Executive Summary

James Shea

11/3/2020

**The Problem and Hypothesis**

The purpose of this analysis is to determine if data collected on current health insurance customers is sufficient to predict customer interest in acquiring automotive insurance from the same insurance company. A Generalized Logistic Model finds relationships between predictor variables and the response variable (Date, 2020). It is hypothesized that a Generalized Logistic Model can be developed using the customer data provided as explanatory variables in order to predict interest in automotive insurance, the outcome variable. A logistic regression model can be helpful in estimating the target variable by determining the strength of predictor variables with respect to loan data (Martinson, 2020). If a relationship between the health insurance customer data and interest in automotive insurance is found, a model to predict future response given acquired data should be possible. (Statistical Consulting Group, 2020).

**The Data-Analysis Process**

The data set was made publicly available on Kaggle.com by Anmol Kumar who was provided the data set by a health insurance company (Kumar, 2020). There are over 300,000 rows and it is necessary to have a sufficiently large quantity of rows for a Logistic Regression Model to fit well (Austin & Steyerberg, 2015). The data set was provided as a CSV file and was imported into SAS as a SAS data set. Data was cleaned by removing unnecessary variables, checking for missing values, and verifying values. No missing data was detected, which is a benefit as missing data can reduce the model accuracy (Swalin, 2018). Univariate, bivariate, and normality tests were performed and found that the data was not normally distributed. A Generalized Logistic Model does not require a test for normality (University of Colorado Boulder, 2018). The data set was split into training and testing data sets. Stepwise logistic regression with fast backward elimination was used to build a logistic model from the training data set. Four variables were included in the model to predict interest in automotive insurance. The ROC curve for the model showed an area under the curve of .8454. The model was then applied to the testing data and predicted 87.8% of observations correctly.

**The Findings**

The accuracy of the logistic regression model was higher on the testing data set than the training set. While surprising, this indicates that the model is generalized and can be used to predict interest in automotive insurance using data from current health insurance customers. The analysis suggests that interest in automotive insurance can be predicted from the variables Vehicle\_Damage, Policy\_Sales\_Channel, Previously\_Insured, and Age.

**Limitations**

One limitation is that the data set only contains observations from a single health insurance company, thus, any problems that occurred during the data collection phase would be persistent throughout the data set. Another limitation is that all data was gathered at a single point in time. Another limitation is that the data set only included observations from the United States.

**Proposed Actions**

The first proposed action for a future study is to include additional predictor variables that could better predict interest in automotive insurance as only four of the variables included in the current data set were included in the final model. Next, including data from multiple health insurance companies could improve the generalizability of the model. Third, gathering data over a longer period of time could reveal potential changes in interest due to seasonal or other time related differences. Finally, gathering data from customers in other countries could generalize the model to a larger segment of the total population.

**Expected Benefits of the Study**

The first expected benefit of this study is that it will allow the prediction of interest in automotive insurance for current health insurance customers. Second, companies will be able to utilize this information to target the correct customers who are likely to be interested in automotive insurance rather than all customers. Third, customers also benefit as a result of customers who are unlikely to be interested to not be targeted, while those who are likely to be interested are targeted. Finally, this analysis will help identify which factors are likely to predict interest in automotive insurance.

**Sources**

Analyzing Non-Normal Data with Generalized Linear Models (GLMs). (2018). Retrieved October 18, 2020, from https://www.colorado.edu/lab/lisa/services/short-courses/analyzing-non-normal-data-generalized-linear-models-glms

Austin, P., & Steyerberg, E. (2015, January 22). The number of subjects per variable required in linear regression analyses. Retrieved October 13, 2020, from https://www.sciencedirect.com/science/article/pii/S0895435615000141

Coding Systems for Categorical Variables in Regression Analysis. (n.d.). Retrieved October 22, 2020, from https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis-2/

Date, S. (2020, May 10). Generalized Linear Models. Retrieved October 13, 2020, from https://towardsdatascience.com/generalized-linear-models-9ec4dfe3dc3f

Generalized linear model. (2020, September 26). Retrieved October 13, 2020, from https://en.wikipedia.org/wiki/Generalized\_linear\_model

Kumar, A. (2020, August). Health Insurance Cross Sell Prediction. Retrieved October 29, 2020, from https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction

Swalin, A. (2018, March 19). How to Handle Missing Data. Retrieved October 13, 2020, from https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

What is Linear Regression? (n.d.). Retrieved October 13, 2020, from https://www.statisticssolutions.com/what-is-linear-regression/