Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Alisson Santos Ribeiro e Irlan Wallace dos Santos Mattos

17/11/2024

Resumo

Este relatório aborda a análise e construção de um modelo preditivo para prever o engajamento dos seguidores dos influenciadores do Instagram, utilizando o conjunto de dados “Dados dos principais influenciadores do Instagram (limpos), O objetivo deste estudo foi entender as relações entre as variáveis do dataset e a taxa de engajamento dos influenciadores e entender os fatores que influenciam essa métrica e como pode fornecer insights valiosos sobre o comportamento dos influenciadores e o impacto que eles têm sobre suas audiências, O processo de análise incluiu a exploração inicial dos dados, a limpeza e transformação das variáveis, e a criação de novas variáveis que pudessem representar melhor o problema e a criação e validação de um modelo preditivo paraa taxa de engajamento dos influenciadores.

Introdução

O conjunto de dados abordado foi o “Dados dos principais influenciadores do Instagram (limpos)”, que se trata de um dataset sobre os principais influenciadores do Instagram e conta com dados como seguidores, total de curtidas, média de curtidas, posts e engajamento de cada um dos dozentos influenciados do conjunto de dados, baixamos o conjunto de dados do site Kaggle seguindo a orientação para esse projeto e a descrição no site sobre o dataset é ”O Instagram é um serviço de rede social americano de compartilhamento de fotos e vídeos fundado em 2010 por Kevin Systrom e Mike Krieger, e posteriormente adquirido pelo Facebook Inc. O aplicativo permite que os usuários carreguem mídias que podem ser editadas com filtros e organizadas por hashtags e tags geográficas. As postagens podem ser compartilhadas publicamente ou com seguidores pré-aprovados. Os usuários podem navegar pelo conteúdo de outros usuários por tag e local, visualizar conteúdo de tendências, como fotos, e seguir outros usuários para adicionar seu conteúdo a um feed pessoal. A rede do Instagram é muito usada para influenciar as pessoas (os seguidores dos usuários) de uma maneira particular para um problema específico - o que pode impactar a ordem de algumas maneiras”. Nós analisamos o dataset, vendo a ligação entre as variáveis e notamos a necessidade de algo para explicar a relação entre a taxa de engajamento com as outras variáveis do conjunto de dados, para resolver esse problema nós escolhemos usar o algoritmo de regressão linear por se tratar de um dataset onde a maioria dos valores são numéricos e aparentam ter uma relação linear entre eles.

Metodologia

Começamos a análise para ver do que se tratava o dataset, carregamos os dados, plotamos uma tabela, vimos os tipos de dados que tinha nela, e do que se tratava aquelas informações e se tinha alguma informação nula, logo após essa análise inicial, fizemos algumas transformações, traduzimos os nomes das variáveis para o portugues, criamos a variável continente para podermos ver a distribuição dos seguidores e curtidas dos influences ao redor do mundo de maneira mais geral, notamos que a maioria das variáveis era do tipo objeto e mudamos o tipo das variáveis apresentavam um valor numérico para float para facilita a criação de gráficos e tabelas futuras, depois dessas mudanças analisamos de novo a tabelas com as informações do dataset e a partir dali começamos a pensar nas relações entre as variáveis e o quanto elas influenciam no engajamento de cada usuario, pensamos no que fazia um influencer está acima do outronos ranks, o que fazia um ter um engajamento maior , curtidas nos seus posts maiores, começamos a explorando a relação entre o país e continente com outras variáveis, plotamos gráficos e tabelas para entender quais países e regiões tinham mais seguidores e curtidas, podemos visualizar que os Estados Unidos era o país que mais consentrava numeros de seguidores e curtidas, ou seja tinha mais influenciadores na tabela e mais influenciadores no alto do rank de usuario com mais seguidores. Depois de dessas análises nos viramos para a variável mais importante para resolvermos o nosso problema, a taxa de engajamente, que representava o percentual de engajamento de cada usuário em um período de 60 dias, pensamos quais variáveis eram calculadas para chegar naquela porcentagem de engajamento, fizemos um Lasso para selecionar as melhores variáveis e continuamos a analisar a tabela de dados, vimos que tinha usuários que com muitos seguidores, mas que em contrapartida tinham um número muito menor de curtidas por post ou seja apesar de terem um grande número de seguidores as suas postagens não alcançaram muitas pessoas, Pensamos em um coeficiente que divide o número de seguidores pelo total de likes, porque assim conseguimos saber o quanto cada seguido geraria de likes para o influenciado, mas depois de vermos os dados da Taylor Swift que tinha uma média total de curtidas de 2.4 milhões e uma média de curtidas em novos post de 2.3 milhões, o que significa que tudo que ela posta e curtido quase de imediato o que se refletia na taxa de engajamento dela ser altíssima decidimos mudar e fazer um coeficiente que pegasse o número de curtidas por post é dividir-se pelo número de seguidores, depois disso descobrimos que aquele valor da taxa de engajamento era resultado da divisão do número de curtidas por post pelo número de seguidores de cada usuário, depois dessa descoberta tomamos uma importante decisão que foi a de criar uma variável chamada engajamento_seguido e usar ela como variável dependente no modelo de regressão linear no lugar da taxa de engajamento, pensamos nisso porque a taxa de engajamento era uma porcentagem normalizada por código que ficava com um valor aproximado do coeficiente o que pode enviesar o modelo, já a nova variável que criamos era um valor real nao aproximado que nao foi transformado em porcentagem o que iria dar resultados gráficos melhores.

Após todas essas mudanças criamos um modelo de regressão linear para prever a variável dependente engajamento_seguido com base nas variáveis seguidores e posts, escolhemos essas variáveis por serem as que têm mais relação com o engajamento por seguidor dividimos 80% dos dados para treinamento e os 20% restantes para teste. Fizemos a normalização dos dados e tambem uma engenharia de atributos o que melhorou levemente o MSE e R2 .depois uma fizemos a validação cruzada(MSE por FOLD) e podemos ver o Erro Quadrático Médio (MSE) calculado em cada fold, o MSE se mantém consistente sem muitas variações o que prova que o modelo é confiável.

Resultados

As métricas usadas por nós para avaliar o modelo foram o Erro Quadrático Médio (MSE) , Erro Absoluto Médio (MAE) e o Coeficiente de Determinação (R^2). nós calculamos os valores do MSE e do R2 tres vezes a primeira logo após o treinamento a segunda depois de normalizamos os dados e a última após a adição do recurso 'seguidores_posts' os resultados mostram que o MSE geralmente está em uma escala relativamente alta. O que sugere que o modelo pode não estar capturando totalmente a complexidade dos dados. O R2 também varia, mas geralmente está abaixo de 0.5. Isso sugere que o modelo está explicando apenas uma parte moderada da variância. O MAE apresentou um valor baixo que significa que o modelo está errando pouco nas previsões

Nós utilizamos gráficos que mostram a relação entre Valores Reais e Valores Previstos, os Resíduos e os Valores Previstos e a Importância das Variáveis, o gráfico de valores reais e valores preditivos nos apresenta um gráfico com um começo muito bom com os pontos formando um linha vertical, mas os pontos vão se dispersando. O gráfico de resíduos e valores previstos apresenta um resultado melhor onde apesar dos pontos se dispensar em algumas áreas a maioria está distribuída ao redor do 0 e por último o gráfico da importância das variáveis apresentou um resultado que nos chocou inicialmente mas logo foi entendido, o gráfico de barras mostra que a variável posts tem a única importância isso se deve, porque no dataset tinha usuários que números muito pequenos de posts e altos valores de seguidores e curtidas por posts o que levou a taxa de engajamento e engajamento seguidor desses usuários serem altíssimas com isso acreditamos que o valor da variável post subiu por esse motivo usuários com poucos posts e muitas curtidas por post quase sempre tinham uma grande taxa de engajamento

Discussão

Nos como dupla estamos felizes com os resultados, apesar dos resultados não terem sidos sempre bons como alguns gráficos mostram, nos gostamos do resultado final, tivemos algumas limitações para desenvolver o projeto é a principal foi o tempo, nós estávamos muito ocupados nos dias de semana, essas duas últimas semanas em especial tivemos várias tarefas acadêmicas de projetos de extensão e etc, mas conseguimos nos reunir analisar os dados explorar tenta várias possibilidades e

combinações de variáveis para pode resolver o problema em questão da predição da taxa de engajamento, olhando para o resultado depois do modelo já pronto vemos que escolhas que fizemos durante a criação do modelo fizeram diferença no resultado de forma positiva e negativa, como quando priorizamos uma variável que nos meus criamos para contornar os valores em porcentagem que estava no dataset e isso melhorou o resultado pois os dados dessa variável eram dados reais e não aproximados , como dito anteriormente não ficou sem falhas, mas o resultado foi muito satisfatório

Conclusão e Trabalhos Futuros

Foi um trabalho gratificante, tivemos a chance de pôr a “mão” do código, por mais que já tenhamos feito os desafios e praticado no colab durante as unidades, fazer o relatório foi uma experiência diferente, foi um desafio, exigiu uma concentração a mais de nós o que melhorou as nossas habilidades de visualização exploração e treinamento de modelos uma sugestão para um melhoria no projeto e amplia os dados e variáveis do dataset como tipos de post conteúdos específicos e uma variedade maior de dados

Referências

[Dados dos Influenciadores Mais Importantes do Instagram | Kaggle]