**Final-Term Project sxa151231(Sunny Anand)**

**Title: To build and evaluate a model for predicting a target variable (whether they own a mobile home insurance policy)**

**Purpose:** This is the final-term project which focuses on performing end-to-end activities on the given dataset and concentrates on model building and evaluation of the model once the model is built. The goal is to build an accurate classification model that will predict the target attribute.

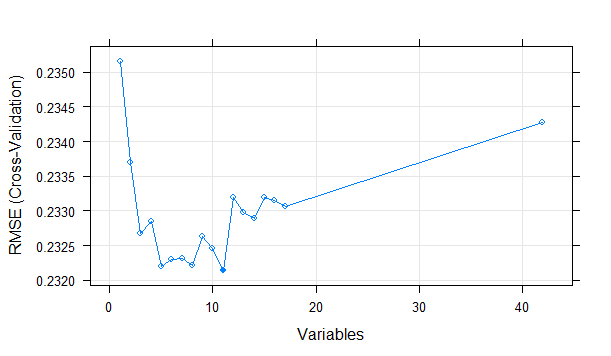
**Dataset(s):** There are several data files: A training set (TICDATA2000.txt) which contains predictors and a target value, a test set (TICEVAL2000.txt) which does not contain a target value, and a set of target values for the test set (TICTGTS2000.txt). The datasets, as well as explanatory information: <https://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+%28COIL+2000%29>

**Pre-processing on the given dataset:**

* Load the 3 datasets that is the training, test and target attribute dataset.
* This dataset has no missing data. There are no NA’s in this dataset as well.
* There are 3 categories of features in this dataset. Continuous, Nominal and Ordinal. The Nominal and Ordinal variables need to be converted to factors and then be used to create a new dataframe consisting of either factors or integers.
* The newly created dataset seems to have few fields which are highly consisting of 0’s. I decided to keep some of them as they did not contribute to a large variance in the spread of the data.
* Overall the data set was very consistent with regard to pre-processing and it was now ready to apply the next steps of feature selection.

**Feature Selection**- Finding the set of best attributes for building the model. As the number of attributes were large I decided to apply a few techniques to remove attributes which did not carry much information and removing them would in turn help find variables which actually do play a role in predicting the target variable.

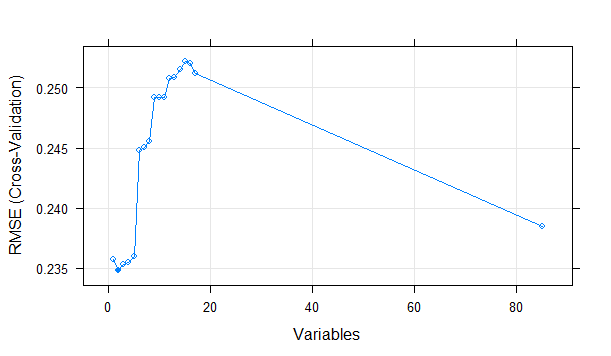
* The next important thing is to apply feature selection techniques in order to determine the best possible features. A popular automatic method for feature selection provided by the caret R package is called Recursive Feature Elimination or RFE. A Random Forest algorithm is used on each iteration to evaluate the model. The algorithm is configured to explore all possible subsets of the attributes. All 85 attributes are selected in this example, although in the plot showing the accuracy of the different attribute subset sizes, we can see that just 5 attributes gives almost comparable results. Please note I have run these in 2 sets : first from attributes 1-43 and second from 44-86. The graphs below are for the 2 sets.



The top 5 variables (out of 11):

V47, V44, V59, V65, V82

"V47" "V44" "V59" "V65" "V82" "V49" "V62" "V61" "V70" "V85" "V68"

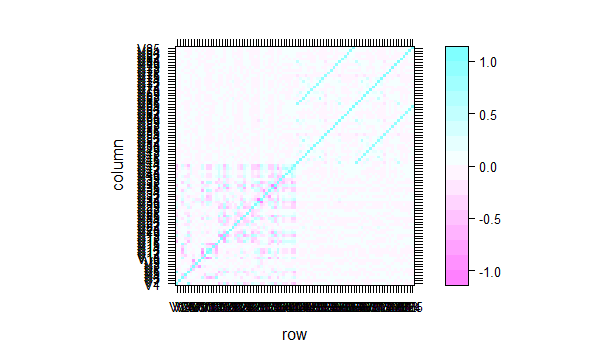


"V82" is an important feature using rfe technique from the caret package from second run.

* I will now check for co-linearity of features to remove the features which are overlapping in model impact. This will help us to make the feature selection better for a better model. I applied the idea of removing all those variables which are more than 90% correlated. Below is the list of variables which were removed by this approach.

Most closely related attributes which we will discard now.

V36 V31 V1 V52 V46 V65 V74 V48 V72 V48 V54 V85 V79 V71 V78 V62 V83 V70



* Each record consists of 86 attributes, containing sociodemographic data (attribute 1-43) and product ownership (attributes 44-86).The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same sociodemographic attributes. Attribute 86, "CARAVAN:Number of mobile home policies", is the target variable. The set of collected variables using the collinarity and feature selection technique eliminates all the varibales based on the sociodemographic attributes(1-43).
* The most important features which are selected are the ones below:
  + PBRAND, MOSHOOFD, MOSTYPE,PPERSAUT and APERSAUT
  + All the variable V59+V47+V82+V65+V68

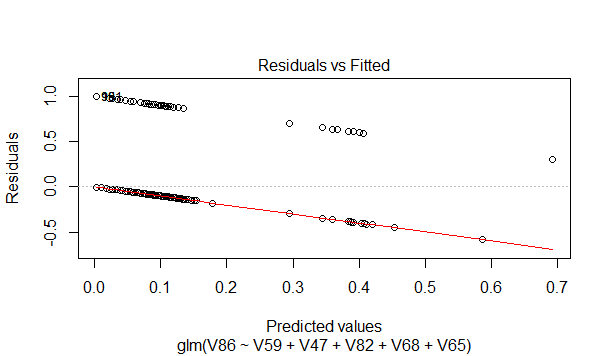
**MODEL BUILDING**

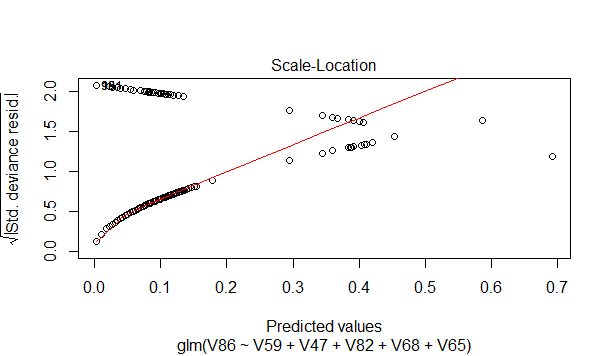
**MODEL#1,2,3 Based on Linear Regression**

* The first model is a linear regression model built on the train data. Based on this simple linear model it was enough for us to determine which of the features are not statistically significant using their p-values and the overall model p-value. The model consisted of high R2 value and small p-value.
* Since a continuous prediction model with categorical features present I used the glm() method to create the model.
* summary(fit)
* Call:
* glm(formula = V86 ~ V59 + V47 + V82 + V68 + V65, data = train\_data)
* Deviance Residuals:
* Min 1Q Median 3Q Max
* -0.58667 -0.08843 -0.05347 -0.00329 0.99671
* Coefficients:
* Estimate Std. Error t value Pr(>|t|)
* (Intercept) 0.003288 0.005188 0.634 0.526283
* V59 0.007610 0.001858 4.097 4.25e-05 \*\*\*
* V47 0.009685 0.002608 3.714 0.000206 \*\*\*
* V82 0.291692 0.037360 7.807 6.87e-15 \*\*\*
* V68 0.007289 0.012591 0.579 0.562679
* V65 0.019747 0.007121 2.773 0.005570 \*\*
* ---
* Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
* (Dispersion parameter for gaussian family taken to be 0.05397058)
* Null deviance: 327.20 on 5821 degrees of freedom
* Residual deviance: 313.89 on 5816 degrees of freedom
* AIC: -466.14
* Number of Fisher Scoring iterations: 2

It is important to note that V68 and V65 are not very significant in making the prediction. Also the prediction for this model was not very good below 60 people which where going to opt for the insurance. I made three linear models to finally decide that the most important attributes were V59, V47 and V82.

**Graphs for the base model:**





**MODEL#2 based on Naïve Bayes**

* From the simple linear model the information I got about the more informant attributes were clear. V86~ V47+ V59+ V82
* The result of this model was better than the glm() model. This predicted 84 customers to buy the insurance.
* I further tweaked this formula to only test with the formula . V86~ V47+ V59 and it gave much better result as 104.

Below are the confusion matrix for the 2 models.

model <- naiveBayes(V86 ~ V59 + V47 , data = train\_data)

> table(predict(model, test\_data),test\_data\_res$V1)

0 1

0 3658 238

1 104 0

model <- naiveBayes(V86 ~ V59 +V47 + V82, data = train\_data)

> table(predict(model, test\_data),test\_data\_res$V1)

0 1

0 3676 238

1 86 0

**Summary and final analysis of the result**

Naive Bayes gave a very good result in comparison to basic linear models. Though I could have applied more complex techniques to this approach and I did. I used the neuralnet() however the results were not as good as the simple and easy to understand naïve bayes and so I decided to drop it in my final report writeup. I strongly believe after manual analysis of the test data that naïve bayes is indeed a very good classifier on this dataset. I had the idea of using interactions but as I was not able to figure out a way to do interactions for V47 and V59 I decided to let this be my final model. I did check on the deviation if I had used the interactions and it seemed to be promising. Hopefully this is still a very good result for the train data and the test data.