google Colab*, que apresenta o entendimento dos dados, a preparação dos dados, a modelagem e a avaliação, com os *scripts* já rodados e os resultados já disponíveis.

5 DESENVOLVIMENTO

5.1 Entendimento do negócio

Conforme citado anteriormente, um algoritmo que, a partir de vários dados sobre clientes, conseguisse encontrar padrões a ponto de conseguir separar os clientes em grupos de acordo com suas características seria de grande valia para alcançar uma segmentação de clientes e o objetivo principal deste trabalho é demonstrar como alcançar tal feito. Como recurso disponível, temos uma base de dados, retirada do Kaggle, contendo várias informações de clientes de um supermercado, de onde é possível extrair desde informações sociodemográficas até informações específicas do comportamento de compra nesse supermercado, como a quantidade comprada de certos produtos, o local utilizado para efetuar a compra (website ou loja física, por exemplo) e se fez uso de promoções. O *dataset* possui 29 atributos, mas o número de *features* a ser considerado será definido no decorrer da análise exploratória, já que talvez valha a pena aumentar a granularidade dos dados. Antes de prosseguir para uma análise exploratória dos dados ou a segmentação propriamente dita, os dados precisam passar por uma limpeza. Os riscos aqui estão associados ao resultado da segmentação não trazer informações relevantes.

5.2 Entendimento dos dados

A base de dados bruta pode ser encontrada no seguinte link e, conforme citado anteriormente, está em formato *csv*. A base contém informações de mais de 2000 clientes de um supermercado e sobre cada cliente existem essas informações: identificador exclusivo do cliente, ano de nascimento, nível de educação, estado civil, renda familiar anual, número de crianças em casa, número de adolescentes em casa, data do cadastro do cliente na empresa, número de dias desde a última compra, se o cliente reclamou nos últimos 2 anos, valor gasto em vinho, frutas, carne, pescado, e doce nos últimos 2 anos, número de compras feitas com desconto, informações sobre aceitação em campanhas ocorridas, número de compras realizadas através do site da empresa, diretamente nas lojas, utilizando catálogo e número de visitas ao site no último mês. Dessas informações, algumas foram retiradas e outras foram derivadas a partir de *features* já existentes, são elas:

a idade, se mora sozinho ou com parceiro, se possui filhos, o número de crianças em casa, a quantia total gasta no estabelecimento, se já acessou o website da empresa, o total de campanhas aceitas e o tempo de cliente. Várias visualizações foram construídas a partir das *features* existentes e alguns *insights* foram tirados e podem ser vistos no link do notebook utilizado.

.

5.3 Preparação dos dados

A fim de preparar os dados para aplicação do modelo, os dados desnecessários foram removidos, os dados a serem utilizados foram convertidos para tipos numéricos e foram reescalados.

5.4 Modelagem

Dado que mesmo após a remoção de algumas *features*, a base ainda permaneceu com muitas colunas (18), viu-se a necessidade de fazer redução de dimensionalidade através da técnica PCA a fim de obter os componentes principais baseado nas *features* mais importantes.

A fim de escolher o número de *clusters* a ser formado, foi utilizado o *Elbow method*, que considera a soma dos quadrados das distâncias de cada ponto para o centro do seu cluster, testando vários valores de *k* para decidir qual o valor ótimo.

O algoritmo *K-Means* foi escolhido para fazer o agrupamento por sua simplicidade e velocidade e, após sua aplicação, foram gerados 5 *clusters*, que foram relacionados com as diversas *features* presentes na base a fim de descobrir os perfis de cada grupo. Os resultados podem ser encontrados na próxima seção.

6 RESULTADOS/CONCLUSÕES

6.1 Avaliação e Implantação

Conforme citado anteriormente, após aplicar o *K-Means* e formar os *clusters*, foram construídas várias visualizações relacionando os *clusters* formados com as diversas *features* a fim de descobrir as características comuns de cada grupo e, assim, montar o perfil de cada segmento.

O resultado obtido, para cada um dos 5 *clusters*, foi:

- **Grupo 1**: clientes sem ou com apenas graduação; com o menor poder aquisitivo; que menos gastaram; relativamente mais jovens e que possuem um ou mais filhos.
- **Grupo 2**: clientes com graduação ou pós; com poder aquisitivo elevado; que mais gastaram; que menos fazem compras com desconto; de todas as idades e que não possuem filhos.
- **Grupo 3**: clientes com possuem graduação ou pós; com poder aquisitivo médio; com gastos medianos; relativamente mais velhos e que possuem um filho.
- **Grupo 4**: clientes com possuem graduação ou pós; com poder aquisitivo baixo; com gastos baixos; na meia idade e que possuem um ou mais filhos.
- **Grupo 5**: clientes com possuem graduação ou pós; com o maior poder aquisitivo; que mais gastaram; que menos fizeram compras com desconto; de todas as idades e que não possuem filhos.

Portanto, foi possível obter resultados interessantes a partir dos dados, fazendo com que o tempo e as técnicas empregadas se mostrassem de valor e serviu para consolidar o entendimento e aumentar a experiência com aprendizado de máquina não supervisionado.

Dada uma base real, o processo construído nesse trabalho pode ser utilizado como molde para a obtenção de uma segmentação de clientes por donos de negócio interessados nessa implantação.

REFERÊNCIAS

Scikit-Learn. K-Means. 2022. Disponível em < https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html >. Acesso em 01 nov. 2022.

Yellow Brick. Elbow Method. 2022. Disponível em < https://www.scikit-yb.org/en/latest/api/cluster/elbow.html />. Acesso em 01 nov. 2022.

CS ACADEMY. *Segmentação de clientes: quais são os tipos, benefícios e exemplos*. Disponível em < https://www.csacademy.com.br/segmentacao-de-clientes-quais-sao-os-tipos-beneficios-e-exemplos>. Acesso em 01 nov. 2022.

KALER, INNA. *So You Have Some Clusters, Now What?* How to Add Value to Your Clusters. Disponível em < https://developer.squareup.com/blog/so-you-have-some-clusters-now-what/>. Acesso em 02 nov. 2022.