Datapalooza 2019

Team: It's Not the Data Point but the Size of the Error Bars That Count

(Karl Ahrendsen, Giovani Baez, Will Newman, and Andrew Vikartofsky)

1 Outline

Given the challenge "to create a potential score for each U.S. zip code that quantifies the opportunity for helping customers who may be a good fit for the product (Mutual of Omaha whole life insurance)," we were left with a wide range of avenues to explore to find a solution. We first thought of generating a score based on parameters we would deem fit for the insurance but this seemed to be too arbitrary and subjective which lead us to the idea of creating a machine learning algorithm to generate the scores for us. We found a set of external data that was used to train a set of machine learning algorithms to estimate the percentage of people buying whole life insurance based on various parameters [1]. Now with an idea of how our scores would be generated, we found census data on statistics of parameters that our model uses to generate scores for zip codes. This method gave us a data set which mapped percentages to zip codes to be called and futher enabled us to create a heat map to display the geographic regions where the product would be the best fit.

2 External Data

For our external data, we had some issues in finding a source that was both open use and applicable to the question at hand. We chose to use ZIP Code Tabulation Areas (ZCTAs) to categorize the census data as apposed to the standard postal U.S. ZIP Codes. The reason for this is that ZCTAs have open use files that are available for use while ZIP Codes do not have such information available. The ZCTA data set was obtained from American Fact Finder, a data tool available from the US Census Bureau [2]. From this service we obtained the data sets (with their 5 digit designation code from Fact Finder), percentages of age ranges (S0101), level of education (S1501), income ranges (S2503), housing ownership (S1101), vehicle ownership (S0802), and marital status (S1201). This was the basis for our generation of the scores of probable insurance purchases.

Using the ZCTA for our specific task of finding the likelihood of individuals buying Mutual of Omaha's guaranteed whole life insurance involved sorting through the tabulation format and information of the census data. This was done by running Python code which parsed through the Census data and organized the various parameters that applied to our model and fit the Mutual of Omahas life insurance policy description.

More specifically, Python with Pandas was used to clean the census data of any categories that were non-plottable values and conform the census data to match with the required inputs for our machine learning algorithm. Once organized the code is then fed to the algorithm to generate our scores.

3 Supervised Learning Algorithms

The CoIL data set contained demographic information for various unnamed US regions, each of which also reported a proportion of the surveyed population owning life insurance. This insurance fraction was used as a labeling metric for supervised learning, with training features based on the remaining CoIL parameters. The CoIL data was too sparse to determine any strong correlation between the training features and the target, thus linear models were assumed for the predictive algorithms.

Using 5-fold cross-validation to optimize the hyperparameters of six linear regression models, a Bayesian Ridge regression was ultimately selected to predict score values. The feature vector from which a score was predicted contained demographic information by ZCTA for education, age, income, marital status, car ownership, and home ownership. It was determined that the ideal candidate for Whole Life Insurance has at least a Bachelors degree, has an income of at least $50,000 per year, is married, is a home owner, owns at least one car, and 45-55 years of age.

4 Conclusion

The obtained scores range from 0-1 with 0 being the least likely to buy the insurance and 1 being the highest. Once the scores per ZCTA were found we generated a heat map which displays the scores in a nice visualization of where the highest density of potential life insurance buyers are on a larger geographic scale. Of course, for a more specific answer one could call a ZIP code and retrieve a score. Additionally we were able to take Census data sets ranging from 2011-2017 to create an animated blend of the score distribution that would better help understand where the trend of potential buyers were going. With these methods of analysis we can better understand the potential parameters contributing the most to buying insurance, where these potential costumers are located, and an overall trend of where they are moving to.

References

[1] P. van der Putten and M. van Someren (eds) . CoIL Challenge 2000: The Insurance Company Case. Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report 2000-09. June 22, 2000

[2] United States Census Bureau: American Fact Finder, https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml