Datapalooza Competition Report

Team: ACCFS

1. Methodology

This problem requires us to predict scores of fit to buy whole life insurance for each zip code without knowing the ground truth. This is similar as unsupervised learning process, which could be approached by clustering or principal component analysis (PCA), etc. But both clustering and PCA only identify natural groups within observations. Potential scores for these groups are still out of reach.

Our approach is to build a rating system based on our collected features e.g., family type, economic status, disability population, etc. The weight for each feature can be adjusted depending on different prediction needs. In addition, we validate the rating system by k-means clustering to see similarities between high score groups and clusters. We also include an extra zip code list of special needs care organizations across US, which we think their clients are highly possible to buy this insurance. This rating system makes most use of the collected features and it could be easily adapted to train prediction models where ground truth value is known.

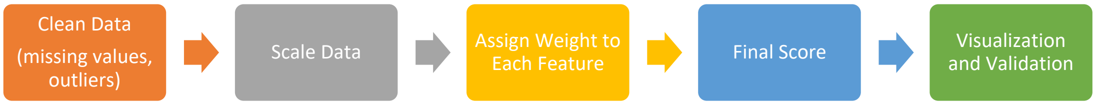
2. Identify Variables

The first question is who might choose whole life insurance. Based on provided information and online sources, it seems that economy and heath status are two major factors prompting people to buy this insurance product. In addition, family type and population size should also be considered. In the end, we identified ten variables and the reason why we selected each one of them are in the following table.

Variables Names	What & Why	Type
median household income dollars	Median household income. • Important economy indicator	Economy
Percent_Unemployment Rate	Unemployment rate • Important economy indicator	Economy
number_total_househlds	Number of total households • More households → higher needs	Size
grandparent_no_parent_fam_pct	Grandparent responsible for grandchildren - parents not present(%) • Special types of family → more likely to buy to benefit children	Family Type
single_chi18_fam_pct	Single family's percentage with under 18 years old children • Special types of family → more likely to buy to benefit children	Family Type
chi18_fam_pct2	Families with under 18 years old children – parent both working (%) • Special types of family → more likely to buy to benefit children	Family Type
one_working_chi18_fam_pct	Married couple families with <18 children – only one working (%) • Special types of family → more likely to buy to benefit children	Family Type
retir_income_pct	Families with retirement income (%) • Not likely to by if they already have retirement income	Family Type
no_child_one_working_fam_pct	Families with no <18 children – only one working (%) • Special types of family → more likely to buy to benefit spouse	Family Type
Disability_pct	Disability population (%) • Preconditioned people → more likely to buy to benefit otherers or themselves	Health Status

3. Format Dataset and Build Rating System

After selecting the variables, we need to prepare our data for the rating system. The dataset, which has each ZCTA code as a unit, were collected from US Census Bureau. It's 2013-2017 American Community Survey 5-Year estimates and there are around 3,3000 zip codes in total. Our data processing flow can be shown as follows.



Clean Data (missing values, outliers) → Scale Data → Assign Weight to Each Feature → Final Score → Visualization and Validation

In this dataset, missing data occurs usually because there were no sample observations or too few sample observations were available to compute an estimate. Thus, all missing data were treated as zero. As for outliers, raw data histogram (fig. 1) shows several features are strongly right skewed, especially number of total households. Because we plan to use min-max scaler to obtain scores for each feature. Outliers at far-right end would decrease rating sensitivity. We then capped all features at their 95% quantiles (winsorization) and take the log of number of total households (fig. 2). We also make sure these features have no strong correlations between each other to avoid repetitive information in the model during rating (seen fig. 3).

R

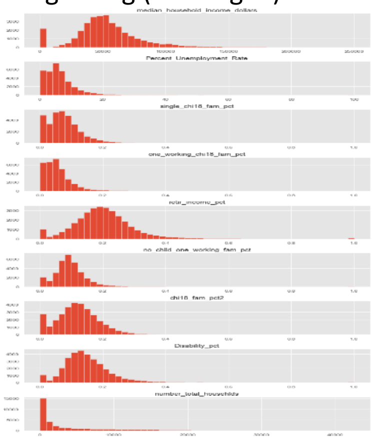


Fig. 1 Histogram of all variables –

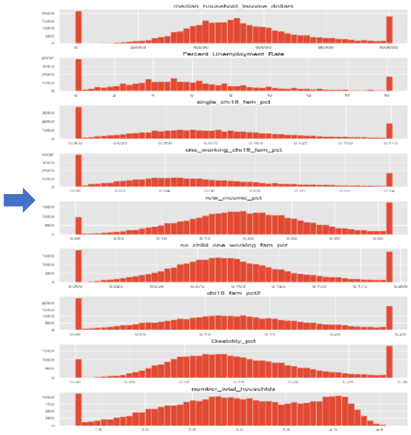


Fig.2 Histogram of all variables – winsorizec

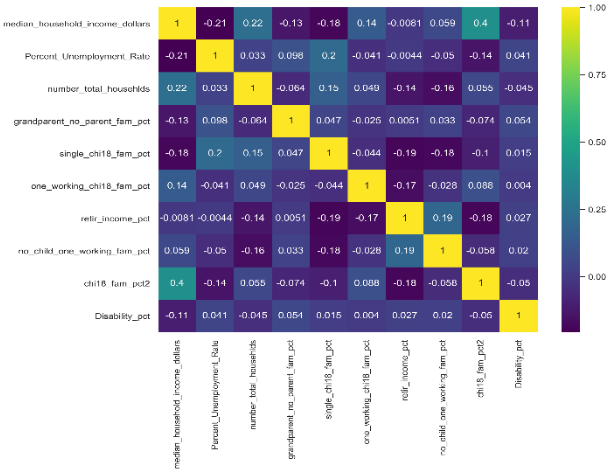


Fig. 3 Correlation heatmap of all variables

The rating model for the final score S is $S = w_1S_1 + w_2S_2 + \cdots + w_pS_p$, where $w_1 .. w_s$ are weights, and $S_1 \ldots S_p$ are scaled features (min = 0, max = 10). As ground truth is not available, the weight assigned to each feature is mainly based on subjective prediction purpose. Currently, we assume each variable type – family, health and economy contribute equally to people's tendency to buy whole life insurance. Variables under each type gets equal weight. You can put more weight on economy if you think it's the most important factor.

4. Results

The score map for the U.S is shown in fig. 4. To validate our rating system, we also used k-means clustering to see if high score zip codes could fall into the same natural groups produced by k-means. It shows that cluster 9 (fig. 5) matches most of the high score zip codes. This means our rating system could complement the limitations of clustering. It can identify high score groups from unlabeled clusters produced by unsupervised learning.

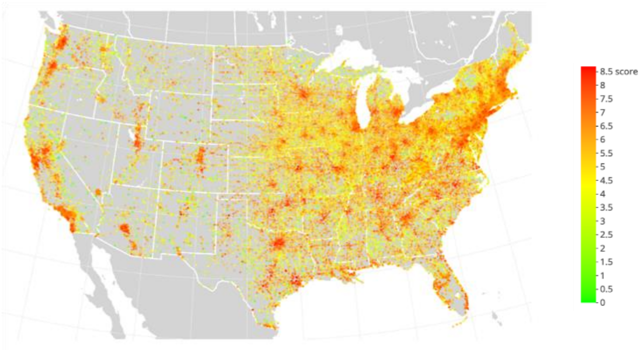


Fig. 4 Final Score by zip code

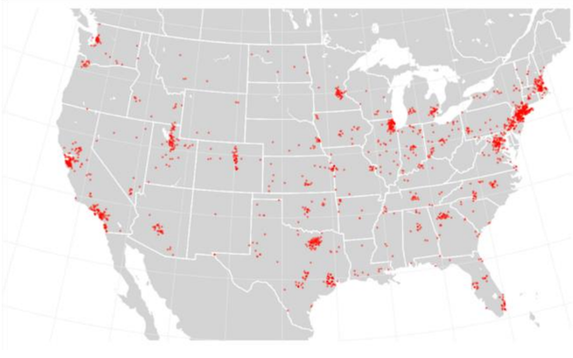


Fig. 5 Group of zip code from cluster 9

Data and Information Source:

https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml (sources for zip code data)

https://www.irs.gov/charities-non-profits/tax-exempt-organization-search-bulk-data-downloads (list of zip codes for special needs care organizations)

https://www.thesimpledollar.com/insurance/life/whole-life-insurance/ (sources for feature selection)