Health vs Insurance

How different factors influence knowledge

Group 18

Katlyn Rhodes, katlyn.j.rhodes@gmail.com
Chris Gutierrez, josephcgutierrez@gmail.com
Michal Chowaniak, michal.chowaniak@gmail.com

Abstract

Lower socioeconomic status individuals along with those who are seemingly healthy have lower knowledge and less comfortability when it comes to researching health insurance. By researching these areas and finding specifics that could group these individuals together, this could be used as a resource for health insurance companies. Health insurance companies would be able to take this information and better market their services while having a better understanding of their target audience. The goal of this project was to be able to target areas of people that have low to little knowledge about health insurance. The target areas that were address and discovered either had low to little knowledge about insurance or they were not comfortable with researching health insurance on their own.

Preprocessing and binning was used before any of the analyses could be ran. The data set that was used for the project was very wide and had a large number of responses recorded. The cleaning of the data came from removing outliers, data with no responses, and also when people refused a response.

After the preprocessing, decision tree analysis were used first to predict the High and Low confidence levels of researching using the internet or other online sources to compare health insurance plans. The dependent variable of "How comfortable are you researching and comparing insurancing market plans using the internet or other online sources" was binned to High Confidence researching and Low Confidence researching. Multiple decision trees were ran using different specifications in order to find the best tree and data set. The decision trees did provide a lot of data, however it was not as expected. Further testing and decision tree analysis would be needed for the future.

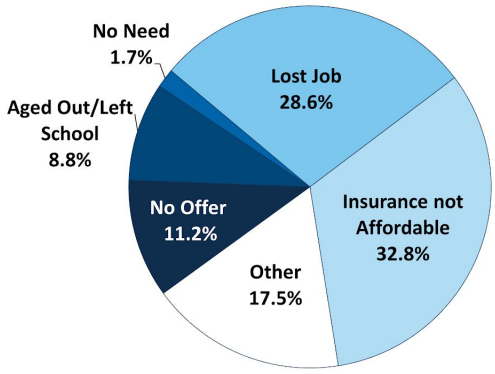
After testing our original hypothesis with the decision tree analysis, cluster analysis was used to do some unsupervised learning. Two versions of cluster analysis were used for this project. The first version was using an unbinned version of the comfortable variable. The second version was based on the binned version of the comfort variable. Both versions were similar but the binned version was the easiest to manage. Utilizing the data in these steps led to four decision trees based on the two sets of clusters. These data sets were higher comfort segments binned and unbinned along with lower comfort segments binned and unbinned. These were segments that showed higher and lower level of comfort and the goal was to be able to see what the rest of the individual looked like for segmentation.

Problem Description

The business goal of this project was that the predicted forecast could be used to help insurance companies target those with low knowledge of insurance and help them get the better care and coverage that they might need. The forecasting goal was that this project would be able to provide insurance companies with forecasts of how socioeconomic factors along with health status correlate with knowledge of health insurance. Specifically targeting those groups that have little to low knowledge. The interest in this topic came from research on different insurance levels and the groups that correlated with them. It was found that lower socioeconomic status individuals don't have the knowledge base to have "good" insurance. There are plenty of factors that would be the cause of this. Some of those factors would be the cost and ability to obtain health care. This can limit lower socioeconomic status individuals with their treatment and prevent them from receiving quality medical care on the event that they do become sick. Another factor is that typically healthy people tend to not know as much about different types of health insurance available to them as their peers who already have some sort of an illness. This project's aim was to find those that do have good knowledge and then turn and focus on those that do not. Individuals with chronic illnesses such as cancers, were seen to have more knowledge because this is something they have to actively deal with. The following pie chart gives a look at reasons for being uninsured among uninsured nonelderly adults.

Figure 3

Reasons for Being Uninsured among Uninsured Nonelderly Adults, 2013



SOURCE: KCMU analysis of 2014 National Health Interview Survey.

THE HENRY J. KAISER FAMILY FOUNDATION

By examining this, it is apparent that there are quite a few reasons that adults are not insured. This project aimed to find patterns within groups of individuals of those were not comfortable or knowledgeable when it came to health insurance and planned to give the insurance companies this information for them to better target their audiences.

Data Set

The data set for this project came from Health Americas 2014 survey done by the Robert Woods Johnson Foundation and Health Americas Foundation. This paper was aimed at Hispanics but the data also included White and Blacks. This paper originally examined knowledge of the Affordable Care Act but for this project took one of the broader variables as the Y. Likert Scale. This was "What is your comfortability researching health insurance online?" This data set included demographic information as well as behavior and diagnosis for certain chronic illnesses. By refocusing the data, this project was able to find broader insights than what the original researchers were after.

The dependent variable used for this project was "How comfortable do you feel using the internet or other online sources to compare health insurance plans." Which the project referred to the dependent variable as comfort level researching insurance online.

The following two charts are the independent variables that were used.

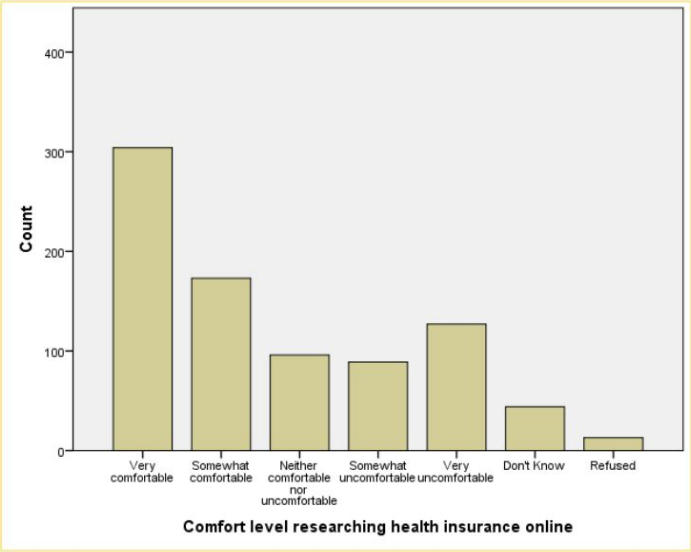
Marital status
Employment status
Last grade of schoold completed
Annual household income
Gender
General health
Do you smoke
How many days of a week do you do physical activities like exercise, sport active hobbies
How many servings of fruits and vegetables do you eat each day
How ofter do you drink soda or pop
Are you watching your sodium intake
Diabetes

Hypertension
High cholesterol
Lung or breathing problems
Depression or mental health condition
Heart or coronary artery desease
Cancer
Obesity
Race
How many days during last 30 days your health was not good
Body mass index
Age

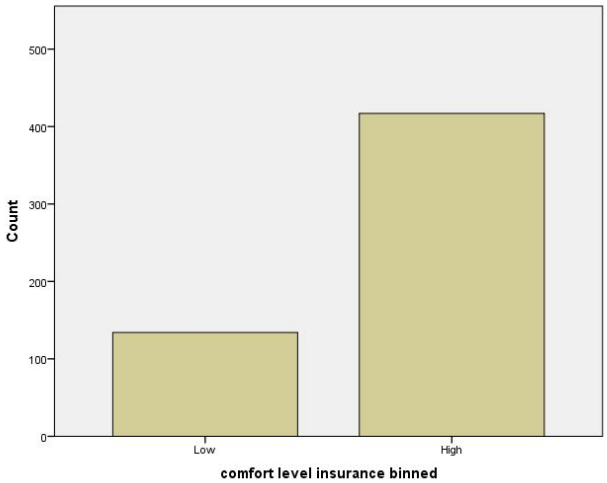
Preprocessing Steps

There were quite a few preprocessing steps that were taken to get the data ready for analysis. First, we went through the codebook for the original data set to remove any variables that weren't directly related to the new study. This can be seen in the previous section.

Binning was of the major steps that lead to improving the quality of the data in this project. The dependent variable was binned. The data that was used for this project was mostly categorical. How comfortable do you feel using the internet or other online sources to compare health insurance plans was changed to high and low comfortability.



GGraph



Additionally, race required binning. The original data looked at Hispanics, Whites, and Blacks. However, Hispanics were broken into two variables: insured and uninsured. This made sense for the original study that created this dataset. It was focused on Hispanics, with Whites and Blacks acting more or less as controls. The new study's broader focus necessitated the creation of one Hispanic variable.

Several of the independent variables had responses such as "refused" and "do not know." These responses were cleaned and removed out of the data set prior to the data analysis. The numerical variables were normalized using z-scores and the outliers were removed so to prevent any false clusters or data sets.

In order to run different models of decision tree, we created 4 binary variables, which essentially were splitting out data set to training set and test sets. Two variables randomly split data set to 60% training and 40% test sets and two of the variables split data set to 70% training and 30% test.

The first few attempts at creating the decision trees, in a way, acted as a step in the data preprocessing and variable selection process. The decision trees produced different models with different variables each time.

model	Comf Lev Binned (1 or 2)	Split (60/40 or 70/30 and version	Variables chosen be decistion tree	Risk Training	Risk Test	%Correct Training Low	%Correct Training High	%Correct Training Overall	%Correct Test Low	%Correct Test High	%Correct Test Overall
a	1	60/40 v1	General Health, income, BMI, fruit serving, marital status,	0.15	0.25	45.00%	97.00%	85.00%	31.50%	90.00%	75.00%
b	1	60/40 v2	General Health, Employment status, age, marital status, do you smoke	0.16	0.28	41.50%	95.80%	83.70%	21.20%	92.40%	72.30%
c	2	60/40 v1	General health, income, age, how often do you drink soda, do you smoke, bmi, how many days do you exercise, how many servings of fruit	0.21	0.43	64.60%	88.20%	79.50%	38.70%	67.40%	56.90%
d	2	60/40 v2	Income, Employment status, fruit servings, general health last grade of school	0.20	0.36	56.60%	92.00%	79.60%	37.00%	81.10%	63.60%
e	1	70/30 v1	general health, age, do you smoke, income, soda, exercise, race, bmi	0.15	0.27	34.10%	99.00%	84.80%	17.30%	96.70%	73.10%
f	1	70/30 v2	general health,, age, last grade of school, exercise, hypertension	0.19	0.19	38.00%	97.00%	81.10%	26.50%	93.20%	80.70%
g	2	70/30 v1	age, income, exercise, employment status, maritas status, diabetes, fruits, race	0.21	0.33	50.80%	93.50%	79.00%	32.40%	92.10%	66.90%
h	2	70/30 v2	income, age, last grade of school, exercise, employment status, marital status, bmi,	0.20	0.35	66.20%	88.30%	80.00%	52.40%	72.00%	65.20%

The above table shows the results of the eight decision trees that were ran. Each data split variable was used in two decision trees, with differently binned Y variable. Once with middle section (Neither comfortable nor uncomfortable) binned with high values (Very comfortable, somewhat comfortable) and for a second time binned with low values (Somewhat uncomfortable, very uncomfortable).

This project chose to focus on a data set which produced the overall highest percent correct for the test. There was a problem with a very high error for prediction low confidence level researching insurance. This could have been because the Y variable was not in balance.

Data Mining Techniques

Decision Trees

In order to try and fix the problem with high error for low confidence level researching insurance online, different decision trees were ran. Chosen data sets were modified by changing the minimum parent node parameter and the minimum child node parameter. However this produced bigger trees with many branches and overall lower percent correct, even though post pruning was selected.

Classification

Sample	Observed	Predicted Low	Predicted High	Percent Correct
Training	Low	58	42	58.0%
	High	11	259	95.9%
	Overall Percentage	18.6%	81.4%	85.7%
Test	Low	12	22	35.3%
	High	21	126	85.7%
	Overall Percentage	18.2%	81.8%	76.2%

Growing Method: CRT
Dependent Variable: Comfort level reasearching insurance online binned

Classification

Sample	Observed	Predicted Low	Predicted High	Percent Correct
Training	Low	60	40	60.0%
	High	11	259	95.9%
	Overall Percentage	19.2%	80.8%	86.2%
Test	Low	13	21	38.2%
	High	21	126	85.7%
	Overall Percentage	18.8%	81.2%	76.8%

Growing Method: CRT
Dependent Variable: Comfort level reasearching insurance online binned

Because of that we decided to choose our original tree produced at preprocessing stage. The table below shows that Low Confidence can be correctly predicted in 26.5% and that High Confidence can be correctly predicted in 93.2%.

Classification

Sample	Observed	Predicted Low	Predicted High	Percent Correct
Training	Low	38	62	38.0%
	High	8	262	97.0%
	Overall Percentage	12.4%	87.6%	81.1%
Test	Low	9	25	26.5%
	High	10	137	93.2%
	Overall Percentage	10.5%	89.5%	80.7%

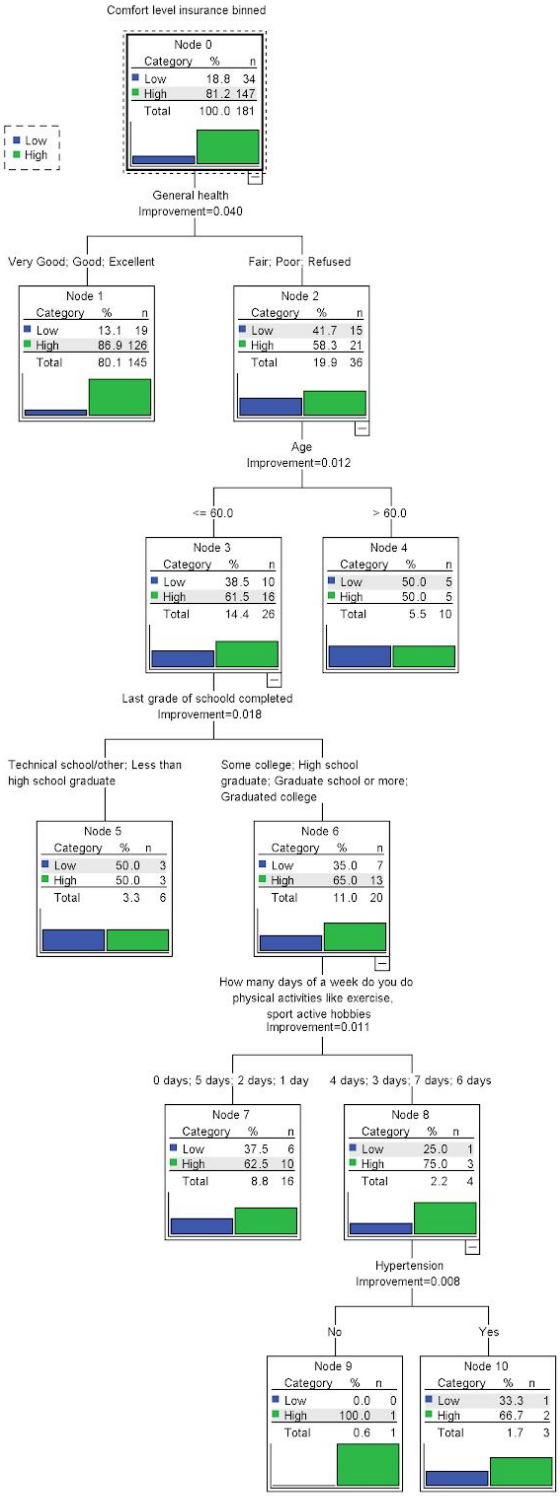
Growing Method: CRT
Dependent Variable: Comfort level reasearching insurance online binned

Based on the table below, it is seen that General Health (how do you feel about your health) is the most influential variable, followed by education level, marital status, exercise, and hypertension. This model predicts very well on High Confidence level in the comfortability of researching insurance online.

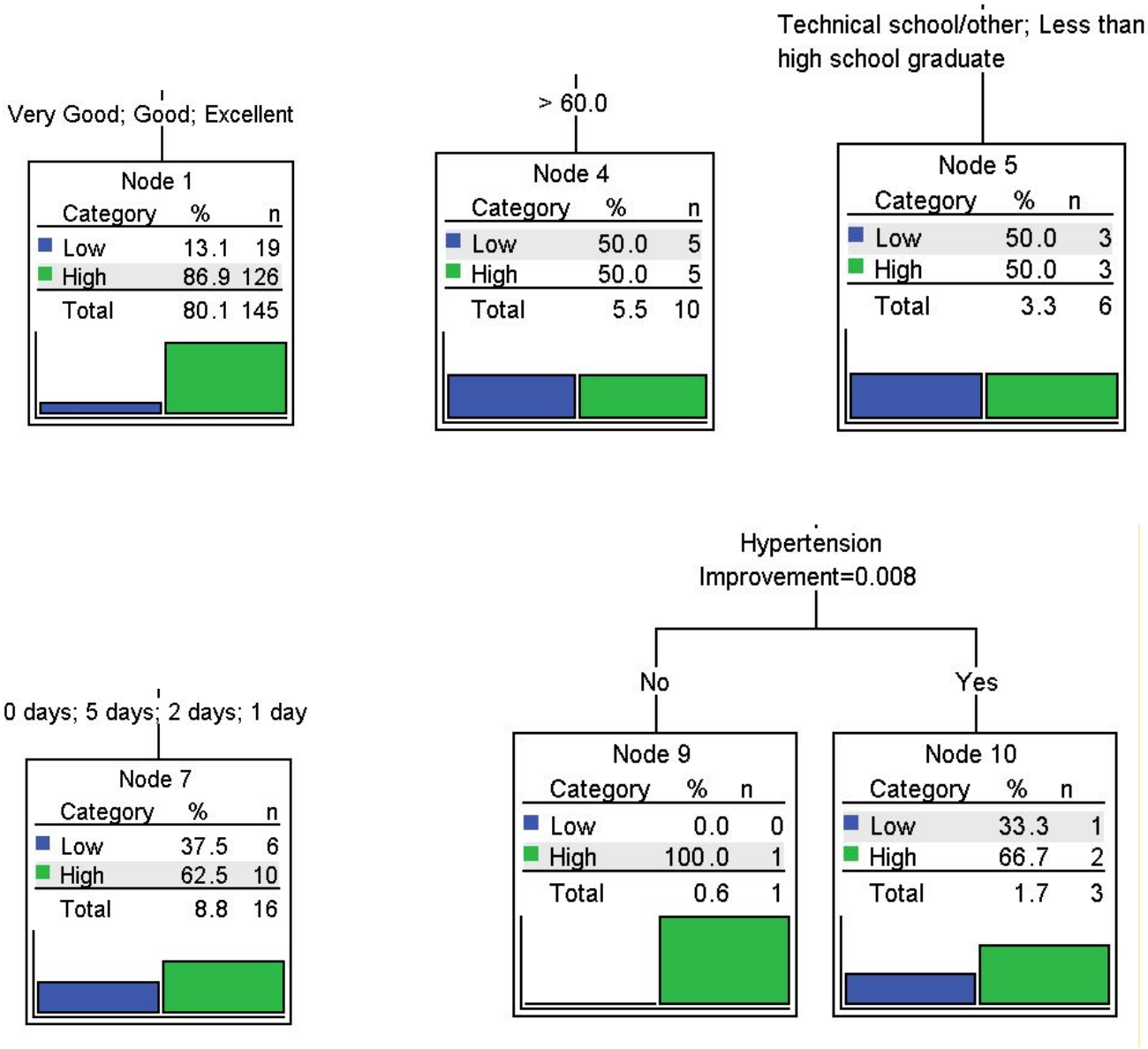
Independent Variable Importance

Independent Variable	Importance	Normalized Importance
General health	.043	100.0%
Last grade of schoold completed	.021	47.6%
Marital status	.018	41.5%
How many days of a week do you do physical activities like exercise, sport active hobbies	.016	37.2%
Hypertension	.015	35.1%
Employment status	.015	34.3%
Age	.015	33.9%
Annual household income	.008	18.1%
Cancer	.007	16.7%
How many days during last 30 days your health was not good	.006	13.4%
Body mass index	.005	10.7%

The accompanying decision tree that was produced by the model did not have many branches. It was a fairly simple tree.



The tree produced 6 terminal nodes, the most important was Node 1, which could explain 86.8% for High (126 tuples out of 181) and 13.1% for Low for general health equal to very good and excellent



Decision rules created by the tree for High/Low confidence level researching insurance online were as follow:

1. If a person has a very good and excellent health then the model can predict 86.9 High and 13.1 Low confidence level.
2. If a person refused to answer or had fair or poor general health and was over 60 years old, then the model could predict 50% for High and 50% of Low.

3. If a person refused to answer or had fair or poor general health and was less than 60 years old and had technical school or less than high school then the model could predict 50% for High and 50% of Low.
4. If a person refused to answer or had fair or poor general health and was less than 60 years old and had some college or high school graduate or graduate school or more then the model could predict 65% for High and 35% for Low.
5. If a person refused to answer or had fair or poor general health and was less than 60 years old and had some college or high school graduate or graduate school or more and exercised 0,1,2,5 days per week then the model could predict 62.5% for High and 37.5% for Low.
6. If a person refused to answer or had fair or poor general health and was less than 60 years old and had some college or high school graduate or graduate school or more and exercised 3,4,6,7 days per week and had no hypertension then the model could predict 100% for High and 0% for Low.
7. If a person refused to answer or had fair or poor general health and was less than 60 years old and had some college or high school graduate or graduate school or more and exercised 3,4,6,7 days per week and had a hypertension then the model could predict 66.7% for High and 33.3% for Low.

At the end we run confusion matrix to find out model measures.

Comfort level reasearching insurance online binned * Predicted Value Crosstabulation

Count

		Predicted Value		
		.00	1.00	Total
Comfort level reasearching insurance online binned	Low	47	87	134
	High	18	399	417
Total		65	486	551

TN(0,0) = 47, TP(1,1) = 399, FN(1,0) = 18, FP(0,1) = 87.

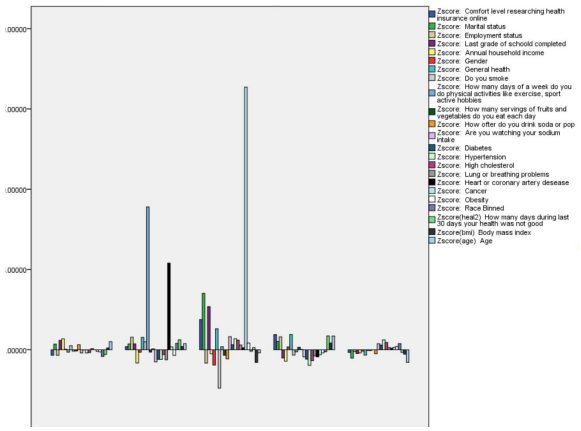
Accuracy (TP+TN)/ALL = 399+47/551 = 0.79
Sensitivity/Recall TP/P = TP/(TP+FN) = 399/(399+18) = 0.95
Specificity TN/N = TN/(TN+FP) = 47/(47+87) = 0.35
Precision TP/(TP+FP) = 399/(399+87) = 0.82
F-measure = 2 x Precision x Recall / (Precision + Recall) = 2*0.82*0.95/(0.82+0.95) = 1.558/1.77 = 0.88

Based on above measures, especially low specificity and high sensitivity, our Decision Tree model could predict well true positives and could not predict very vell true negatives. It possibly could be used by a marketing company to target people with high confidence in researching insurance online, which have a good health to switch to different insurance company offering better rates or better terms.

In general we were not satisfied with Decision Tree model and we decided to also do a Cluster Analysis.

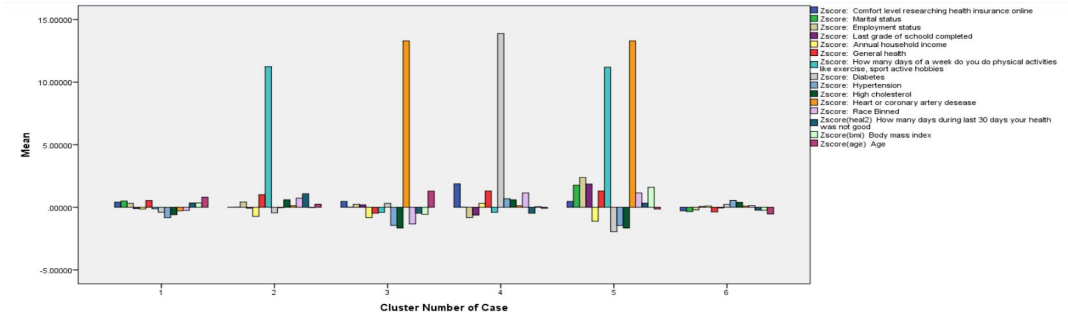
Cluster Analysis

After the decision tree analysis, to further this project cluster analysis was used. The decision trees analysis showed that knowledge might not have had that strong of an effect on other factors but that other factors might have a strong effect on knowledge. The first attempt at cluster analysis showed a the elbow having a large spike. This led to some of the variables to be removed with one of them being cancer. Five clusters were chosen from the graph and the cluster bar was used to help decide what variables to drop. These variables were dropped for various reasons one of which was redundant data.

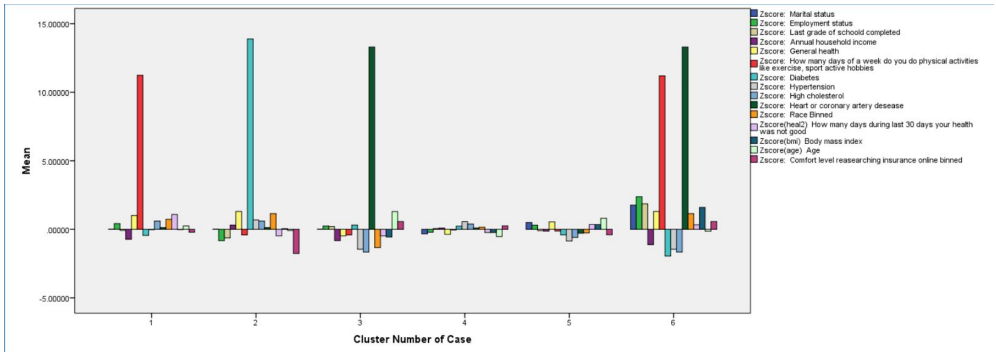


During the second attempt at the clusters, the data was taken a little further. In addition to reducing variables, two sets of analyses were done with two different interpretations of the Y variable. The unbinned Likert Scale 1 equaling high comfort to 7 equaling uncomfortable. The binned binary variable as used in the previous decision tree analyses categorized those with low to high comfort. In both cases for the second cluster analysis attempt the elbow graph was soother than the first attempt. Both of these graphs were relatively similar to each other and six clusters were chosen for each seeing that was where the elbow was located. The following two graphs show the second cluster attempt elbow.

The following graph shows the cluster bar unbinned. It the native format the comfort is from high to low comfort which means that a positive in this chart actually means it is negative. Looking at this chart for unbinned it is seen that diabetes, heart disease, and coronary artery disease have a very low comfort level.



The following graph shows the cluster bar binned. This graph has comfort as the last variable. The binary version is based on having a mid-high comfort level. This graph has a true relationship when negative is negative and positive is positive.



In deciding which clusters to use for this project, different factors were taken into place. In a typical cluster analysis the small clusters are the more interesting. This project in particular, the smaller clusters were too small meaning there weren't enough cases to use for the decision trees. Both binned and unbinned each had two large clusters. These clusters corresponded with higher or lower comfort. The decision tree analyses were ran based on these four clusters.

There were four variables that showed to have the highest importance in each of the four clusters. These variables were age, hypertension, employment, and high cholesterol. Following these three variables were a mix of disease, health, and marital status. The follow charts show the variable importance in each of the four clusters.

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
Age	.236	100.0%
Hypertension	.214	90.4%
Employment status	.188	79.4%
High cholesterol	.117	49.7%
Marital status	.094	39.8%
General health	.077	32.5%
Diabetes	.067	28.4%
Heart or coronary artery desease	.053	22.3%
How many days during last 30 days your health was not good	.046	19.6%
Body mass index	.034	14.2%
How many days of a week do you do physical activities like exercise, sport active hobbies	.022	9.2%
Race Binned	.013	5.7%
Annual household income	.011	4.5%
Last grade of schoold completed	.003	1.3%

Growing Method: CRT
Dependent Variable: Un_High

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
Age	.240	100.0%
Hypertension	.223	92.9%
Employment status	.180	75.3%
High cholesterol	.131	54.6%
Marital status	.100	41.7%
General health	.095	39.5%
Diabetes	.092	38.2%
Heart or coronary artery desease	.065	27.3%
How many days during last 30 days your health was not good	.058	24.2%
Body mass index	.041	17.0%
How many days of a week do you do physical activities like exercise, sport active hobbies	.034	14.0%
Annual household income	.022	9.1%
Last grade of schoold completed	.006	2.5%
Race Binned	.005	2.1%

Growing Method: CRT
Dependent Variable: Bin_High

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
Age	.243	100.0%
Hypertension	.219	90.4%
Employment status	.188	77.3%
High cholesterol	.139	57.1%
Diabetes	.137	56.6%
General health	.100	41.1%
Marital status	.084	34.6%
How many days during last 30 days your health was not good	.071	29.4%
Heart or coronary artery desease	.069	28.3%
Body mass index	.038	15.6%
How many days of a week do you do physical activities like exercise, sport active hobbies	.036	14.9%
Annual household income	.026	10.8%
Last grade of schoold completed	.004	1.5%
Race Binned	.001	0.4%

Growing Method: CRT
Dependent Variable: Un_Low

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
Age	.227	100.0%
Hypertension	.217	95.6%
Employment status	.183	80.7%
High cholesterol	.129	56.7%
Marital status	.095	42.0%
Diabetes	.085	37.2%
General health	.079	34.7%
How many days during last 30 days your health was not good	.063	27.8%
Heart or coronary artery desease	.061	26.7%
Body mass index	.033	14.7%
How many days of a week do you do physical activities like exercise, sport active hobbies	.024	10.6%
Annual household income	.014	6.3%
Last grade of schoold completed	.009	4.2%
Race Binned	.007	3.1%

Growing Method: CRT
Dependent Variable: Bin_Low

Conclusion

The decision tree did not predict the Low Confidence correctly for this project. The High Confidence in researching insurance online would not be helpful for the purpose of this project. However, the High Confidence levels that were found could be useful in other areas such as sociology or some other group study. The most important variable for predicting High Confidence was an individual's on feelings and if they felt good about their own health. This factor makes logical sense. When exploring this project in more depth in the future the decision tree model would need to be improved.

The cluster analysis looked at both the binned and unbinned data. This data was then sorted into high and low levels of comfort. Classification percentages were good for what was needed on the project but should the analysis be ran again, some more variables should be taken away to get strong percentages. The binned variables gave smoother data. The binned high showed more tiers when it came to the decision trees. The more tiers gave more options which were stronger correlations than the unbinned.

All four of the decision trees using the cluster analysis had the same three most important variables. These variables were age, hypertension, and employment status. Age makes sense as being the most important variable in being comfortable with researching insurance. Younger individuals have the option to be on their parents health care until the age of 26. They tend to be less informed with what is out there and how to research health insurance. Older individuals have more life experience and tend to be sick more often or have some other sort of health ailment such as joint replacements. These individuals would be more versed when it comes to health insurance.

Hypertension was the highest disease variable and the second overall variable for the importance. There is a positive correlation between stress and hypertension. Stress can also be related to age and employment.

Employment was the third most important variable seen by the cluster analyses. Individuals with full time jobs can have the option to get their health insurance coverage through their employer. These individuals would (assumed to) have researched the health insurance plans that were made available to them prior to signing up with their employer. Those individuals that fall under the unemployed or part time employees would not have this option and would be left to research health insurance plans on their own.

Chronic illness individuals require more trips to a physician's office and more medical monitoring. And example of a chronic illness would be diabetes. Seeing as these individuals require more medical attention, they would be better versed in healthcare coverage to cover their needs.

Ultimately, the findings from this project were that knowledge does not have that big of an effect on health and health insurance. On the other hand, health and health factors do play a major role in knowledge of health insurance. Furthermore, it shows the importance of running data through multiple processes, to see more of the bigger picture.

Citations

Arpey, Nicholas C. “How Socioeconomic Status Affects Patient Perceptions of HealthCare: A qualitative Study.” 2017. Journal of Primary Care & Community Health. 169-175 http://journals.sagepub.com/doi/full/10.1177/2150131917697439#articleCitationDownloadContainer

Falcon, Adolph. Healthy Americas Survey, 2014. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2016-08-31. https://doi.org/10.3886/ICPSR36433.v1

KCMU Analysis of 2014 National Health Interview Survey