The University of North Carolina at Charlotte

DATA ANALYSIS TO PREDICT TRAVELERS' INTEREST IN TRAVEL INSURANCE

Final Course Group Project Report

Course: Big Data Analytics for Competitive Advantage

Authors:

Hope Grubb (hgrubb2@uncc.edu), Lenka Raslova (lraslova@uncc.edu), Sean Oberer (soberer@uncc.edu), Connor Brown (cbrow316@uncc.edu)

Charlotte, 2021

Introduction

The main problem that consumer-based industries, such as the insurance industry, face is the initial attainment and then retainment of clients. Building out a strong client base is dire for the success of a consumer-driven company because studies show client attainment is more difficult than simply keeping the individual as a client over time. Therefore the goal of this analysis is to determine the model best fit for predicting whether an individual will become a client for a travel insurance agency or not.

Background/Literature review

Travel insurance is becoming increasingly popular among people who travel as there are more risk factors now than ever with the COVID-19 pandemic. Travel insurance plans cover a variety of different things such as trip cancellations, health insurance, and other health related emergencies, lost bags, and so on (*What is travel insurance and what does it cover, n.d.)*. However, travel insurance plan prices can vary depending on the amount of people covered on a plan, the ages of those covered under a plan, where someone is traveling, and how many days someone is traveling for (*How much is travel insurance? it's more affordable than you think, n.d.)*. An insurance company named Battleface conducted a survey to see how many people had to pay for additional costs or losses when they traveled, if someone would purchase travel insurance when traveling domestically, or if someone would pay for

travel insurance when traveling internationally (Kagan, 2021). That survey found that nearly half of Americans had paid additional costs or losses when traveling. Additionally, it also found that 34% of Americans would purchase travel insurance when traveling domestically and 36% of Americans would purchase travel insurance when traveling internationally (Kagan, 2021). With travel insurance becoming increasingly popular especially with the COVID-19 pandemic, insurance companies need to know who they should target based on relative factors such as a customer's demographics and how often/much they travel for their travel insurance plans.

Data Description

Our dataset was sourced from Kaggle.com and the dataset name is Travel Insurance Prediction Data. It contains information about travel insurance consumers, some of these customers are open to travel insurance, and some are not.

The data is of cross-sectional type and has a size of 1,986 observations. The data file was uploaded to Kaggle on 09/07/2021 for study purposes. The description of the data file states that the data comes from the Performance/Sales of The Travel Insurance Package in the year 2019. No further information about the data file has been published. For this reason, the method of data collection is not known. However, it can be assumed that the data is of a primary type. The data in the file is commonly used for an analysis of which customers would be interested in buying new travel insurance from its travel company.

The data contains 10 variables - 4 quantitative and 6 categorical. Quantitative variables include the variable Age which is the age of the customer where the

minimum age is 25 and the maximum is 35. Further, the annual income of the customer in Indian rupees (variable AnualIncome), number of family members living with the customer where the minimum number is 2 and the maximum is 9 (variable Family Members), and variable Index which is the customer identification number.

Categorical variables include Employment Type which indicates where the customer is employed. It has two values – Private sector/Self-employed and Government Sector. Another variable is Graduate Or Not which indicates if the customer is a college graduate or not. Chronic Diseases indicates if the customer has a chronic disease or not. Further, categorical variables include whether the customer frequently books air tickets (Frequent Flyer), whether the customer has ever traveled to a foreign country (Ever Traveled Abroad), and if the customer has ever bought travel insurance (Travel Insurance).

Data Preparation

Our dataset does not contain any missing values. We used the 3 sigma rule for the detection outliers in quantitative variables and we did not find any.

We also checked if our dataset is imbalanced. We found out that 1,277 customers did not purchase travel insurance and 710 customers purchased travel insurance (the total number of customers was 1,987). Based on that, we concluded that the difference between the numbers of positive and negative samples is a case of imbalanced data. To balance data, we used the basic sampling method over-sampling. This method involves making exact copies of minority class samples (Burez & Van den Poel, 2009). We randomly picked 500 samples from positive

samples and added them again to the database. The balanced dataset consists of 2,487 observations (negative - 1,277 and positive - 1,210).

Methods

For purposes of our research problem, we decided to use two models - Decision Tree and Logistic Regression. The logistic regression model was also broken out into two separate models, one that utilizes backward selection and the other using stepwise selection.

Logistic Regression

For achieving our business problem, we decided to use logistic regression. Our dependent variable is whether the customer bought the travel insurance or not. We used Age, Employment type, Graduate or not, Annual income, Family members, Chronic Diseases, Frequent flyer, and Ever Traveled Abroad as independent variables.

We divided our data into two parts - Training and Validation. Training data consists of 70 % of data and Validation consists of 30 % of data. Training data was used for building models, validation data was used for the evaluation of the model.

We used SAS Enterprise Miner for building the regression model. Specifically, we built two regression models. The first model was based on Backward selection and the second one was based on Stepwise selection. For both models, we used the selection criterion Validation Error.

First, we ran these models with imbalanced data. The backward regression model scored 76.55 % on accuracy, sensitivity 47.73 %, and specificity 93.21 % (positive samples - 214, negative samples - 383). The stepwise regression model scored 76.88 % on accuracy, sensitivity 47.66 %, and specificity 93.21 % (positive samples - 214, negative samples - 383). In both models, the sensitivities were too low. After using balanced data, the sensitivities improved. Therefore, we decided to use just models with balanced data.

The first model - Backward

The first model based on the Backward method consists of Age, Annual Income, Chronic Diseases, Employment Type, Ever Traveled Abroad, Family Members, and Frequent flyer as independent variables. Only the variables Age, Annual Income, Ever Traveled Abroad, Family Members, and Frequent flyer are individually significant on a 5 % level of significance. Variables Chronic Diseases and Employment Type are not individually significant on a 5 % level of significance.

The first regression model - Backward

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-4.3909	0.6237	49.56	<.0001		0.012
Age		1	0.0912	0.0189	23.31	<.0001	0.1510	1.095
AnnualIncome		1	1.372E-6	1.775E-7	59.70	<.0001	0.2864	1.000
ChronicDiseases	0	1	-0.0562	0.0607	0.86	0.3543		0.945
EmploymentType	Government Sector	1	-0.0667	0.0652	1.05	0.3061		0.935
EverTravelledAbroad	No	1	-0.8310	0.0844	97.01	<.0001		0.436
FamilyMembers		1	0.1994	0.0338	34.77	<.0001	0.1800	1.221
FrequentFlyer	No	1	-0.3059	0.0713	18.41	<.0001		0.736

Classification Table		Actual Class	
		Positive	Negative
Predicted Class	Positive	241	74
	Negative	123	310

The accuracy of the second model is 73.66 %, the sensitivity 66.21 %, and the specificity is 80.73 % (positive samples - 364, negative samples - 384).

The second model - Stepwise

The second model based on the Stepwise method consists of Age, Annual Income, Ever Traveled Abroad, and Family Members as independent variables. All of these variables are individually significant on a 5 % level of significance.

The second regression model - Stepwise

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-4.5466	0.6144	54.75	<.0001		0.011
Age	1	0.0845	0.0185	20.81	<.0001	0.1400	1.088
AnnualIncome	1	1.606E-6	1.664E-7	93.08	<.0001	0.3352	1.000
EverTravelledAbroad No	1	-0.8507	0.0836	103.43	<.0001		0.427
FamilyMembers	1	0.1944	0.0337	33.30	<.0001	0.1755	1.215

Classification Table		Actual Class	
		Positive	Negative
Predicted Class	Positive	239	69
	Negative	125	315

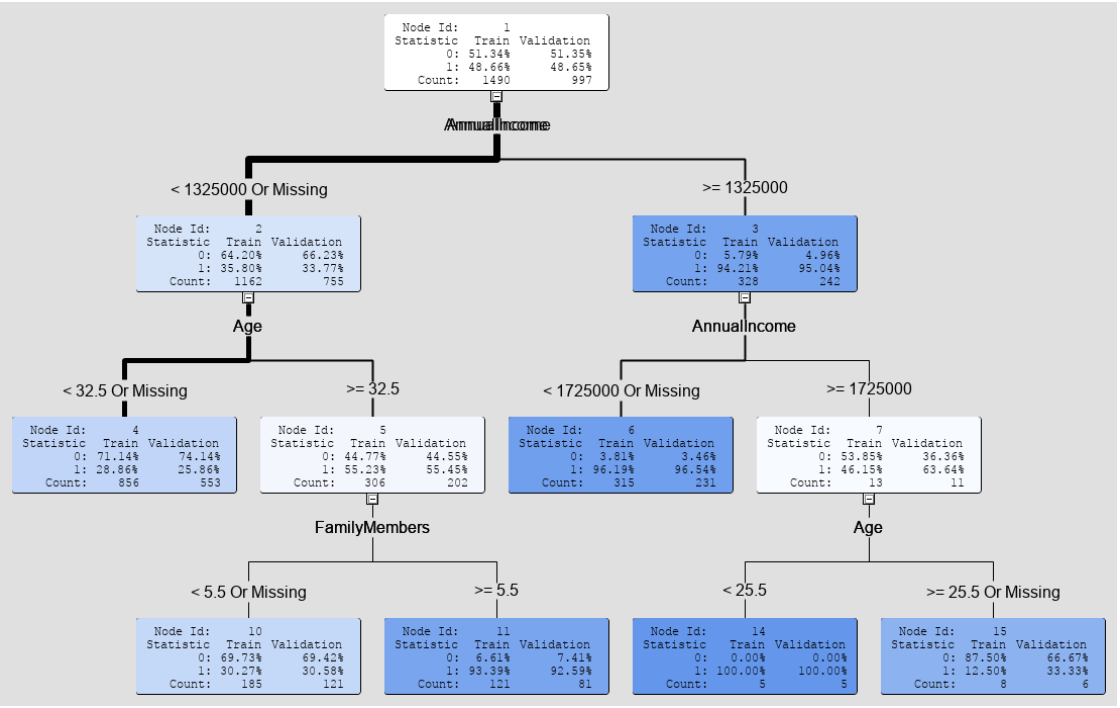
The accuracy of the second model is 74.06 %, the sensitivity 65.66 %, and the specificity is 82.03 % (positive samples - 364, negative samples - 384).

Decision Tree

As stated above, the other model we decided to use for this analysis is a decision tree. The dependent/target variable for this model is whether or not a customer purchased travel insurance. The independent variables for this model used are Age, Annual Income, Chronic Diseases, Employment Type, Ever Travelled Abroad, Family Members, Frequent Flyer, and Graduate or Not.

We divided our dataset into two parts for this model - training and validation. We used 60% of our dataset for training and 40% of our dataset for validation. The training part of the dataset is used to train our model and our validation dataset is used to determine how accurate our predictions are.

The decision tree



Based on the model above, the most influential variables for whether someone has purchased travel insurance or not are Annual Income, Age, and FamilyMembers. From the decision tree above, we predict that only 25.86% of people younger than 32.5 with an income of less than 1,325,000 rupees will purchase travel insurance. Additionally, only 30.58% of people older than 32.5, have an annual income of 1,325,000 rupees or less, and with less than 5.5 family members will purchase travel insurance. Furthermore, we find that 92.59% of people with an income less than 1,325,000 rupees older than 32.5, and have a family size of 5.5 people or more will purchase travel insurance.

On the other side of our decision tree, we predicted that 96.54% of people with an annual income greater than 1,325,000 rupees but less than 1,725,000 rupees will purchase travel insurance. Additionally, 100% of people younger than 25.5 with an annual income greater than 1,725,000 rupees will purchase travel insurance. Lastly, only 33.33% of people older than 25.5 with an income greater than 1,725,000 rupees will purchase travel insurance.

Below, we have our confusion matrix for our decision tree.

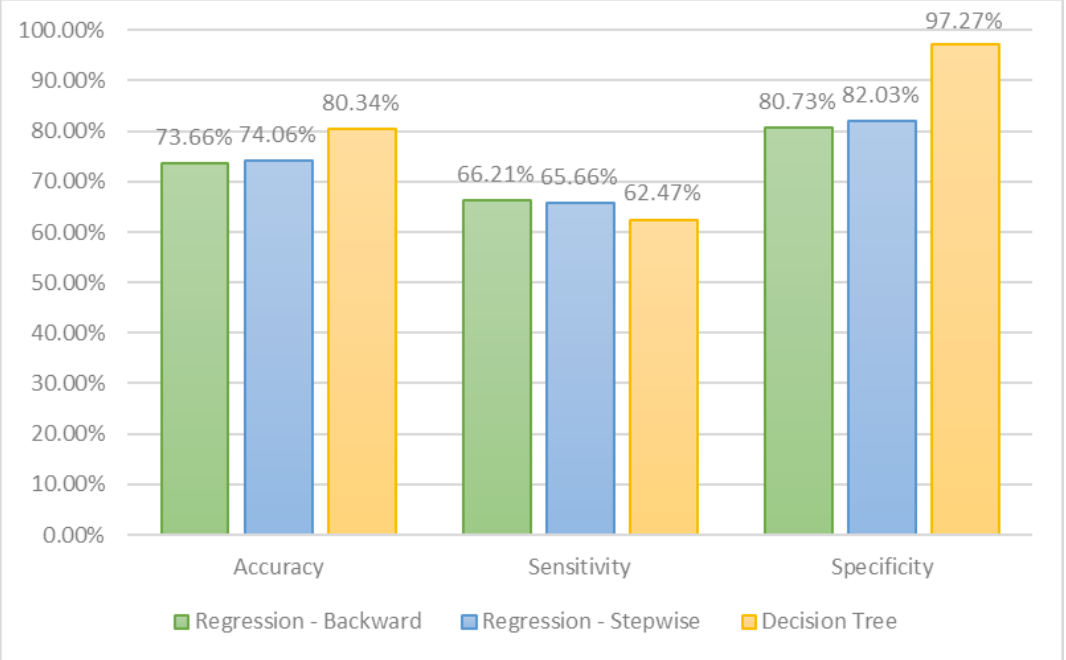
Classification Table		**Actual Class**	
		Positive	Negative
Predicted Class	Positive	303	14
	Negative	182	498

Based on the confusion matrix of our decision tree, our model has an accuracy of 80.34%. Additionally, the sensitivity of our model is 62.47%. Lastly, the specificity of our model is 97.27% (positive samples - 485, negative samples - 512).

Model Evaluation

All of the models were similar in performance, but the top performer in accuracy was the decision tree model meaning the percentage of total correct predictions when compared to the total number of predictions was the highest. The decision tree model was also right at 97% accurate in predicting who would *not* buy insurance. Which for this specific research problem isn't relevant, but can provide a base to potentially show an opportunity for expansion by reaching out to another target population of possible clients. The backward selection regression model earned the highest percentage in sensitivity, or the number of true positives compared to the number of actual positives meaning that it did the best job in predicting who *will* purchase travel insurance. Since the main goal of our research and analysis is to predict *who* will buy travel insurance, the model best fit would be the backward regression model as it can accurately predict the highest number of potential clients. Though, the backward regression model was only a few percentages away from the stepwise model that also performed relatively well in accuracy and sensitivity while surpassing the backward model in specificity, or the true negatives compared to the actual negatives.

Comparison of models



Results

Backward Regression

Our first regression model, backward regression, gave us five different non-correlated statistically significant variables. It showed that generally, there is a positive correlation between age and someone's likelihood of purchasing travel insurance. This could be due to the fact that the older someone is, the more likely they are to have experienced a travel disaster, which would likely make someone purchase travel insurance. The second significant variable was annual income. Our model showed that there was a positive correlation between annual income and someone's likelihood to purchase travel insurance. This makes sense because if you

have the money to protect your investment in your travel, you'd probably like to. If you were, however, already breaking the bank to afford the trip, you'd be less likely to splurge on travel insurance. The third significant variable was whether or not someone had ever traveled abroad. The model indicated that someone never having traveled abroad decreased their chances of purchasing travel insurance. Trips abroad are generally more complicated and have more moving parts than domestic travel, and a problem traveling to a place outside of your home country not only has the potential to be worse, but more frightening as well. The fourth significant variable this model gave us was family members. This is likely due to the fact that a larger family would mean that more people would be going on the trip and because of that it would be more expensive, incentivising someone to protect the money they spent on their travel. The final significant variable was the frequent flyer variable. This variable indicates that someone is less likely to purchase travel insurance if they are not a frequent flyer. This could be explained by frequent flyers taking more trips than non frequent flyers, meaning they spend more money on travel and are more likely to want to protect it.

Stepwise Regression

Our stepwise regression model used four variables, and showed that they were all statistically significant. The first variable was age, and showed a similar, but slightly lower correlation to someone's likelihood of purchasing travel insurance as the previous model. The second variable was annual income which, like the previous model, showed a positive correlation to someone's likelihood of purchasing travel

insurance. This correlation, however, was stronger than in the previous model. The third variable was the ever traveled abroad variable. Like the previously mentioned variables in this model, the correlation between this variable and someone's likelihood of purchasing travel insurance was the same sign as the previous model. This correlation was notably stronger, however. The fourth variable was the number of family members. The correlation between this variable and someone's likelihood of purchasing traveling insurance was positive, and in fact, identical to the previous model.

Decision Tree

The decision tree model showed, in accordance with the other two models, that annual income, number of family members and age were the most important variables in predicting whether or not someone would purchase travel insurance. Key findings show a few things. First, that people with income above the lower threshold, 1,325,000 rupees, were overwhelmingly likely to purchase travel insurance, however, if they were below the higher income threshold they were actually far less likely to purchase travel insurance. Second, large families below the lower income threshold are almost certain to purchase travel insurance. Third, older people are more likely to purchase travel insurance, with the notable exception of very young, wealthy people who are the most likely of any group to purchase travel insurance.

Discussion

Business Suggestions

The primary goal of our research was to understand the factors that influenced someone buying travel insurance. Our models generally concluded that age, annual income, and family size were the biggest influential factors. A few business suggestions came out of our research. 1. Young, wealthy people are a good group to market towards. 2. Big families are a good group to market towards. 3. Middle class people are a strong group to market towards. Beyond those three marketing based suggestions, we have a few suggestions to help target different market groups. 1. Create plans to target lower income customers. 2. Create plans to target elderly customers

Limitations

Our research was impacted by a few minor limitations. Namely, the data was rather unbalanced originally, and required oversampling. To continue this research, we would need new data. Our other limitation was the limited demographic variables we had at our disposal. To continue this research, we would prefer to have variables like gender, marital status, and job sector beyond whether or not the customer worked in the government.

Contributions of each group member to the project and project report

Data analysis, results section, discussion section - Connor

Data analysis, Background/Literature review, Decision tree - Sean

Data analysis, Data Description, Logistic regression - Lenka

Data analysis, Introduction, Model Evaluation - Hope

References

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3), 4626-4636.

How much is travel insurance? it's more affordable than you think. Berkshire Hathaway Travel Protection. (n.d.). Retrieved December 5, 2021, from https://www.bhtp.com/how-much-does-travel-insurance-cost.

Kagan, J. (2021, September 15). *What is Travel Insurance?* Investopedia. Retrieved October 5, 2021, from https://www.investopedia.com/terms/t/travel-insurance.asp.

What is travel insurance and what does it cover? Nationwide. (n.d.). Retrieved December 5, 2021, from https://www.nationwide.com/lc/resources/home/articles/what-is-travel-insurance.

Data

Travel Insurance Prediction Data | Kaggle. Kaggle: Your Machine Learning and Data Science Community [online]. Copyright © Original Authors [cit. 04.10.2021]. Available from: https://www.kaggle.com/tejashvi14/travel-insurance-prediction-data