How Do We Increase Positive Responses To Emails?

https://github.com/DLeenheer/CSPB_4502_project

Dana Leenheer
Applied Computer Science
University of Colorado
Boulder, Colorado, United States
dana.leenheer@colorado.edu

Abstract

This project investigates the various ways of increasing positive responses to emails in business-to-business sales. A positive response includes the contact asking for a meeting, referring the email to a coworker, or a response asking for a follow up on a future date.

Sailes.com is an AI startup company that uses AI to automate the sales process so that enterprise sales teams can focus their energy on high value tasks. This investigation will investigate current data of positive responses to see what trends are present and explore ways to optimize campaigns for future success. The other focus of this project will be to investigate job titles within specific industries and business sizes (small, medium, and large) to see what trends and information is present. Knowing which job titles tend to be most frequent within an industry will help the optimization of the contact sourcing process.

Problem Statement/Motivation

The current campaign setup process at Sailes includes utilizing ideal job titles either provided by the client, internal customer success team or operations team. This investigation aims to review the current success rates of job titles by various metrics including location and day/week/month of the year. This job title and company data will be compared with external data focused on various target countries to determine if the current strategy is effectively covering the intended target population or if further refinements are needed such as using general data on prevalence and density of job titles by region to better inform campaigns.

A key focus of this project will be identifying the most common job titles for different criteria such as industry, revenue, job seniority level, and region/location. Given that the previously successful responses are heavily skewed by the focus of individual campaigns, target markets, target regions, and product or service sold by the client the previous successful email responses will be reviewed but not used as a baseline for future campaigns. Aligning a future campaign from a different company with the specifications of a previous company would not be successful or the most ideal strategy. Sometimes the campaign is focused on the senior level of the company, other times the focus would be on specific departments within the company based on the product or service sold. Given this the variety of specific job titles responding positively to emails would differ based on the targets sought after.

This investigation will try to answer questions like, what is the most common job title? Do job titles vary by region or company size? Do different industries have different varieties of job titles? What job title(s) respond positively to emails most often? Is there a variance in the day of the week or month for responses? What metrics affect positive responses the most? What job titles are most frequent withing different industries?

Literature Survey (Previous Work)

The literature review for this project will focus on a couple different areas including, sales prospecting strategies for business-to-business sales and artificial intelligence (specific to emails if found), job title

creation/assignment and/or general information on job titles (via websites like Indeed), job title lists for key industries, and business size classification.

An overall review of positive response trends was completed by Clive Cadogan, Sailes CTO in 2022. He reviewed the overall trends of positive responses via various monthly metrics across all clients for Sailes [Cadogan 2022].

My role as a Data Operations Specialist at Sailes also lends insights to this project and helps direct the dataset selection. My understanding of the current internal data confirms that it is focused on the specific requirements of the client specifications for campaigns. Additional external data will be compared against the internal data to account for this.

Proposed Work

Data cleaning and preprocessing are the key first steps for my project. There are two main data sets for my project, one coming from QuickSight featuring data that has already gone through one round of cleaning prior to getting utilized by the API and another set of data comprised from a set of CSV files. Based on the initial review of this dataset a third much smaller dataset will be generated from Apollo.io focusing on manager level job titles.

The data in QuickSight will be pulled from the data lake that Sailes maintains using SQL and currently resides in 3 different internal tables (this method and the 3 associated tables were identified after meeting with one of the software engineers at Saile). These tables also include columns that go outside the scope of this project so they will be excluded from the data set created. This is partially why a custom query will be created, so that only the specified data columns will be added to the data set for this project vs. the entire table. I will review the data once it is pulled from the data lake to confirm whether cleaning is needed. I will also utilize QuickSight to mine the positive response rate per title and/or location. The details of this data set will be further discussed in my data set section.

The other main data set is a collection of CSV files from Apollo.io featuring 50k rows each of contact data. These files were exported from this software in August 2023. These files feature the same column order and format which simplifies the data cleaning and processing steps needed to assemble this project. The first step will be to combine each of the CSV files together, confirm that the columns do line up exactly, trim the white space, and remove the duplicate rows, both processes are common tools I routinely utilize in Google Sheets.

My approach to this project will focus on the characteristics of each contact, without focusing on the campaign they are associated with (beyond the general category of industry). Once the two data sets are ready for mining, I will review the applicable data mining algorithms and utilize the corresponding ones. One initial algorithm is the InfoGain algorithm with positive response, yes or no being the class under consideration. The CSV file data set lends itself to classification to help form conclusions about the job title information.

A qualitative review of the job titles is the focus of this investigation. The previous successful email responses are useful information. However, discovering the trends within key industries of focus will provide Sailes with valuable information to consider as part of the overall strategy of campaigns. The qualitative review will be comprised of different drilled down views of the external dataset focusing on metrics of interest including location/region, industry, job seniority level, revenue range of the prospected company, and the number of employees at the company.

Data Set

There are two main data sets for this investigation. The first data set is data that has not been uploaded to a campaign the second data set is being sourced from a data lake via QuickSight.

The first data set is a collection of 45 files of contact data sourced from Apollo.io that contain at most 50k lines of data in each file. This data was sourced in August 2023. Each of the files has identical column headers which assisted in the compiling of the files into one main file. The selected columns or attributes of these files are:

- Title
- # Employees (this is the number of employees at a respective contact's company)
- Seniority (this is the seniority level of the job title, ex. Senior, Vice President, Director, etc.)
- Industry
- State (of the employee)
- Country (of the employee)
- Company State
- Company Country
- Annual Revenue
- Apollo Contact Id (this attribute was kept while compiling the files and was used to remove duplicate contacts/lines of data)

The second data set will be sourced from QuickSight and will also feature some of the above attributes (minus the Apollo Contact ID) but will also have:

- Contact ID
- Is_deliverable
- Email_response
- # of Email_response

The second dataset timeframe will be either the past 6 months or from January 2023 to October 31st of this year. This is due to staff and API structure being consistent and improved from pre-2023. Data sourced from 2021-2022 could still be valuable but 2023 will be the focus of this investigation. Both data sets will be kept on my personal computer.

Evaluation Methods

I will utilize the graphs and charts available in QuickSight to help gain understanding of the trends with the data. The timeline information will be displayed with a line chart to show positive response trends over time. An initial thought is to break this out by location as well. Athena also has a querying function that I will use to test my query on a sample set prior to querying the data lake. This gives me the ability to focus on the SQL query components first and then apply the specific data lake references for tables second.

I will utilize Excel pivot tables to gain understanding of the most common job titles. I will investigate ways to discover the most common keywords used among the titles. My initial thought was to create my own word cloud with Python for the keywords but if that is outside the scope of this class my next option would be to find a word cloud website online and copy the words in. Prior to creating the word cloud though I will create drilled down views of individual industries. My initial review of the external dataset showed significant variation between the job titles of different industries, so an individual industry view is needed to clarify trends.

Amazon QuickSight also gives me the ability to create various charts and graphs with the data so a separate python library or framework will not be needed for this project. Any fields I need to calculate (like the positive responses per location or job title) will be created with the custom calculated field option within QuickSight. I will also utilize the available functions within Excel as needed (ex. Count).

The qualitative review of the job titles will be a key focus of this investigation. The qualitative review will take place in Excel. Drilling into the trends within job titles of key industries is a key focus of this investigation.

Tools

I will utilize Excel, Amazon Athena, Amazon QuickSight, VS Code (if needed), and Google Sheets. The data cleaning will take place within Excel and Google Sheets and be uploaded via CSV to QuickSight for evaluation if needed. The summary functions combined with my domain knowledge might be enough to find some interesting conclusions. Given the qualitative focus of this investigation the main tool used will be Excel for the external data.

Milestones

I have separated the milestones into those that still need to be completed and those that have been completed.

Milestones to Complete

- Data cleaning – further cleaning of the job titles data set will be completed within the next week.
- Some of the job titles include phrases like 'Operations Director at ABC Company'. I will need to remove the 'at ABC Company' portion of the title.
- Data selection – the data in the data lake will be selected and organized within the next week to 10 days
- A literature review of past work related to this project will be conducted over the next week. This review will be purposefully inclusive of the different aspects of this project. The initial topics that will be searched for include research done on business-to-business sales in general, business-to-business sales and artificial intelligence, job title creation, job title importance, and business size classification.
- Data transformation will take place this week, both QuickSight, Excel, and Google Sheets have summary formulas that are user friendly.
- Data mining will take place shortly after the data transformation is completed.
- A drilled down version of the dataset focusing on key industries of interest will be created. The key industries will be selected based on internal company focuses and my current knowledge for the recurring industries focused on by clients. Given the variance in job titles observed in the initial review of the dataset individual views of five industries will be completed so that trends within the industry are clearer. Comparing the job titles between the information technology and health care industries could be interesting but doing a more detailed analysis within an industry will also provide information specific to the industry.
- Drilled down views of the titles by location and by revenue range will also be created. Domain knowledge of common clients and requests are the driver for these views.
- Region specific views of the job title dataset will be created. I will create a United States view and a EMEA view. These two views are based on domain knowledge and the commonly requested focuses of clients. The United States view will include Canada as North America is a popular region of focus.
- The # Employees and Annual Revenue columns need to be grouped. I am going to review the partitioning methods in the book and pick the one most suited to this task. Another option is to graph the annual revenue column as a line graph and decide on partitions based on the shape of the line. I could also divide the revenue intervals based on my own domain knowledge.
- Further pattern evaluation will take place shortly after the mining is completed, further mining may take place after the initial mining is completed.
- There will be a final report, a presentation slide deck including my findings and recommendations, and a video of the presentation created and submitted.

Completed Milestones

- Data cleaning – an initial cleaning of the individual job title files has taken place. This initial cleaning included deleting the columns of the files that go beyond the scope of this project or/and make the dataset harder to sort because of the additional cells used for the data. The removed columns included data such as contact's first name, last name, company name, various phone numbers, and the company ID number.
- Data integration - The individual job titles documents have been combined into one main document in Excel. This process included the following steps changing the file type from a CSV to a Google sheet file, removing the

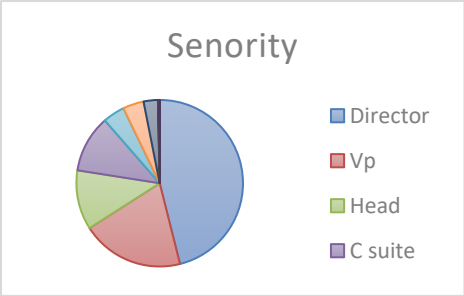
columns that go beyond the scope of the project, deleting all rows that feature an Employee number count of 0 (per my domain knowledge these rows of data have been commonly inaccurate), changing the file type back to a CSV, emailing the CSV to my school email address, downloading the CSV on my school laptop, importing the CSV to a temporary Excel sheet, copying this data into the main dataset for the project.

- Duplicate rows have been deleted from the dataset. There was a unique contact ID number column that each of the precompiled job titles had. The Excel 'Remove Duplicates' function was run on the entire dataset with the contact ID number field used as the comparison field. Once this operation was completed the contact ID number field was deleted to help reduce the total file size of the remaining dataset.
- GitHub account link for this project - https://github.com/DLeenheer/CSPB_4502_project

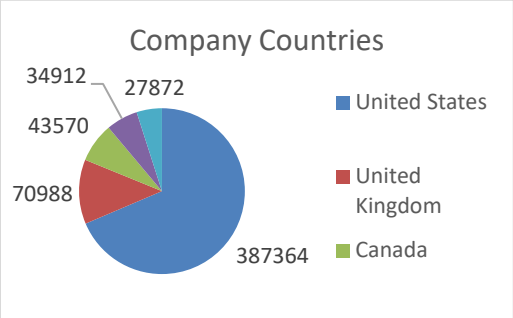
Results

The Excel pivot table functionality has provided some initial information about the job title dataset.

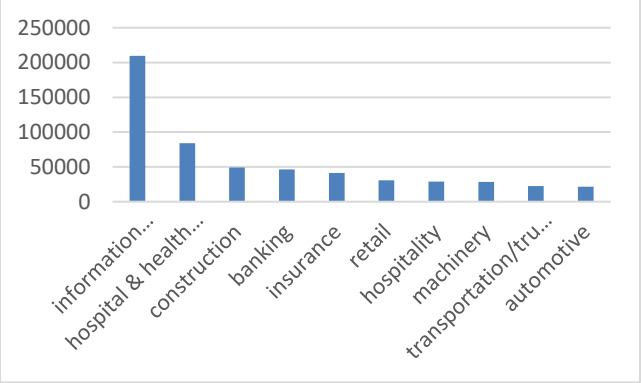
- The top 3 most common job seniority levels in the dataset are Director, Vice President, and Head (ex. Head of Engineering)



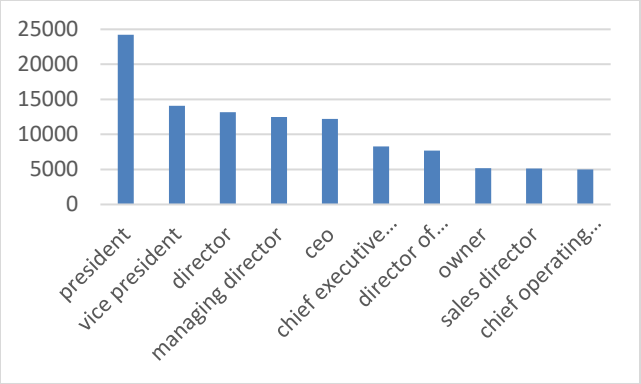
- The companies in the dataset are located in 210 countries. The top 5 countries are the United States, the United Kingdom, Canada, France, and Germany. The United States has the most companies in the dataset with 387,364 companies.



- The top 10 most frequent industries in the job title dataset are information technology & services, hospital & healthcare, construction, banking, insurance, retail, hospitality, transportation, and automotive. Information technology & services is the most common industry in the dataset with 209579 people.



- Currently the top five most common job titles across all industries in the data set are President, Vice President, Director, Managing Director, and CEO. My hypothesis is that they are general due to the variety of job titles found within specific industries and the titles that currently have company references in them.



REFERENCES

[1] Clive Cadogan. 2022. ML to Targeting: Optimizing Positive Engagement. Kansas City, MO. (Confluence document).