Datapalooza 2019

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

(Karl Ahrendsen, Giovanni 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 explored a wide range of approaches in determining a solution. Ultimately, machine learning algorithms were leveraged to generate unique scores for each US ZIP code area. The COIL data set [1] was selected to train a set of supervised learning algorithms to estimate a score for people buying whole life insurance based on various parameters. U.S. Census data [2] provided a set of features that our model uses to generate these scores for each U.S. ZIP code. These results are presented both as set of tabulated data, and visually as a heat map displaying the geographic regions where ideal customers can be found.

2 External Data

Finding a source of data which was both open-source and applicable to training our model was nontrivial. The problem statement did not specify a particular zip code class to be used (ie, ZCTA vs USPS ZIP code), thus ZIP Code Tabulation Areas (ZCTAs) were mapped to ech U.S. geospatial region. This conformed well to the default format of the Census data, and provided a simple means of plotting our final scores across the country. The ZCTA-labeled Census data sets were obtained from American Fact Finder, a data tool available from the US Census Bureau. These data sets contained demographic information (with their 5 digit designation code) for age ranges (S0101), level of education (S1501), yearly income (S2503), housing ownership (S1101), vehicle ownership (S0802), and marital status (S1201) of citizens in each region. These data were munged to form the feature vector from which scores for probable insurance purchases were predicted.

Use of the ZCTA-labeled Census data to predict a score for buying Mutual of Omaha's guaranteed Whole Life insurance required extensive cleaning and sorting of said data. To accomplish this task, numerous scripts were developed to create and manipulate pandas dataframe objects in Python.

Additionally, geospatial plots required geometric data related to ZCTA region boundaries. This data was imported from Wolfram databases with the aid of Mathematica, wherein heat map plots were also produced to visualize our score results.

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, a home owner, owns at least one car, and is 45-55 years of age.

4 Conclusion

Our predictive algorithm produced scores ranging from 0-1. A score of 0 indicates low interest of insurance purchasing, while the maximum score predicted across the country served as a point to normalize the scale to unity. A heat map was generated to visualize these scores across the continental U.S, and a score for a specific ZCTA region can be retrieved using the provided csv file.

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