P07 – Udacity

Lieby Cardoso

Experiment Design

Metric Choice

Invariant metrics – Metricas invariantes

As métricas invariantes não sofrem alteração no grupo de controle ou de experimento, com base nesta orientação, foram escolhidas como métrica:

Number of Cookies: O número de cookies é a unidade de divergência deste experimento, espera-se permaneça sem variação entre o grupo de controle e o experimento;
Number of Clicks: O total de cliques no botão "Start Free trial" é contabilizado antes do momento do experimento, então o seu valor permanecerá constante entre o grupo de controle e experimento.
Click-through Probability: O valor da probabilidade de cliques é baseado no número de cookies e cliques, e assim como seus valores de referência permanecerão sem alteração. É uma opção de métrica interessante para nos ajudar a verificar se houve variação em ambos os grupos.

Evaluation metrics – Metricas de avaliação

1. Gross conversion: Para o propósito do teste a taxa de conversão bruta é interessante porque computa o total de usuários (user-ids) que se matricularam durante o período de experiência dividido pelo total de cookies que clicaram no botão "Start free trial"; Se a hipótese for verdadeira, espera-se uma diminuição no valor desta métrica porque algumas pessoas podem optar por não se inscreverem após a mensagem informando o tempo de comprometimento superior a 5 horas esperado do aluno.
2. Net conversion: Esta métrica mede o total de alunos que fizeram pelo menos um pagamento em relação ao total de cookies que clicaram no botão "Start free trial". Esta métrica complementa a de conversão bruta porque ela capta informações do segundo momento do teste, quando estamos mais interessados no total de alunos que permaneceram matriculados e efetuaram o pagamento após os 14 dias de teste do curso. Não espero uma grande alteração nesta métrica em relação ao grupo de controle, porque neste período de teste estou supondo que após os 14 dias, mesmo no grupo de controle, ficam sempre os alunos com mais tempo de dedicação disponível.

Em resumo, o lançamento do experimento será recomendado quando houver no grupo de experimento:

a. Uma diminuição mais expressiva no valor de Gross conversion, já que esperamos que o alerta na tela reduza a quantidade de alunos sem tempo disponível para se dedicar ao curso cadastrados;
b. Uma redução no resultado de Net conversion, mas esperamos, para que não haja perdas financeiras para empresa, que ela seja menos expressiva e esteja dentro do limite de significância prática.

Metricas descartadas

1. Number of user-ids: O número de identificadores dos usuários não foi uma boa métrica invariante porque só é possível capturar este valor após o clique no botão de matrícula. Esta métrica poderia ser usada como métrica de avaliação, mas seria redundante já que estou usando este dado na métrica de conversão bruta (Gross conversion). Minha preferência por usar Gross conversion está no fato de que os dados estão em forma normalizada e não em seu estado bruto como no user-ids.
2. Retention: Poderia ser uma boa métrica, mas como pode ser confirmado nos cálculos realizados neste experimento, para obter o valor desta métrica foram estimados 119 dias e mais de 4 milhões de visualizações da página. Caso fosse escolhida, supomos que haveria um aumento no total de alunos matriculados que fizeram pelo menos um pagamento, visto que, os alunos que não despunham de tempo foram alertados no início do experimento.

Measuring Standard Deviation

Probability of enrolling, given click	0.20625
Click-through-probability	0.08
Sample	5.000

Metric evaluation	value	n	SD (RAIZ(p*(1-p)/n))	m
Gross conversion	0.20625	400	0.020230604	0.039652
Net Conversion	0.109313	400	0.015601545	0.030579

Para o teste estou assumindo que a distribuição dos dados é binomial e se aproxima de uma distribuição normal a medida que o tamanho da amostra aumenta. Por esta razão foi utilizada a formula (RAIZ(p*(1-p)/n)) para obtenção do desvio padrão.

O total de cookies por dia foi utilizado no cálculo do desvio e da análise, desta forma podemos assumir que a estimativa analítica é bem próxima da medida da variabilidade empírica e será considerada para análise do teste.

Sizing

Number of Samples vs. Power

O teste ou correção Bonferroni não foi aplicado porque as duas métricas escolhidas estão correlacionadas e a margem de erro é menor do que a diferença observada, então para estes casos a mudança não é significativa.

Metric evaluation	value	p	d_min	Sample size	#Pageviews
Gross conversion	0.20625	0.08	0.01	25835	645875

Retention	0.53	0.0165	0.01	39115	4.741.212.121
Net Conversion	0.10931	0.08	0.0075	27413	685325

Devido ao elevado número de dias para teste e o volume de pageviews necessários, Retention foi descartado como métrica.

Duration vs. Exposure

O experimento consiste na apresentação de uma tela para o usuário com um questionamento que não envolve questões éticas. É considerado de baixo risco para o aluno e para a Udacity e não envolve a manipulação de dados sensíveis ou que possam comprometer o aluno. Com um tráfego de 100%, estão previstos 18 dias para a execução, o que pode ser um tempo muito longo para este teste.
Considerando o baixo risco oferecido e o tempo para execução, recomento que 100% do trafego seja utilizado nos testes.

Dias (Pageviews ÷ tráfego ÷ cookies)

Gross conversion: 645.875 ÷ 1.0 ÷ 40.000 = 16 dias
Net Conversion: 685.325 ÷ 1.0 ÷ 40.000 = 18 dias

Experiment Analysis

Sanity Checks

	Controle	Experimento
Pageviews	345543	344660
Clicks	28378	28325

	Probabilidade	SE	m	CI_Lower	CI_upper	Valor observado	Status
Pageviews	0.5	0.0006018	0.00118	0.49882	0.50117961	0.500639667	Passou
Clicks	0.5	0.0020997	0.00412	0.49588	0.5041155	0.500467347	Passou

As duas invariantes Pageviews e Clicks passaram na avaliação de sanidade.
Pageviews: CI (0.49882, 0.50117961) , valor observado = 0.500639667
Clicks: CI (0.49588, 0.5041155) , valor observado = 0.500467347

Como Click-through-probability é baseada nas invariantes Pageviews e clicks e as duas passaram na avaliação de sanidade, estou considerando que Click-through-probability também atende aos valores do intervalo de confiança (CI).

Result Analysis

Effect Size Tests

	Controle	Experimento
Clicks	17293	17260
Enrollments	3785	3423
Payments	2033	1945

					CI		
	p^	SE	m	d^	Lower bound	Upper bound	d_min
Gross conversion	0.20860707	0.004371675	0.008568484	-0.02055	-0.0291	-0.0120	0.01
Net conversion	0.11512749	0.003434134	0.006730902	0.004874	-0.0116	0.0019	0.0075

Gross conversion: Com o valor observado de -0.0205, CI (-0.0291, -0.012), dmin=0.01, o resultado tem significância estatística e prática.
Net conversion: Com o valor observado de 0.004874, CI (-0.0116, 0.0019), dmin=0.0075, o resultado não tem significância estatística, nem prática.

Sign Tests

Os dados foram calculados no site: https://www.graphpad.com/quickcalcs/binomial2/

Probabilidade	Experimentos	Sucesso Gross	Sucesso Net
0.5	23	4	10

Gross Conversion: Para o teste de duas caudas foi obtido um valor de p = 0.0026 < α=0.05, demonstrando uma significância estatística.
Net conversion: Para o teste de duas caudas foi obtido um valor de p = 0.6776 > α=0.05, com este valor maior que o α, não há significância estatística.

Summary

A correção de Bonferroni é indicada quando o experimento é composto por múltiplos testes estatísticos e se deseja evitar que a hipótese nula seja rejeitada mesmo quando é verdadeira. Neste experimento optei por não aplicar a correção de Boferroni, uma vez que:

- Ao aplicar a correção de Bonferroni o valor da alfa é dividido pelo total de testes, como temos duas métricas de teste teríamos um alfa de 0.05/2 = 0.0025, resultando num intervalo de confiança (CI) de 97.5% para cada teste. Este CI é muito conservador para o resultado que queremos alcançar, uma vez que, foi estabelecido em 0.05 para cada métrica;

- A correção de Bonferroni pretende identificar pelo menos um resultado significativo no grupo de testes, mas para este experimento defini que as duas métricas (Gross e Net conversion) precisam ter um resultado com significância estatística para que o teste seja lançado;
- Os dois testes estão correlacionados e usam a mesma unidade de divergência, o que somado aos dois fatores anteriores colabora para o descarte da correção de Bonferroni neste teste.

Não foi identificada nenhuma discrepância entre os resultados obtidos nos testes de tamanho do efeito e de sinal. Ambos demonstraram que havia significância estatística nos testes realizados para a métrica de conversão bruta (Gross conversion) e que não havia a mesma significância para a conversão liquida (Net conversion).

Recommendation

As métricas escolhidas para o teste A/B foram a conversão bruta (Gross Conversion) e a conversão liquida (Net conversion), com base nos resultados encontrados não recomendo o lançamento desta mudança no site da Udacity.
Foi observado que no grupo experimental menos pessoas clicaram no botão de inscrição, sugerindo que menos alunos ficariam frustrados futuramente no curso por não terem tido tempo suficiente para se dedicar. Este resultado foi apoiado por um valor de p maior que o alfa de 0.05, demonstrando sua significância estatística. Por outro lado, com um intervalo de confiança de 95% não obtivemos um nível de significância estatística e prática para a métrica de conversão líquida. Seria muito arriscado recomendar o prosseguimento desta mudança no site, uma vez que, não conseguimos demonstrar com um nível de certeza que haveria ganho para a Udacity em seu objetivo de reduzir o número de alunos frustrados que abandonam o período de teste gratuito, sem que houve perda significativa no número de alunos inscritos após este período.

A ausência de ganhos financeiros que podem decorrer do lançamento desta mudança, conforme sugere a falta de significância estatística na conversão líquida deve ser avaliada pela equipe Udacity, caso a empresa decida seguir com os testes.

Follow-Up Experiment

Vou usar minha própria experiência com o nanodegree da Udacity para sugerir um experimento futuro; acredito que exista um grupo de alunos como eu que possa ter experimentado uma frustação inicial, não por não ter o a disponibilidade necessária, mas por não ter a base de Python que era esperada. Para este grupo atender os primeiros meses de curso são necessárias mais que 5 horas semanais e provavelmente existe um outro de alunos com mais experiência que poderá atender todos os requisitos do curso com menos tempo de dedicação.

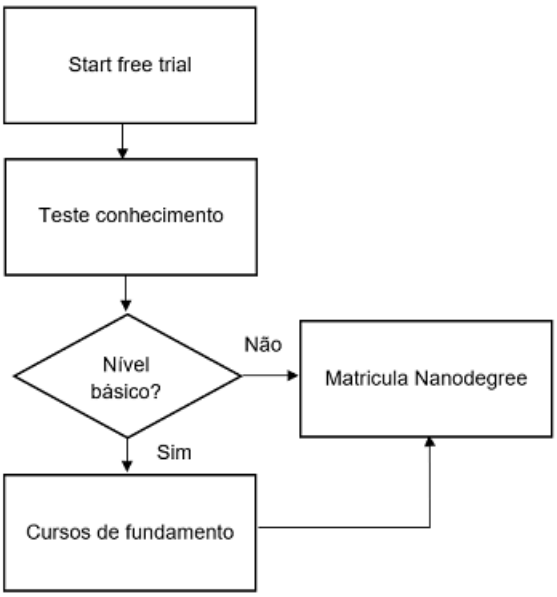
Neste sentido a Udacity poderia testar um experimento em que:

1. Ao demonstrar o interesse no curso, o potencial aluno é direcionado para um teste de conhecimento;
2. Após o teste de conhecimento o aluno é alertado quanto ao nível de comprometimento necessário para sua atualização de conhecimento e para execução do nanodegree;
3. Para os alunos com baixa pontuação a Udacity pode sugerir cursos complementares;.

A hipótese é que após o teste de conhecimento, o aluno que decidir se matricular no curso estará mais ciente do nível de exigência/dedicação e será mais provável que complete todo o curso, aumentando o percentual de retenção de alunos.

Métricas sugeridas para o experimento:

1. A unidade de divergência mais adequada é o cookie, por permanecer sem alteração tanto no grupo de controle, quanto de experimento. É a variável que é possível controlar antes mesmo do clique no botão do free trial e da criação do user-id;
2. Invariantes: Número de visualização de páginas (pageviews) e o número de clicks no botão de início do teste (Free trial) de 14 dias;
3. Avaliação: As mesmas métricas de conversão bruta (# user-ids matriculados / # cookies clicaram no botão Free Trial) e líquida (# user-ids matriculados e pagante / # cookies clicaram no botão Free Trial) usadas e descritas neste teste.



Para que o experimento seja lançado é necessário que tenhamos um crescimento estatisticamente significativo (CI=95%) na taxa de conversão bruta após clicar no star free trial. Para complementar a decisão é necessário que haja um crescimento menos expressivo, mas ainda com significância estatística (CI=95%) na conversão líquida de matriculados que fizeram o pagamento vindos do curso

de fundamento ou não. As duas métricas combinadas têm capacidade de demonstrar se o teste de conhecimento foi capaz de desencorajar a matricula de alunos que ainda não estavam preparados para o curso, sem que houvesse uma perda financeira para a empresa capturada através da taxa de conversão líquida.

Observações:
Considerando que no teste sugerido teríamos em média o mesmo volume de participantes do teste descrito neste trabalho, usar a métrica de retenção seria inviável por causa do total de pageviews necessários para o teste. Uma possível solução seria diminuir a expectativa em relação à taxa de conversão, mas caberá à empresa definir e comunicar suas expectativas em relação ao teste.

Apesar do teste não envolver questões éticas, legais e de risco para os participantes, seria mais prudente considerar um tráfego de 80% para este experimento, estendendo, para a mesma massa de dados, o tempo do experimento de 18 para 22 dias. Ao propor um teste de conhecimento estamos inserindo uma nova etapa ao processo de inscrição, e isto pode desanimar algumas pessoas de participarem, portanto, é importante expor somente parte do grupo ao experimento para que possamos fazer uma comparação.

Referências

http://www.evanmiller.org/ab-testing/sample-size.html
https://rpubs.com/superseer/ab_testing
https://www.graphpad.com/quickcalcs/binomial2/
http://onlinelibrary.wiley.com/doi/10.1111/opo.12131/pdf
http://www.portalaction.com.br/anova/teste-de-comparacoes-multiplas
http://adalee2future.github.io/udacity_data_analyst/AB_Test.pdf
http://lilychang.net/AB_Testing/