Predictive Data modeling for health care: Comparative performance study of different prediction models

Shivanand Hiremath hiremat.nitie@gmail.com

National Institute of Industrial Engineering (NITIE) Vihar Lake Mumbai

Rahul Gupta

National Institute of Industrial Engineering (NITIE) Vihar Lake Mumbai

Abstract

Data mining techniques have been successfully applied in numerous fields. and predictive analytics and machine learning algorithms have been commercially utilized. Health care sector has recently seen a surge in the use of predictive analytics due to accuracy of the predictive and prescription models for the heart diseases. Out of all the fields in the healthcare domain, insurance attains a far more commercial importance. The dataset used for this paper has a target variable which indicates whether a person will buy insurance or not.

The paper delves on the performance of different popular predictive models to identify the most suitable data model for predicting the same. Confusion matrix is generated for 9 data models for comparison and overall error of each data model is used to decide the most suitable data model for the insurance dataset.

Keywords

Confusion Matrix, ROC Curve

Introduction

Insurance is a very competitive field. For insurance products to succeed, it is essential to predict its attractiveness to prospective customers. Data mining has the potential to identify the market acceptance of an insurance product. This can help insurance companies design a better portfolio of insurance products. For example, a prediction model to predict the probability of a car accident happening within a particular span of time based on customer data can help insurance companies to arrive at the pricing for their products. Also, a good prediction model based on the hospitalization needs can help the insurance companies to devise more relevant products for their prospective customers. The dataset used in this paper helps the insurance company to identify customers more likely to buy their products. Thus, a better and targeted marketing plan could be developed to attract those customers.

Predictive Data mining

Two most common modeling techniques are classification and prediction. Classification models predict categorical variables (discrete, unordered) while prediction models predict continuous – valued functions.

Decision trees and neural networks use classification algorithms while regression, association rules and clustering use prediction algorithms. Naïve Bayes algorithm is used to create models with predictive capabilities and it learns from the “evidence’ by calculations the correlation between the dependent and the independent variables.

Neural networks involve three layers viz; input, hidden and output units. Connection between input units, hidden and output units are based on relevance of the assigned value (weight) of that particular input unit, the higher the weight the better the network.

Data Mining techniques were applied to Health Care Data by (Obenshain M K infect controlHosp 2004)

P. van der Putten. M. van Someren (eds).andCharles Elkan(2, 3) (2000) discussed the Naïve Bayesian classifiers with CoIL Challenge 2000 data. But, the most important issue for any such model would be its accuracy. We have built some 9 popular data models and validated the accuracy of each data model to arrive at the best model.

Data Source

Dataset used in the analysis is from THE INSURANCE COMPANY (TIC) 2000 (c) Sentient Machine Research 2000. This dataset has been used to train and validate prediction models and build a description (5822 customer records). Each record consists of 86 attributes, containing socio-demographic data (attribute 1-43) and product ownership (attributes 44-86) data. The socio-demographic data is derived from zip codes. All customers living in areas with the same zip code have the same socio-demographic attributes. Attribute 86 named “CARAVAN: Number of mobile home policies", is the target variable.

Target Variable: CARAVAN - Number of mobile home policies

Input Variables Description

Customer Subtype, Number of houses, Average size household, Average age, Customer main type
Religion Details, Relationship Details, Children Details, Education Details, Working Class,
Social Class, House Ownership, No of Cars, National Health Service, Private health insurance,
Income Category, Average income, Purchasing power class, Insurance Contribution, Number of third party insurances

Figure 1 - Dataset Description

Methodology

Rattle is a Graphical User Interface tool for Data Mining. Rattle can be used to present statistical and visual summaries of data. It can also be used for transformation and build both unsupervised and supervised models.

We have utilized Rattle to generate the data models and then to evaluate the generated data models for their accuracy. This was done by creating confusion matrices.

Confusion Matrix in Data Mining: For models with two values of the depended attribute, these counts are false positives and negatives.

Findings

Run time is calculated for the generated models and formulae are generated.

<u>Naive Bayes</u>

naiveBayes.default(x = ticdata2000 [, 1:85], y = ticdata2000 [, 86]) **(1)**

```
A-priori probabilities:
Ticdata2000 [, 86]
         0          1
0.94022673 0.05977327
```

Time taken: 1.20mins

Decision Tree

```
rpart(formula = CARAVAN ~ ., data = crs$dataset[crs$train, c(crs$input, crs$target)], weights =
(crs$dataset$MOSTYPE)[crs$train], method = "class", parms = list(split = "information"),
control = rpart.control(usesurrogate = 0,maxsurrogate = 0))                                (2)
```

Time taken: 0.39 seconds

Random Forest

```
randomForest(formula = as.factor(CARAVAN) ~ ., data = crs$dataset[crs$sample, c(crs$input,
crs$target)][rep(row.names(crs$dataset[crs$sample, c(crs$input, crs$target)]),
as.integer(eval(parse(text = "crs$dataset$MOSTYPE"))[crs$sample])), ],ntree = 500, mtry = 9,
sampsize = c(100), importance = TRUE, replace = FALSE, na.action = na.roughfix)            (3)
```

Time taken: 4.20 minutes

Ada Boost

```
ada(CARAVAN~.,data=crs$dataset[crs$train,c(crs$input,crs$target)][rep(row.names(crs$dataset
[crs$train,c(crs$input,crs$target)]),as.integer(eval(parse(text="crs$dataset$MOSTYPE"))[crs$sm
ple])), ], control =rpart.control(maxdepth = 30, cp = 0.01, minsplit = 20, xval = 10), iter = 50) (4)
```

Time taken: 27.06 minutes

SVM (Support Vector Machine)

```
Support Vector Machine object of class "ksvm"
SV type: C-svc (classification)
parameter: cost C = 1
Gaussian Radial Basis kernel function.
Hyperparameter : sigma =  0.0103569394589033
Number of Support Vectors : 7923
Objective Function Value: -5746.787
Training error: 0.021609
Probability model included.
```

Time taken: 1.20 hours

Logistic Regression

```
  glm(formula = CARAVAN ~ ., family = binomial (link = "logit"), data = crs$dataset[crs$train,
c(crs$input, crs$target)], weights = (crs$dataset$MOSTYPE)[crs$train])                    (5)
```

Time taken: 50.75 seconds

<u>Neural Network:</u>Time taken: 2.03 seconds

<u>Recursive Partition:</u> Time taken: 0.36secs

Overall Error calculated using Confusion Matrix (Error Matrix)

```
Error matrix format:
      Predicted
Actual    0 1 Error
     0 0.93 0     0
     1 0.07 0     1
Overall error: 0.0744559
```

Table 1 - Overall Error observed from Error Matrix for each algorithm

Naive Bayes	Decision Tree	Random Forest	Ada Boost	Support Vector Machine	Logistic Regression	Neural Network
6.2%	7.4%	7.4%	7.4%	8.9%	14.8%	7.4%

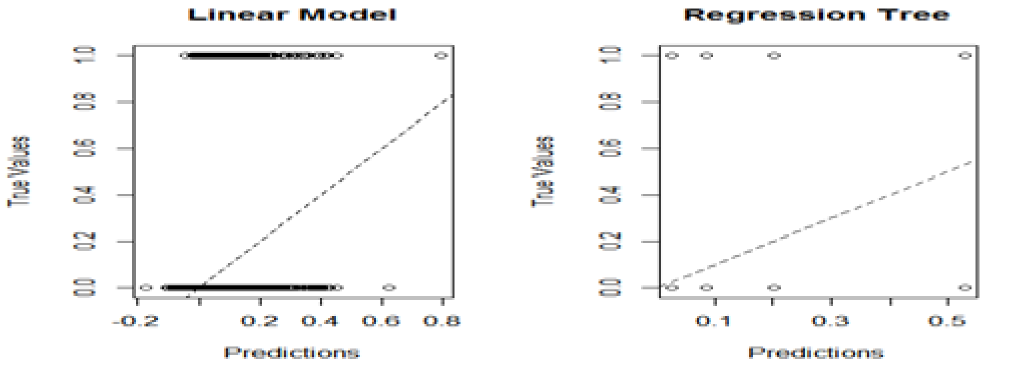

Figure 2 - Error Plot for Linear Model & Regression Tree Models

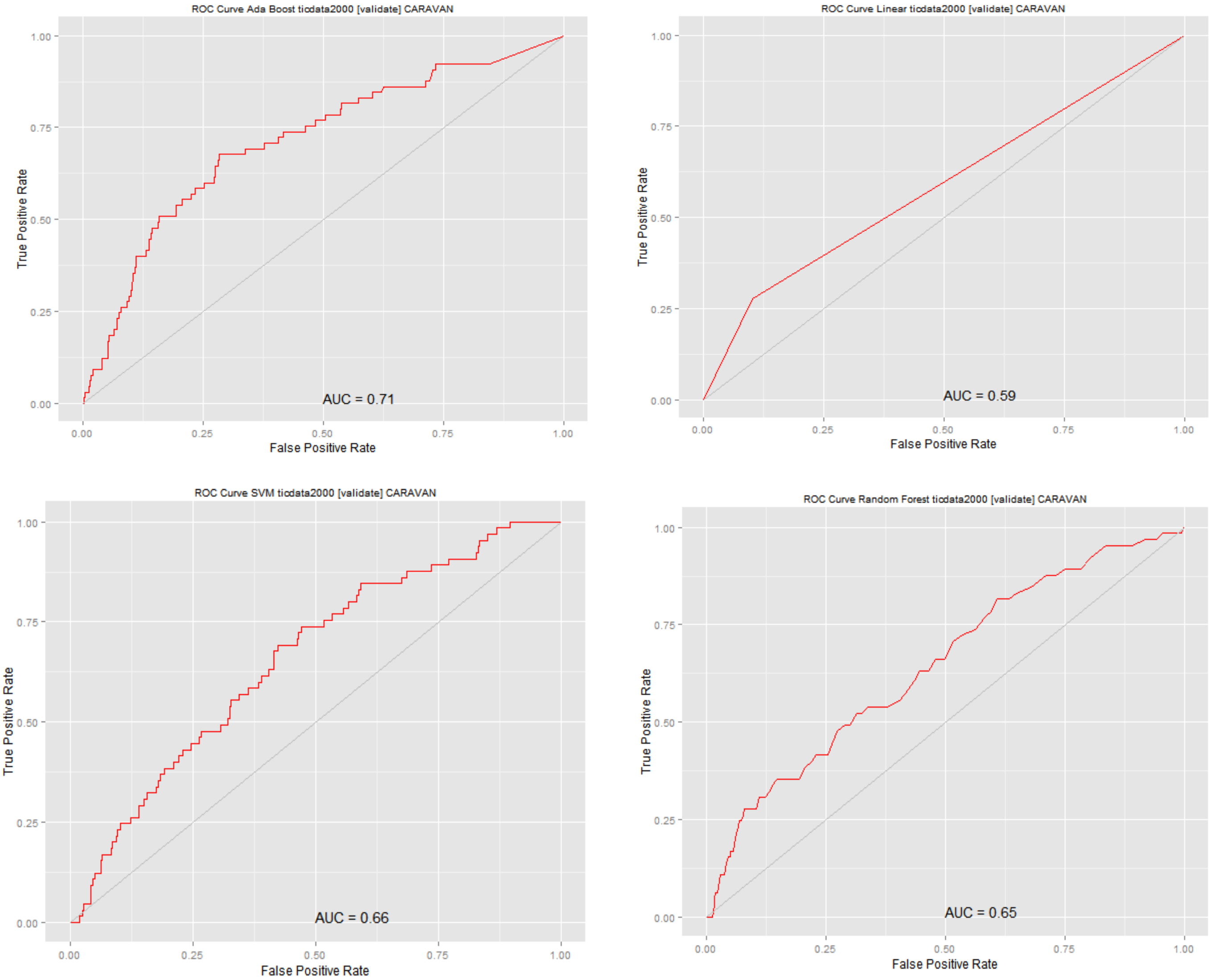

Figure 3 - ROC curves for various prediction models

Result

From the above observations in table 1, it is clear that Naïve Bayes classifier works best for the dataset. This doesn't come as a surprise as other models suffer from over-fitting. Also, the error plot in Figure 2 clearly depicts that the predicted values more closely follow the true values of the target variable for regression tree as compared to a linear model.

Figure 3 shows ROC curves for some of the prediction models. The area under the curve (AUC) is a good estimate of the accuracy of the predictions. This area measures discrimination of the model which signifies its ability to correctly classify an observation, i.e., whether the person will buy the policy or not. The AUC observed from the ROC directly correlates to the error obtained from the confusion matrix as depicted in table 1. For example, the AUC for ada boost model is 0.71 while its overall error obtained from the confusion matrix is 7.4%. Similarly, for SVM model, the error is 8.9% while the AUC from ROC curve is 0.66. Thus, higher the observed error from the confusion matrix, lower will be the area under the ROC curve. This further validates the result we obtained from the error (confusion) matrix.

Conclusion and Future Work

In this paper, a comparative performance analysis of various data mining techniques has been done for insurance data. Naïve Bayes classifier performs the best, though other algorithms results are also quite close. Logistic Regression, on the other hand, gave the worst result.

The proposed work could be further enhanced by eliminating certain attributes (by backward elimination) to improve the result of predictive analytics. This will help us decrease bias and get a more accurate prediction.

Acknowledgement

Authors thank THE INSURANCE COMPANY (TIC) 2000 (c) Sentient Machine Research 2000 for their data.

References

Obenshain, M.K. Application of Data Mining Techniques to Healthcare Data, Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

P. van der Putten and M. van Someren (eds). CoIL Challenge 2000. The Insurance Company Case. Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report 2000-09. June 22, 2000

Charles Elkan. Magical Thinking in Data Mining:Lessons from CoIL Challenge 2000

Rattle: A Graphical User Interface for Data Mining using R. Available at http://rattle.togaware.com (accessed date February 1, 2015)

Edgar, Acuna and Members of the CASTLE group at UPR-Mayaguez, (2009).dprep: Data preprocessing and visualization functions for classification. R package version

Leo Breiman, (2001). Random forests. Machine Learning. 45, 5-32, 2001

Shawe-Taylor, J. and Cristianini, N. (2000). An introduction to support vector machines. Cambridge University Press.

Model Assessment with ROC Curves- Lutz Hamel Department of Computer Science and Statistics University of Rhode Island USA

Kuhn, M. (2008). Building predictive models in R using the caret package, *Journal of Statistical Software*. Available at http://www.jstatsoft.org/v28/i05/(accessed date February 1, 2015)

Altman, D.G., Bland, J.M. (1994). Diagnostic tests 1: sensitivity and specificity, *British Medical Journal*, vol 308, 1552.

Altman, D.G., Bland, J.M. (1994). Diagnostic tests 2: predictive values, *British Medical Journal*, vol 309, 102.

Velez, D.R., et. al. (2008). A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction, *Genetic Epidemiology*, vol 4, 306.