Udacity Engenheiro de Machine Learning

Proposta Projeto Capstone

Daniel Vieira Roberto

Background do Domínio

Segmentação de e clientes é uma das tarefas mais importantes em qualquer empresa de Marketing. Utilizando irá influenciará as decisões de marketing e vendas de forma que seja possível oferecer serviços e produtos de forma mais personalizada para obter maior lucratividade no negócio. Conceito de segmentação de marketing foi cunhado por Wendell R. Smith, que em seu artigo “Diferenciação de Produto e Segmentação de Mercado como Estratégias de Marketing Alternativas” observou “muitos exemplos de segmentação” em 1956.

O papel da segmentação de clientes no marketing digital é justamente identificar esses grupos menores e categorizá-los em seguida aprofundar sobre seus interesses e necessidades, a partir disso, você pode criar suas personas, isto é, personagens únicos que representem cada grupo identificado. As personas servirão de base para que que os serviços e produtos sejam desenvolvidos ou direcionados.

Além disso, você pode usar a segmentação para criar campanhas de publicidade e marketing para as personas categorizadas, com isso é possível aumentar a probabilidade de venda ou taxa de conversão de forma otimizada e data-driven.

O comportamento dos consumidores vem mudando rapidamente nos últimos anos, por isso, a segmentação no marketing digital deve utilizada constantemente, até mesmo para identificar novos nichos de negócio, ou nichos que são mais promissores e a previsibilidade desses cenários é um ponto critico a ser resolvido em todas as empresas que prestam serviços e vendem produtos.

Enunciação do problema

No Marketing digital, uma atividade muito comum é a utilização de ferramentas de anúncios que realizam campanhas tanto em sites de busca (Google, Bing, etc) ou em redes sociais (Facebook, Instagram,etc), em resumo, esses anúncios aparecem para as pessoas que estão buscando algum serviço em questão. As plataformas fornecem opções de filtros para que esses anúncios tenham um público-alvo espesifico, a pergunta feita é como as empresas sabem o perfil desses público alvo? Esse é o problema em questão que esse projeto irá ajudar a resolver.

A métrica mais comum para mensurar o desempenho dessas campanhas é a taxa de conversão, as empresas que gerenciam essas campanhas estão sempre estudando formas de otimizar essas taxas de conversões visto que as empresas que provém o serviço de anuncio cobram por cliques, ou seja, caso a

pessoa clique no anúncio é gerado um custo mesmo se não haja a compra o serviço, ou outras palavras, não temos uma conversão.

O objetivo é aumentar a taxa de conversão otimizando os anúncios para o grupo que teria maior probabilidade de serem convertidas (comprarem o serviço ou produto que está sendo ofertado via anúncio).

Conjunto de Dados e Inputs

Os dados usados neste projeto são da campanha publicitária de mídia social de uma organização anônima.

O arquivo conversion_data.csv contém 1143 observações (Linhas) em 11 variáveis (Colunas). Abaixo estão as descrições das variáveis:

- **ad_id:** ID exclusivo para cada anúncio;
- **xyz_campaign_id:** ID associado a cada campanha publicitária empresa XYZ;
- **fb_campaign_id:** ID associado a campanha;
- **age:** Idade da pessoa a quem o anúncio foi mostrado;
- **gender:** Gênero da pessoa a quem o anúncio foi mostrado;
- **interest:** Código que especifica a categoria de qual o interesse da pessoa pertence;
- **Impressions:** Numero de vezes que aunncio foi mostrado;
- **Clicks:** Número de cliques do anuncio;
- **Spent:** Quantidade pago pela empresa XYZ para o Facebook, para mostrar o anúncio;
- **Total conversion:** Número total de pessoas que se interessaram sobre o produto ou serviço depois de ver o anúncio;
- **Approved conversion:** Número total de pessoas que compraram o produto depois de ver o anúncio.

Explanação da solução

Primeiramente sera realizado uma analise exploratória junto com limpeza (caso necessário), ajuste estrutural (caso necessário) podendo haver feature engineering caso seja identificado que as variáveis disponíveis não descrevem suficientemente bem os dados e por fim pré processamentos como escalonagem e transformação de dados categóricos para numéricos.

Em seguida será aplicado PCA para diminuir as dimensões, o número de dimensões sera definido pelo numero que o algoritmo dizer que representa pelo menos 80% da variabilidade dos dados.

Após a redução de dimensionalidade, será aplicado o algoritmo de clusterização que será selecionado por experimentação, será utilizado a métrica de avaliação coeficiente de silhueta para escolher o algoritmo e o número de clusteres, após o número de clusteres definido, será necessário classifica-los de for que faça sentindo para o negócio.

Utilizando os clusteres classificados, será treinando uma regressão linear para projetar qual segmento será o mais promissor utilizando o “Approved conversion” como variável a ser projetada.

Modelo de benchmark

O coeficiente de silhueta varia de -1 a 1, porém não existe um valor benchmark para ser comparado, apenas durante a analise devemos escolher a quantidade de clursters referente ao ponto de inflexão dos score encontrados no coeficiente de silhueta.

Para a regressão, existe uma competição no keggle que teve como objetivo predizer quanto cada cliente iria gastar utilizando a mesma métrica de avaliação em dados que são similares, o ganhador atingiu o score de 0.88140, sendo mais conversador, para uma regressão, um valor satisfatório para uma regressão seria em torno de 0.7.

Métricas de avaliação

Para o modelo de clusterização sera utilizada validação baseada em densidade, coeficiente de silhueta é calculada com a **média** para todos os pontos utilizando a equação:

$$s = \frac{b - a}{\max(a, b)}$$

s = Coeficiente de silhueta do pronto (vai de -1 até 1)

b = Distância média do ponto e dos outros pontos dos outros clusters.

a = Distância média do ponto e dos pontos do mesmo cluster.

Para o modelo de predição sera usado Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{obs,i} - X_{model,i})^2}{n}}$$

X_{obs} = Valor Real

X_{model} = Valor estimado pelo modelo

Design do projeto

O projeto irá seguir os seguintes passos:

1) **Análise exploratória:** Entendimentos mais aprofundados dos dados, tratamentos de qualidade de estrutura dos dados, pré-processamentos dos dados (escalonamento, one hot coding, etc), identificação de outlies e Feature Engineering.
2) **Redução de Dimensões:** Aplicação de algoritmo de PCA com a biblioteca do sckitlearn *sklearn.decomposition.PCA*
3) **Aplicação dos algoritmos de clusterização:** Utilizar a biblioteca sckitlearn, K-Means (*sklearn.cluster.KMeans*) e Mistura Gaussiano (*sklearn.mixture.GayssianMixtur*e) e utlizar a métrica de avaliação coeficiente de silhueta para definir o número de clusteres.
4) **Aplicação regressão linear:** Utilizar a biblioteca sckitlearn LinearRegression (*sklearn.linear_model. LinearRegression*) para cada cluster encontrado para intentificar qual é a projeção de “Approved conversion” para cada segmento.
5) **Conclusão:** Analise final e sugerir plano de ação utilizando os resultados encontrados para diferenciar a campanha de forma otimizada.