Relatório do segundo trabalho de Grande Volumes de Dados

Fabiano de Paula Martins

4 de Outubro de 2016

1 Nova feature no site.

O site tinha um relatório de quanto tempo os usuários ficavam no site. Essa populção possui os seguintes parâmetros:

Tabela 1: Parâmetros da população

Parâmetro	Valor
Média	1.998297
Desvio Padrão	1.995156
Variância	3.980649

A resposta que queremos dar é se a nova feature tem feitos os usuários ficar mais tempo no site. Podemos três avaliações: a feature aumentou o tempo de permanência, a feature piorou o tempo de permanência ou a feature não fez nenhuma diferença no tempo de permanência. Podemos fazer essa análise em duas etapas: analizar se houve um aumento na média de tempo dos usuários e se essa média produziu lucro para o site.

Se a feature reduziu a média do tempo de acesso, ela deve ser removida por que causou o efeito contrário que deveria causar. Caso ela não tenha alterado ou tenha aumentado a média de tempo ela é boa ou indiferente.

Como o cálculo do tempo de acesso é contabilizado até o momento que o usuário fecha o browser e não tem como garantir se a máquina do usuário está em pleno funcionamento e sem delay, temos que fazer uma análise com vários graus de confiança nas medidas realizadas. Os graus de confiança nos dados serão 5%, 10% e 15%.

A hipótese de nula(H_0) é que a inclusão feature não fez qualquer tipo de alteração na quantidade de acessos ao nosso site. Porém, caso a feature tenha sido ruim para o nosso site a hipótese alternativa é que a feature causou uma alteração na quantidade de acessos do nosso site. Porém, não basta saber se houve alteração ou não. Caso tenha ocorrido uma alteração é preciso que a mesma seja para melhor. Pois caso não tenha sido isso pode implicar na remoção da feature do site.

Foi fornecida uma amostra com os tempos de uso dos usuários depois do lançamento da nova feature. Enquanto as população possui uma média de 1.998297, a amostra tem média igual 3.472058. Os ganhos são respectivamente R0.10 e R0.17. A tabela mostra os resultados com vários graus de confiança.

Tabela 2: Intervalos de avaliação dos resultados

Grau de confiança	Limitante Superior	Limitante Inferior	Conclusão
95%	1.994386	2.002207	Feature aumentou o tempo do usuário
90%	1.995005	2.001589	Feature aumentou o tempo do usuário
85%	1.995264	2.001330	Feature aumentou o tempo do usuário

A tabela mostra que em vários graus de confiança a feature aumentou o tempo de permanência dos usuários no site. Isso indica que a implementação dela foi positiva para o site. A permanência da feature no site é positiva para o site.

2 Clicks no site

As amostras apresentam se as pessoas que acessam o site clicaram ou não no banner do produto. Nessa situção podem criar duas categorias sobre os dados: "yes"por terem clicado no banner do produto e "no"caso não tenham clicado no banner.

A nossa H_0 é que independente de ter sido pelo pop-up o percentual de pessoas que clicaram no banner não será alterado. E a H_1 é que há diferença em exibir pelo pop-up ou não. O grau de confiança nos dado é de 97%, a tabela a seguir informa as frequências das categorias nas amostras:

Tabela 3: Frequências observadas nas amostras

Amostra	yes	No
A	301	682
B	387	621
Total	688	1303

Tendo feito a análise dos dados pelo método *Chi quadrado*, constatou-se que a presença de um pop-up faz diferença no percentual de cliques no banner do site. Devemos rejeitar a hipótese que H_0. Por fim, para maximizar os lucro com venda é indicado usar um pop-up para apresentar os produtos.

3 Produtor de cinema

Para analisar os dados removemos os registro que tem informações insulficientes. Os resgistros que não tinha o facenumber não foram contabilizados na análise. Queremos verificar se existe ou não uma correlação entre o número de faces no poster e o grau de imdb de um filme. Para isso vamos usar a correlação de spearman para chega essa conclusão.

Nossa hipótese nula será que não existe uma correlação entre essas informações, nossa hipótese alternativa é que exite uma correção entre as grandeza. O método de correlação de spearman, usando a implementação scipy.stats.spearmanr, retorno o valor $\rho = -0.087125$ e $p-valor = 6.03915^{-10}$. Como o $p-valor$ é muito pequeno, para qualquer grau de confiança relevante(90%, 95%, 99%) a hipótese nula é rejeitada. Com isso, podemos mostrar que existem uma correlação entre os dados.

Como o ρ é negativo isso implica que a relação é inversamente proporcional. Então quanto mais face o poster possuir, menor será o imdb score desse filme. Logo a recomendação é que se coloque o menos possível de faces no poster.