IST 707

Final Project – Caravan Insurance Prediction

James D. Harris and Jessica Kwon

11 December 2018

**Project Description**

This project comes from the Computational Intelligence and Learning Cluster challenge from the year 2000. The challenge was to predict who would be interested in buying a caravan insurance policy and give an explanation why. The data supplied to participants came from a Dutch data mining company, Sentient Machine Research.

**Data Description**

The dataset is hereby described:

* 9,822 customer records
* 43 socio-demographic variables
* 43 insurance-related variables, including the target variable
* Training/test split (5,522/4,000 respectively)
* All data already discretized
* Positive cases of the target variable represent only 6% of the total cases
* A data dictionary is available at http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/data.html

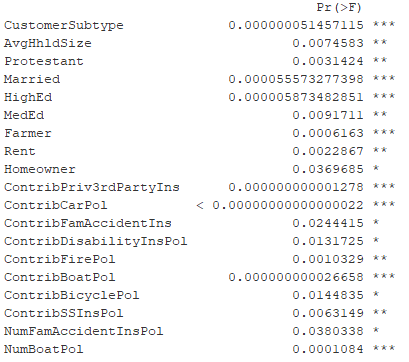
**Process Description**

The goal of this project is to be able to accurately predict the target variable, identified as “CARAVANNumMobileHomePol,” which reflects if the customer did (1) or did not (0) own a caravan insurance policy. As there are 85 potential predictors, we decided to run a regression analysis to get a better idea of which attributes were more closely related to the target variable. Following that, we decided to run a variety of models and compare them to see which one offered the greatest accuracy in predicting the caravan insurance policy. Our analyses are shown in the following sections.

Note: for the analyses that include a confusion matrix as the method of evaluating the model, we have focused on balanced accuracy as the evaluation metric, due to the unbalanced nature of the dataset.

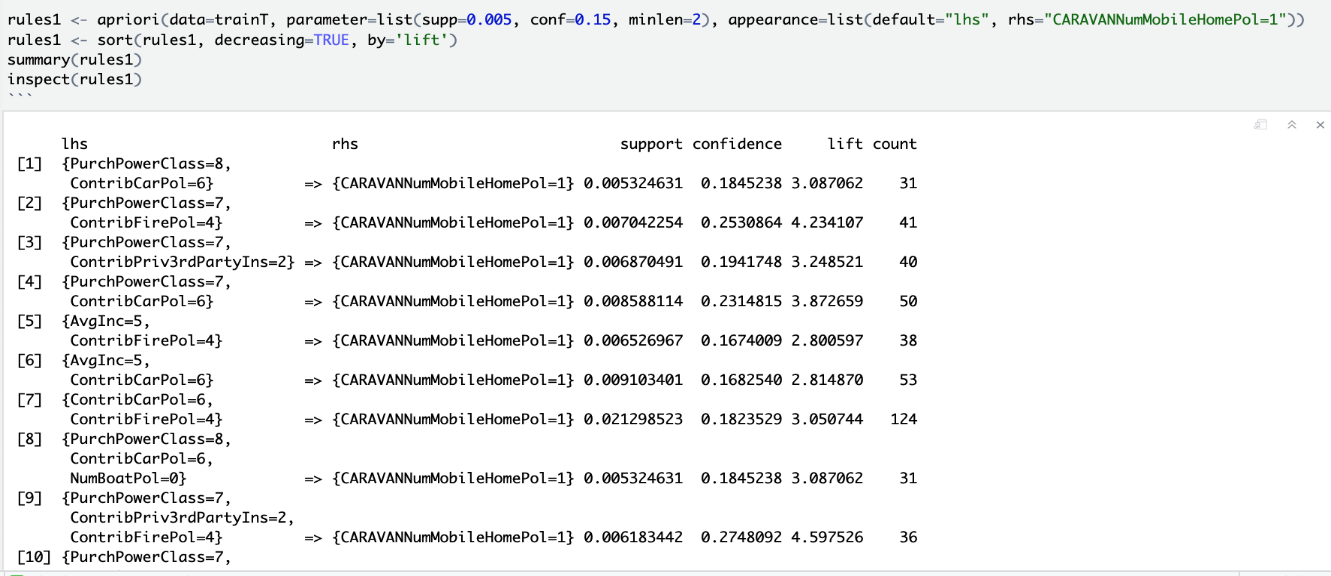
**Regression analysis**

We conducted this analysis mainly to get a better understanding of the data. Of particular interest was the variable ‘ContribCarPol,’ or how much a customer paid into a car insurance policy. This variable was prominent in other analyses. The table that follows is a representation of all variables that were statistically significant (p < 0.05):



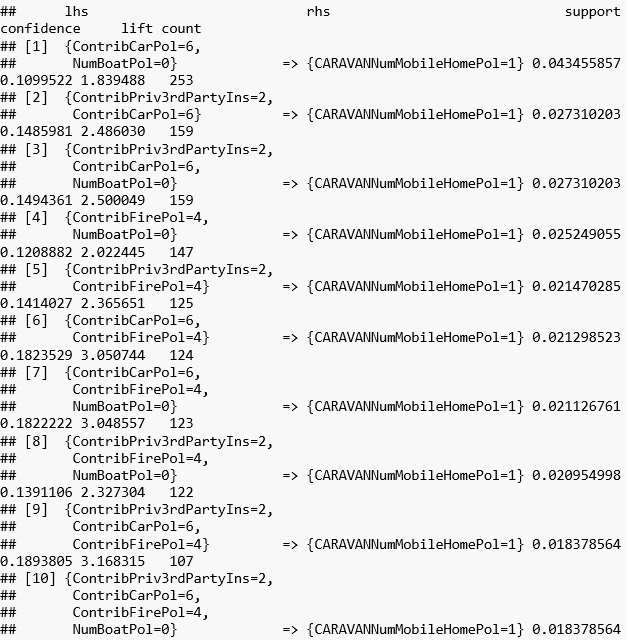
**Association Rule Mining**

This analysis was ideal due to the discrete nature of the values in the data set. The relatively low number of positive cases in the target variable meant support for this set of rules had to be set quite low, but after trying various combinations, we found a set of 5,447 rules with strong lift. For this analysis, support was set to 0.005, confidence to 0.15, minimum length 2, and the target variable as the right hand side.



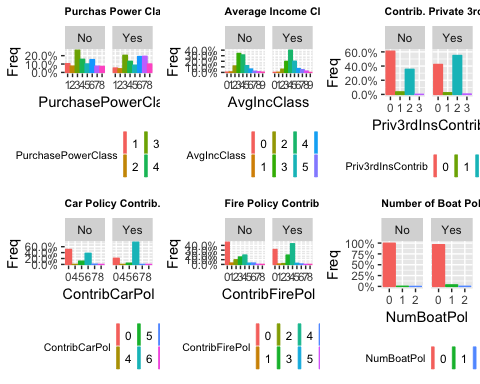
We tried a variety of settings. When we selected the following, we generated a list of 28,144 rules (top 10 shown):

Support = 0.006, conf = 0.1, minlen = 3, target variable = right hand side.



From this analysis, we identified six variables that served as strong predictors of the target variable. These variables were average income, purchasing power class, contributions to third party insurance, contributions to a car insurance policy, contributions to a fire insurance policy, and number of boat insurance policies.

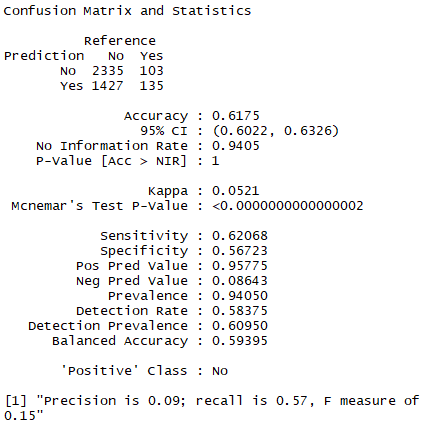
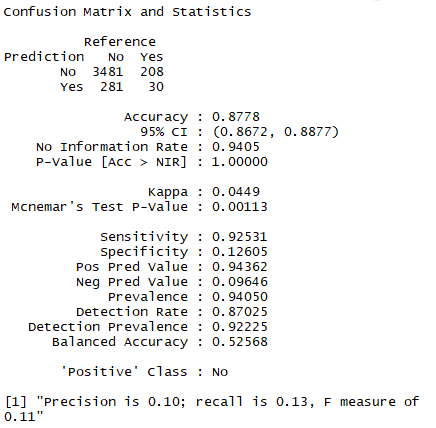
We checked the distribution of these six variables:



**Naïve Bayes**

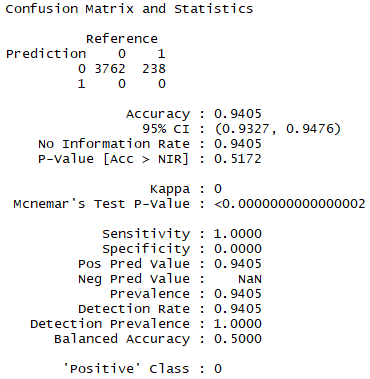
The naïve Bayes method is a classification method that assumes that the predictors are independent of each other. We ran the model twice: first with all the variables, and then with only the statistically significant variables in the previous section:

All variables: Significant variables only:



Balanced accuracy: **52.568%** Balanced accuracy: **59.395%**

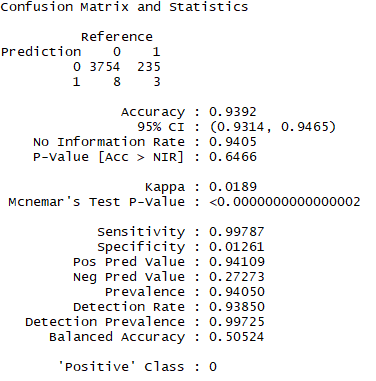
**Decision Tree**



Balanced accuracy: **50%**

The decision tree model predicted a negative outcome 100% of the time.

**Random Forest**

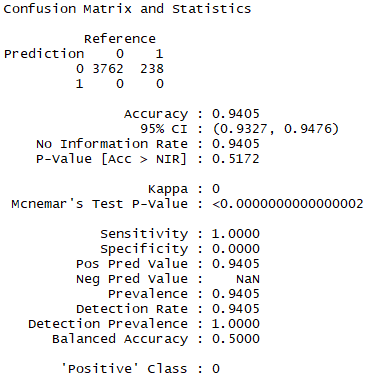
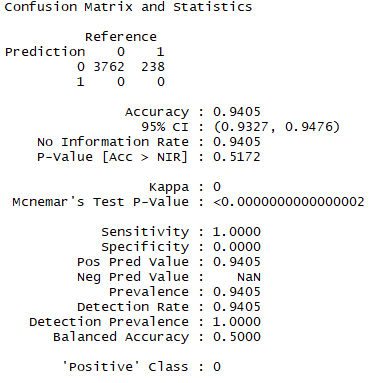


Balanced accuracy: **50.524%**

The random forest model predicted a positive outcome only 11 times out of 4,000, when there were 238 actual incidences of a positive outcome.

**Support Vector Machines**

Linear: Radial:



In using support vector machines, it is clear that the linear and radial models are identical. Moreover, while the SVM model boasts an accuracy rate of 94.05%, it arrived at that conclusion by predicting a negative outcome for every case, and so adds no usable information to the project. For both versions, the balanced accuracy rate is **50%**, the lowest of any of our models.

**Conclusion**

After using regression and association rule mining to identify strong predictors, it is apparent that the naïve Bayes model, using those predictors, is the most accurate model, with a balance accuracy of 59.395%. Other methods, especially the decision tree and support vector machine models, were not able to overcome the relatively low number of positive cases in the dataset in order to give a reliable prediction.

The variable that served as the strongest predictor was the contribution to a car insurance policy—specifically when that contribution was in category 6 or higher ($1,000 or higher). Purchasing power class was a strong influencer when it had a value of 5 or higher, but that value was not reflected on the data dictionary, so its real-world meaning is unclear. Other strong predictors included the purchase of social security insurance, boat insurance, and fire insurance.

The owner of the dataset could find this information actionable, specifically regarding targeted advertising. The customers who purchased caravan insurance were wealthier than most, and carried a greater variety of insurance, often at higher contributions than the average.