Experiment on the Efficacy of

Custom LinkedIn Connection Invitations

BAX 401 - Group 6: Tien Her, Ivy Zhang, Aditya Satpute, Qinyi Qiu

Abstract

This study aims to investigate the influence of customizing invitations on LinkedIn connection request acceptance rates. Using logistic regression, the study examined the impact of AI-generated versus human-crafted messages, message length, and recipient characteristics like gender, number of connections, professional experience, and company size. The analysis found that message type and length, as well as the recipient's number of connections and years of experience, were not significant predictors of acceptance. A power analysis determined that a sample size of 118 was sufficient to give the study a statistical power of 80% needed to detect the effects being investigated. The findings suggest that other factors may influence LinkedIn connection acceptance rates and points to the need for further research into the qualitative elements of professional networking.

Introduction

- **Background**

  LinkedIn, a leading professional networking platform, has become an indispensable tool for individuals and businesses seeking to expand their professional networks and opportunities. Establishing meaningful connections on LinkedIn is a fundamental aspect of leveraging this platform for career growth, business development, and knowledge sharing. As networking strategies evolve, the use of custom messages accompanying connection requests increases your chances of establishing a connection as claimed by LinkedIn.

- **Objective**

  This report aims to investigate the impact of custom messages on response rates when connecting with individuals on LinkedIn.

  To explore this, we designed an experiment with a control group and four treatment groups. The control group will involve sending connection requests without any accompanying message, while the treatment groups will incorporate custom messages of varying types and lengths.

  These treatments include messages generated by ChatGPT, a cutting-edge language model, as well as messages generated by human participants. The variations in custom messages will include both short and long messages to assess how message length impacts response rates.

  By conducting this experiment, we hope to shed light on the following key questions:

  How do personalized custom messages affect response rates compared to standard connection requests?

  Does the source of the custom message (AI-generated or human-generated) influence response rates?

  Is there a significant difference in response rates between short and long custom messages?

Hypothesis

- Hypothesis: Custom messages in connection requests lead to higher acceptance rates.

Methodology

Target Population & Recruitment Strategy:

- Focus on LinkedIn users with the job title 'Data Scientist'.
- Filter on 3rd + connections and set location to New York, New Jersey
- Employ a search strategy based on job titles and random assignment of invitation types.

Experiment Treatments:

- Control Group: Invitations without messages.
- ChatGPT generated templates (50 and 300 characters).
    - Long: Hello, I'm XX, I'm a master's student in business analytics exploring data science roles. I'd love to learn more about your role at [Company] and gain insights into the field. Would you be open to connecting and sharing your experiences with me?
    - Short: Exploring data roles, eager to learn about your work!
- Human-created templates (50 and 300 characters).
    - Long: Hi , I'm XX a UC Davis master's candidate in business analytics, passionate about data science and hoping to pursue it as a career. I want to learn more about your experiences as a data scientist. Excited to connect and learn from your expertise. Thank you for your time!
    - Short: Hi XX, I am currently exploring Data Science careers and am eager to learn from your expertise!

Data Collection:

- Plan for sending 200 to 300 invitations.
- Detail the process of random assignment using Excel function RANDBETWEEN.
- Collection of the LinkedIn user's data including Name(to track response), Gender, Company size, number of connections, year of experience.

Ensuring Randomness and Avoiding Bias:

- Randomization: In our experiment design, we use an Excel function "RANDBETWEEN" to generate random numbers from 1 to 5 and distribute how we should send out the invitation prompt accordingly. For example, if we receive a “1” then we put the profile in the control group, and if we receive a “2” then we send out the chatGPT prompt.

- Non-interference: Ensure that the person's one connection is not connected to the other, we chose the LinkedIn profiles that are 3rd-degree connections to avoid bias, given that it’s 3rd-degree it means that there’s no shared connection and would be more independent on whether the user accepts or not.
- Excludability: 3+ connections, Location changed to east coast, only data scientists

Data Analysis

Minimum sample size:
The minimum sample size was calculated using the power test with maximum allowable type 2 error as 0.2
Logistic regression is designed for binary dependent variables which is appropriate in our situation. In addition, Logistic regression does not require the dependent variable to be normally distributed.
In our data, the outcome is categorical with two possible results and the independent variables are a mix of dummy variables and continuous variables which makes Logistic regression ideal for our analysis.

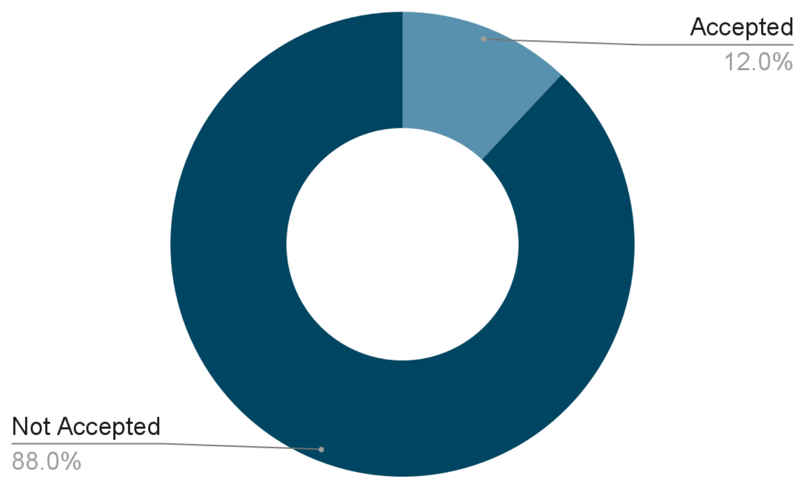
To ensure that our logistic regression model accurately captures the ATE, we decided to use the pwr.f2.test from the pwr package to calculate the minimum sample size.

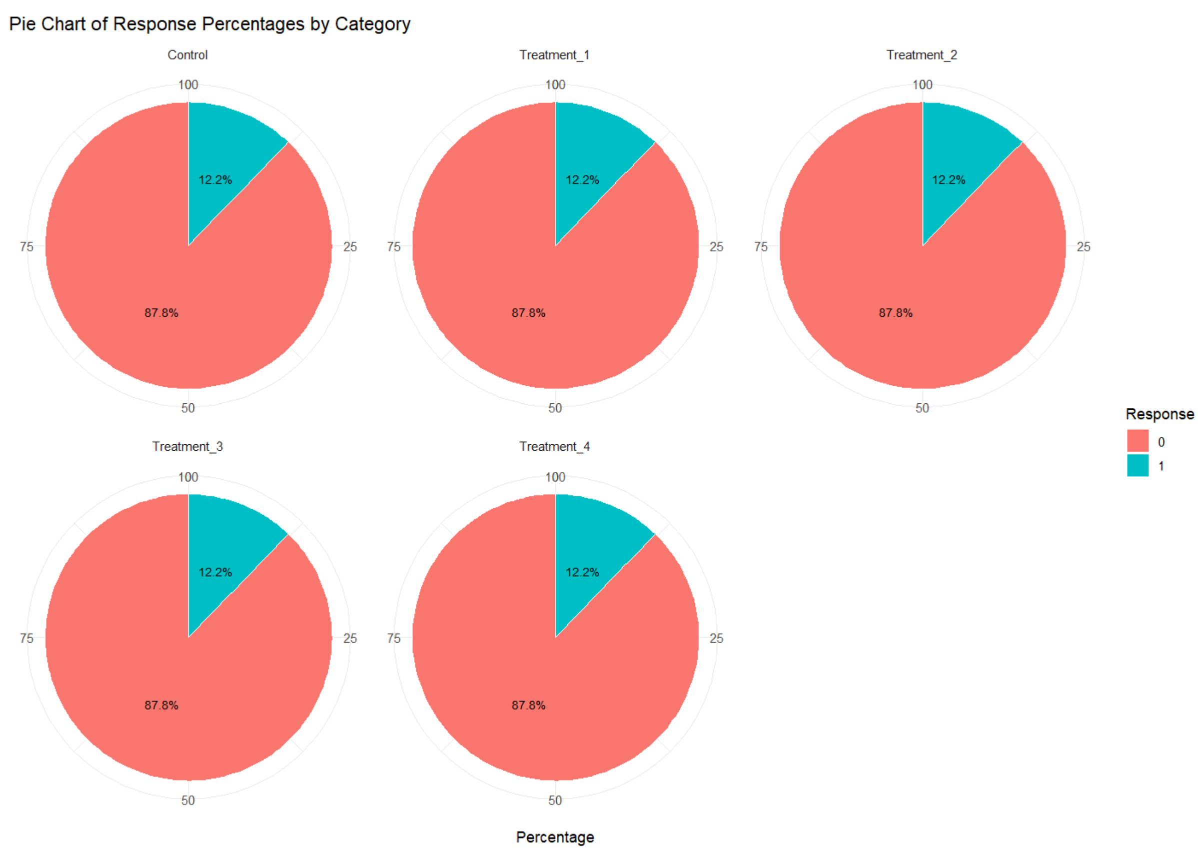
Our pilot indicated a rough acceptance rate of 15% and hence, we considered our effect size to be 0.15.
The following is the output from our power test to get the minimum sample size for a power of 80% with a significance level of 0.05

```
> f2 <- 0.15
> alpha <- 0.05
> power <- 0.8
>
> # Calculate the required sample size
> effect_size <- pwr.f2.test(u = 10, v = NULL, f2 = f2, sig.level = alpha, power = power)
>
> # Print the result
> cat("The required sample size is", ceiling(effect_size$u + effect_size$v + 1), "\n")
The required sample size is 118
```

Hence, the minimum sample size was taken as 118 for our experiment.

EDA:
Before starting regression analysis, we were curious about the results and the following were the aggregated results
The following charts show the total acceptance percentage and the treatment-wise acceptance percentages





We can observe an even distribution.

Next, we move on to regression and further analysis

Regression:

To start with, we run a regression for Acceptances on all our variables. The following is the R output.

$$\begin{aligned} Response = \beta_0 + \beta_1 \times Treatment_1 + \beta_2 \times Treatment_2 + \beta_3 \times Treatment_3 \\ + \beta_4 \times Treatment_4 + \beta_5 \times female_{\text{dummy}} + \beta_6 \times connections_{\text{dummy1}} \\ + \beta_7 \times connections_{\text{dummy2}} + \beta_8 \times years\ of\ experience + \beta_9 \times company\ size_L \\ + \beta_{10} \times company\ size_M \end{aligned}$$

Response: Accepted or rejected invitation

Treatments: custom message on the connection request

Female dummy: dummy variable on gender. If the female dummy = 1, the receiver is female, otherwise, the female dummy = 0

Connections: dummy variable on the number of connections the receivers already have.

Connection 1 = 1, the number of connections of the invitation receivers is greater than or equal to 250 and smaller than 500

Connection 2 = 1, the number of connections of the invitation receivers is greater than or equal to 500

Year of experience: total years of working experience of the receiver

Large/Medium: dummy variables with a base of small-size companies the receiver is working on. If Large = 1, the receiver is working in a large size company, otherwise, Large = 0. If Medium = 1, the receiver is working in a medium size company, otherwise, medium= 0.

```
Call:
glm(formula = Response ~ Treatment_1 + Treatment_2 + Treatment_3 +
    Treatment_4 + female_dummy + connections1 + connections2 +
    Years_of_experience + large + medium, family = binomial,
    data = data)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.510740   0.691947  -0.738  0.46044
Treatment_1         -0.038179   0.656244  -0.058  0.95361
Treatment_2         -0.005785   0.646726  -0.009  0.99286
Treatment_3         -0.566369   0.722143  -0.784  0.43287
Treatment_4          0.119716   0.674461   0.177  0.85912
female_dummy        -0.645997   0.438286  -1.474  0.14050
connections1        -0.687458   0.738498  -0.931  0.35191
connections2         0.163206   0.508561   0.321  0.74827
Years_of_experience -0.016345   0.044640  -0.366  0.71425
large               -1.318564   0.487796  -2.703  0.00687 **
medium              -1.688221   0.642638  -2.627  0.00861 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It can be observed that the company size dummies have a statistically significant effect on the odds of acceptance of invitation.

Next, we dig deeper to ensure that the covariates are balanced between the treatment and control groups.

Balanced covariates were achieved by dropping the company size dummy. The following are the final results from testing various covariate combinations.

```
glm(formula = Control ~ female_dummy + connections1 + connections2 +
    Years_of_experience, family = binomial, data = data)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.58713    0.43001  -1.365    0.172
female_dummy        -0.17846    0.32757  -0.545    0.586
connections1        -0.19409    0.48018  -0.404    0.686
connections2        -0.44985    0.38674  -1.163    0.245
Years_of_experience -0.04776    0.03732  -1.280    0.201
```

```
glm(formula = Treatment_1 ~ female_dummy + connections1 + connections2 +
    Years_of_experience, family = binomial, data = data)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.064572   0.435976  -2.442   0.0146 *
female_dummy        -0.240759   0.330053  -0.729   0.4657
connections1         0.229366   0.474786   0.483   0.6290
connections2        -0.278987   0.401521  -0.695   0.4872
Years_of_experience -0.006788   0.034412  -0.197   0.8436
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
glm(formula = Treatment_2 ~ female_dummy + connections1 + connections2 +
    Years_of_experience, family = binomial, data = data)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.51627    0.45142  -3.359 0.000783 ***
female_dummy        -0.05448    0.31968  -0.170 0.864678
connections1        -0.26370    0.54438  -0.484 0.628101
connections2         0.41434    0.40591   1.021 0.307365
Years_of_experience  0.01357    0.03314   0.409 0.682244
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
glm(formula = Treatment_3 ~ female_dummy + connections1 + connections2 +
    Years_of_experience, family = binomial, data = data)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.463005   0.477739  -3.062   0.0022 **
female_dummy        -0.005343   0.346969  -0.015   0.9877
connections1         0.046156   0.538210   0.086   0.9317
connections2         0.050588   0.432773   0.117   0.9069
Years_of_experience -0.013530   0.037381  -0.362   0.7174
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
glm(formula = Treatment_4 ~ female_dummy + connections1 + connections2 +
    Years_of_experience, family = binomial, data = data)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.49685    0.52641  -4.743  2.1e-06 ***
female_dummy         0.55474    0.35951   1.543    0.123
connections1         0.19296    0.57276   0.337    0.736
connections2         0.32059    0.46932   0.683    0.495
Years_of_experience  0.05225    0.03460   1.510    0.131
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Once this was done, we ran our regression analysis on the final model as follows:

$$Response = \beta_0 + \beta_1 \times Treatment_1 + \beta_2 \times Treatment_2 + \beta_3 \times Treatment_3 + \beta_4 \times Treatment_4 + \beta_5 \times female_{\text{dummy}} + \beta_6 \times connections_{\text{dummy1}} + \beta_7 \times connections_{\text{dummy2}} + \beta_8 \times years\ of\ experience$$

```
glm(formula = Response ~ Treatment_1 + Treatment_2 + Treatment_3 +
    Treatment_4 + female_dummy + connections1 + connections2 +
    Years_of_experience, family = binomial, data = data)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.4962943  0.6179874  -2.421   0.0155 *
Treatment_1          0.0493800  0.6248726   0.079   0.9370
Treatment_2         -0.0841355  0.6268182  -0.134   0.8932
Treatment_3         -0.2451657  0.6924123  -0.354   0.7233
Treatment_4          0.3344279  0.6410982   0.522   0.6019
female_dummy        -0.6145290  0.4263644  -1.441   0.1495
connections1        -0.8861933  0.7166145  -1.237   0.2162
connections2        -0.1484587  0.4740134  -0.313   0.7541
Years_of_experience -0.0009637  0.0434054  -0.022   0.9823
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Observations

It can be clearly seen that the treatment effect on the odds of acceptance are statistically insignificant.

From the slopes, we can see that longer custom messages slightly increase the odds of acceptance compared to no custom message at all, but there is not statistically significant difference between the control and treatments.

Additionally, we can observe that the covariates are also of no statistical significance in improving the acceptance odds.

The observations also closely align with the EDA results which shows little distinction between the treatment and control groups.

Conclusion

- **Treatment Variables (Treatment_1, Treatment_2, Treatment_3, Treatment_4)**: None of these variables are statistically significant since all their p-values are above 0.05. This suggests that none of the treatment messages significantly change the odds of acceptance compared to the baseline.
  It suggests that the content of the messages or the fact that they were written by humans or AI and long or short isn't influencing the acceptance rate in a way that can be detected by this model.
- **Female Dummy**: With a p-value of 0.1495, connecting with a female member does not significantly decrease the odds of acceptance.
- **Connection Variables (connections1, connections2)**: Both have p-values well above 0.05, indicating that the number of connections does not have a significant effect on the acceptance rate within the ranges defined.
- **Years_of_experience**: Also not statistically significant (p-value = 0.1838), suggesting that years of experience do not significantly impact the odds of acceptance.
- **Company Size Variables (large, medium)**: The company size dummy variables were discarded due to covariate imbalance in the treatment and control groups.

We can reasonably conclude that there is not enough evidence to support the claim that a custom message increases the odds of invitation acceptance, especially when connecting with people who have little to no similar attributes in common with the sender.

Improvement

- **Revealing Unseen Biases:** there may be other significant factors that impact acceptance rates in Linkedin, such as the receiver's ethnicity. The effectiveness of certain messages may vary across different ethnic groups due to cultural nuances in communication preferences and norms.

  Solution: Collect more details like the receiver’s ethnicity and ensure that your data collection methods are robust and ethical

- **Generalizability:** The current analysis was run only on people with “Data Scientist” as their current job role

  Solution: Increase sample size and include people from other job roles to provide a more generalizable result.