Jordan Hayes

Professor Schweikert

Data Mining/Predictive Modeling I

12/18/2023

**Segmenting Insurance Market with Data Mining Models**

**Abstract**

The point of this analysis was to segment the insurance market into target markets with descriptive attributes to come up with intentional marketing strategies supported by data. In this analysis, clustering was used to come up with distinct groups that were different from each other while within each group, data points were similar. Once the segments were established, multiple classifier algorithms were run; Decision Tree, naïve Bayes, and Random Forest, to classify new customers into the target markets established from clustering.

After a good classification model was established, the Apriori rule mining algorithm was run to find the attributes that were most prevalent in each cluster, which were also the data points within each cluster that are similar, so that marketing campaigns could be created around those attributes. In the data, the attributes used in the clustering, classifying, and rule mining algorithms were gender, diabetic, smoker, children, region, bmi, blood pressure, age, and claim. From this analysis, four distinct target markets were found that all had very different attributes, create a random forest model that was able to classify customers with 100% accuracy and describe the customers that make up each market segment for effective marketing.

**Related Work**

A Novel Approach to Market Segmentation Using Artificial Intelligence Techniques

             This article studies the use of AI in enterprises to gain a competitive advantage over other companies to better determine if a market segment should be entered by an enterprise. This article argues that market segmentation traditionally is only handled by upper management and has no scientific process to it, that it relies heavily on past success and experience. Because the market is constantly changing at a high pace, relying on past success will not accurately depict the present market. The article goes on to discuss which artificial intelligence methods they used, the results, and how artificial intelligence can “render market targeting more scientific and systematic, improve the quality of marketing decisions, maximize corporate profits, occupy the optimal market with limited resources, and achieve the goal of sustainable business” [4].

             The study uses naïve Bayes, J48, and OneR in Weka to classify markets as yes (to enter) or no (do not enter). The attributes that were used to measure a market’s attractiveness were channel accessibility, profitability, market size, stability, costs, and competition. During the training stage of the models, the naïve Bayes algorithm classified the markets the best at 100%, whereas J48 and OneR classified the markets at 91.7% and 83.8%. During the testing, each model returned differing results as to which markets should be entered or not. What this study found was that naïve Bayes has reliable predictions, J48 is less reliable, and OneR is the least reliable. This study concluded that AI algorithms can be used to optimize selecting market segments to enter, but a combination of actual business data and domain knowledge should be used when selecting market segments.

**Data**

The data used in this analysis was demographic information about insurance customers and their claims. The columns in the dataset were index, patient id, gender, bmi, bloodpressure, diabetic, children (how many children an individual had), smoker, and region. The initial pre-processing step that was taken was removing unique identifies so the columns index and patient id were dropped. The rest of the columns were the demographics relevant to describing each market segment and therefore important for clustering. My next task in processing the data was to find missing values and drop the rows without data. Out of 1,340 rows, only eight rows were dropped due to missing value.

Now that my data contained no missing values, code was run to detect outliers and if the outliers were significant enough to remove from the data. In the picture below, the graphic shows that only age contained no outliers. No rows were removed from bmi, bloodpressure, or claim because their z-scores were not above |3|.

A screenshot of a video

Description automatically generated

­

Next, new features were created for bmi, blood pressure, and age. For bmi, a new feature was created called wtClass that binned data into the categories **underweight** (bmi <18.5), **healthy** (bmi 18.5 - 25), **overweight** (bmi 25 - 30), **Class 1 Obese** (bmi 30 - 35), **Class 2 Obese** (bmi 35 - 40), **and Class 3 Obese** (bmi > 40) [3]. For age, a new feature was created called ageCLass with the ages binned into groups **Young Adult** (age 18-35), **Middle-Aged** (age 36-55), **Older Adults** (age 55+) [7]. For bloodpressure, the recognized level for high blood pressure is over 130 (2), so a new feature was created called bpClass binned into groups **Low Blood Pressure**(bloodpressure <130) and **High Blood Pressure** (bloodpressure >130). The last new feature created was called claimClass binning the claims row into groups **Low Claim** (claim < $25,000), **Medium Claim** (claim $25,000-$50,000), and **High Claim** (claim > $50,000). The new feature claimClass ended up not having much effect on the study, but this could be due to the skewed data. The analysis may have been more effective if claim was binned better, or if it was normalized.

**Methodology**

A graph with a blue line

Description automatically generated             The most important part of this study was clustering the data to create a class that would represent different segments/target market. The k-modes algorithm was used because my features were categorical, and k-means cannot be run on categorical data. Before running the k-modes clustering method, the elbow method was by creating a graph in Python to create the graph pictured below. This graph represents the costs of the algorithm for how many clusters there are. Four clusters were decided as the optimal number because the cost decreases at a slower rate at five than it did at four.

After the optimal number of clusters was determined, a new dataset was created with the row Cluster to add a column to each row that represents the cluster it was grouped into. To make sure that four clusters were the right k-amount, data sets with seven clusters and eleven clusters were created to see if more clusters would affect the accuracy of the decision tree. Now that there were datasets with differing cluster amounts, datasets with dummy columns for each attribute; gender, diabetic, smoker, children, region, bpClass, ageClass, cliamClass were created so that a decision tree and random forest classifier can be run. Now that there were data frames with binarized values, a decision tree classifier was run on each data frame with different k amounts to see which had the best accuracy predicting the Cluster class.

             After running the decision tree classifier, the data frame with four clusters had the highest accuracy in algorithms using information gain and the Gini index. The accuracy results were as follows: Data with four clusters; Information gain of .93 and Gini Index of .94, data with seven clusters; Information gain of .91 and Gini Index of .91, data with 7 clusters, Information Gain of .88 and Gini Index of .86. Below is the final decision tree created in R.

A diagram of lines and dots

Description automatically generated

A screenshot of a computer code

Description automatically generatedNow that it was apparent that four clusters would produce the best model, R was used to run three different classifiers to see which would have the best results. A decision tree, naïve Bayes classifier, and random forest were run. The best model with the best results was the random forest classifier which had a perfect accuracy of 1. This means that the random forest model can be used to accurately group customers into target markets that were segmented during clustering.

A screenshot of a computer

Description automatically generated

             Next, Apriori was run in R to find which attributes lead to customers being grouped into different clusters. When Apriori was run, the rules were sorted by lift so that the attributes that had the highest correlation to a specific cluster were noted. Then, the rules were sorted by confidence to find which attributes were most prevalent in a cluster. The attributes in the rules that had the highest lift and confidence will be used as key attributes in differentiating marketing strategies for different target markets. One thing to note before the results was that bpClass and claimClass were taken out of the data sets because they were heavily skewed to one class. The results may have been better if these columns were normalized.

**Results**

             The segments that were created with their key attributes from the results of Apriori were **Segment 1:**female, diabetic, middle-aged, northwest region, 0 children, overweight, non-smoker**, Segment 2:**female, not diabetic, southwest, class 1 obese, middle-aged, **Segment 3**: male, diabetic, young adult, non-smoker, no children, **Segment 4:** male, not diabetic, 1 child, class 2 obese, southeast, non-smoker. It is worth noting that all attributes used as key attributes were from rules with a lift greater than 2.5 or a confidence greater than .75.

             From the algorithms run, new clusters can now be grouped into a target market that will pertain to them the most. The attributes that are most prevalent in those target markets were found scientifically and are used to come up with customer profiles for advertising strategies. Applying the results from this analysis would look something like this: Come up with an advertisement that is personalized to a segment, for example, an advertisement for **Segment 4** that is targeted towards males. In this advertisement, diabetic and smoking-related insurance policies should not be mentioned because they won’t be relevant to this target market. However, having a man with one child incorporated in the advertisement and mentioning insurance claims that will benefit overweight people that are only shown in the southeast would be an efficient use of my analysis.

When new customers are added to the database, their demographic information can be thrown into my random forest model, and they will accurately group into a segment that fits their demographic information. The algorithms from my model will cover the full cycle of segmenting the marketing, targeting the different markets, and then positioning the advertisements in the relevant areas. From the random forest results, we can see that the model’s predictions and the cluster label are statistically significant because the p-value is 2.2e-16, much less than .005.

References

[1] Bonthu, H. (2023, September 14). KModes clustering algorithm for categorical data. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/>

[2] Center for Drug Evaluation and Research. (n.d.). High blood pressure–understanding the silent killer. U.S. Food and Drug Administration. <https://www.fda.gov/drugs/special-features/high-blood-pressure-understanding-silent-killer#:~:text=Normal%20pressure%20is%20120%2F80,seek%20medical%20treatment%20right%20away>.

[3] Centers for Disease Control and Prevention. (2022, June 3). Defining adult overweight & obesity. Centers for Disease Control and Prevention. <https://www.cdc.gov/obesity/basics/adult-defining.html>

[4] Chang, Y.-T., & Fan, N.-H. (2023). A novel approach to market segmentation selection using artificial intelligence techniques. Journal of Supercomputing, 79(2), 1235–1262. <https://doi-org.jerome.stjohns.edu/10.1007/s11227-022-04666-2>

[5] Devastator, T. (2023, January 7). Insurance claim analysis: Demographic and health. Kaggle. <https://www.kaggle.com/datasets/thedevastator/insurance-claim-analysis-demographic-and-health>

[6] How to plot a large ctree() to avoid overlapping nodes. Stack Overflow. (1959, January 1). <https://stackoverflow.com/questions/13751962/how-to-plot-a-large-ctree-to-avoid-overlapping-nodes>

[7] Petry, N. M. (2002, February). A Comparison of Young, Middle-Aged, and Older Adult Treatment-Seeking Pathological Gamblers. Academic.oup.com. <https://academic.oup.com/gerontologist/article/42/1/92/641498>

[8] Sklearn.model\_selection.train\_test\_split. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.train\_test\_split.html#sklearn.model\_selection.train\_test\_split