DATA 621: Final Project Proposal

Calvin Wong, Juanelle Marks, Kevin Benson, Ravi Itwaru, Sudhan Maharjan

11/12/2019

Introduction

In this final project, we will work with a datset relating to workers compensation insurance in the state of Florida. Workers compensation is a type of commercial insurance in which an employer takes out accident insurance for the benefit of its employees. If an employee is injured on the job or acquires a work-related illness, workers comp will cover medical expenses as well as lost wages until the employee is able to return to work.

We will focus in particular on a co-employment arrangement called a Professional Employer Organization (PEO), in which a typically smaller company outsources its HR, employee benefits, and payroll to a larger company that specialize in these functions. One such company is ADP's TotalSource PEO.

Research Question

In this project we will attempt to answer the following research questions:

- To what extent might TotalSource PEO have opportunities to grow market share in the South Florida business market?

- Specifically, can we identify companies that resemble those already represented by TotalSource PEO, and which therefore might be likely prospects to be converted to clients of TotalSource PEO?

Data Source

The dataset that we will use for this project is a publicly available dataset from the State of Florida Workers' Compensation Department, which contains workers comp coverage data over 2018 and 2019.
The dataset is sourced from the Florida workers comp website (https://www.myfloridacfo.com/division/wc/), and has been downloaded and saved as a raw CSV file at:

https://raw.githubusercontent.com/cwong79/DATA621/Calvin/Final%20Project/PEO1.csv

This dataset includes 10,730 cases of 25 variables, with each case relating to a specific company's workers comp insurance policy. The 25 variables include policy information (policy dates, insurance agent, carrier, etc.) as well as insured company information (name, location, number of employees, etc.).

```
## Observations: 10,730
## Variables: 25
## $ X1                          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1...
## $ `POLICY NUMBER`             <chr> "AMW0210001003", "AMW0210001003"...
## $ `EFFECTIVE DATE`            <chr> "4/1/19", "4/1/19", "4/1/19", "4...
## $ `CANCEL DATE`               <chr> NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ `EXPIRATION DATE`           <chr> "4/1/20", "4/1/20", "4/1/20", "4...
## $ `GOVERNING CLASS CODE`      <dbl> 8835, 8835, 8835, 8835, 8835, 88...
## $ `NAMED INSURED`             <chr> "COHESIVE NETWORKS 2 INC", "COHE...
## $ `CARRIER NAME`              <chr> "STATE NATIONAL INSURANCE COMPAN...
## $ `AGENCY NAME`               <chr> "AMWINS SPECIALTY CASUALTY SOLU"...
## $ `AGENCY CITY`                 <chr> "CHICAGO", "CHICAGO", "CHICAGO",...
## $ `AGENCY STATE`                <chr> "IL", "IL", "IL", "IL", "IL", "I...
## $ EMPLOYER                      <chr> "ANOLAZE CORP", "CARE HEALTH SER...
## $ STREET1                       <chr> "6985 GARDEN RD", "2290 10TH AVE...
## $ STREET2                       <chr> NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ CITY                          <chr> "RIVIERA BEACH", "LAKE WORTH", "...
## $ STATE                         <chr> "FL", "FL", "FL", "FL", "FL", "F...
## $ ZIP                           <dbl> 334045905, 334616607, 333042522,...
## $ COUNTY                        <chr> "PALM BEACH", "PALM BEACH", "BRO...
## $ PHONE                         <dbl> NA, NA, 9197891616, NA, NA, NA, ...
## $ `# OF EMPLOYEES FOR LOCATION` <dbl> 2, 1, 12, 4, 0, 2, 1, 1, 1, 1, 0...
## $ `LOCATION EFFECTIVE DATE`     <chr> "4/1/19", "4/1/19", "4/1/19", "4...
## $ `LOCATION CANCEL DATE`        <chr> NA, "9/21/19", NA, NA, NA, NA, N...
## $ NAICS                         <dbl> 623110, 623110, 111421, 561990, ...
## $ `WRAP-UP INDICATOR`           <chr> "N", "N", "N", "N", "N", "N", "N...
## $ `PEO INDICATOR`               <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y...
```

For purpose of this project, our target variable will be derived from `NAMED INSURED`, which is the beneficiary listed for each workers comp policy. The named insured may be the primary employer company, or it may be the PEO in the case of a PEO arrangement. We can identify the companies that currently use TotalSource PEO by filtering the target variable `NAMED INSURED == 'ADP TOTAL SOURCE INC'`.

Methodology

In this project we will develop a regression model to predict whether a company in the workers comp dataset is a client of TotalSource PEO, as labeled in the target variable `NAMED INSURED`. The regression model will most likely be a logistic regression, in which the binary response variable is:

- 1: `NAMED INSURED == 'ADP TOTAL SOURCE INC'`
- 0: otherwise

Potential predictor variables include various information about the companies, including industry, number of employees, location, and other data. Some of this data may have to be pulled from other data sources, such as business information websites.

Once we develop the logistic model, we can use this to predict the probability of a company being a client of TotalSource PEO. By comparing the predicted probabilities to the actual values of the target variable, we can identify cases where the probability is high (>0.5) but where the company is not a client. These are companies that resemble TotalSource's clients based on the predictor variables, but which are not yet clients. These could be likely candidates for conversion, based on similarity to TotalSource's current client base.