Comparing the different data models and predicting future risk of a member using healthcare data

Biswas Lohani, Santosh Khanal and **Rabindra Bista**

Kathmandu University, Nepal

Date: 10-13-2015

Contents

Introduction

- Generation of billions of healthcare data
- Includes:
  - Patient eligibility, medical claim, pharmacy claim, biometric records, HRA records, and lab
- Utilize a scalable processing platform (Hadoop/Cascading) and distributed search platform (Elastic search) for storage
- Can prepare various algorithms to do supervised learning for various predictions

Problem Statement

- Huge amount of raw healthcare data
- Not utilized for any other analysis and prediction
- Utilization of these data can give meaningful information like
    - future risks of particular patient,
    - future diagnosis for particular member,
    - whether the patient will be admitted or not, and so on

Methodology

- Data -semi-time series
- Exploring existing algorithms that are suitable to such data
- Enhancing these algorithms to more closely meet the research
- Modeling our dataset to fit the algorithms' input
- Comparing the output of the algorithms and deciding the best fit for each of our prediction need

Identified Algorithms

- Random Forest
- Regression Tree
- Naïve Bayes

Random forest

- A powerful new approach to data exploration, data analysis, and predictive modelling
- An ensemble classifier using many decision tree models
- Can be used for classification or regression
- **Features:**
  - It runs efficiently on large data bases
  - It can handle thousands of input variables without variable deletion
  - It gives estimates of what variables are important in the classification

Results

- Since the sample size is small, for reliability 1000 trees are grown using mtry0=150
- Since look=100, the results are output every 100 trees in terms of percentage misclassified

Trees	Error Rate
100	2.47
200	2.47
300	2.47
400	1.23
500	1.23
600	1.23
700	1.23
800	1.23
900	1.23
1000	1.23

Depth	Trees	Accuracy
1	50	0.6550713
1	100	0.6779153
1	150	0.6799633
2	50	0.7000791
2	100	0.6984858
2	150	0.6886874
3	50	0.6838721
3	100	0.6992044
3	150	0.6976292

Advantages of random forests

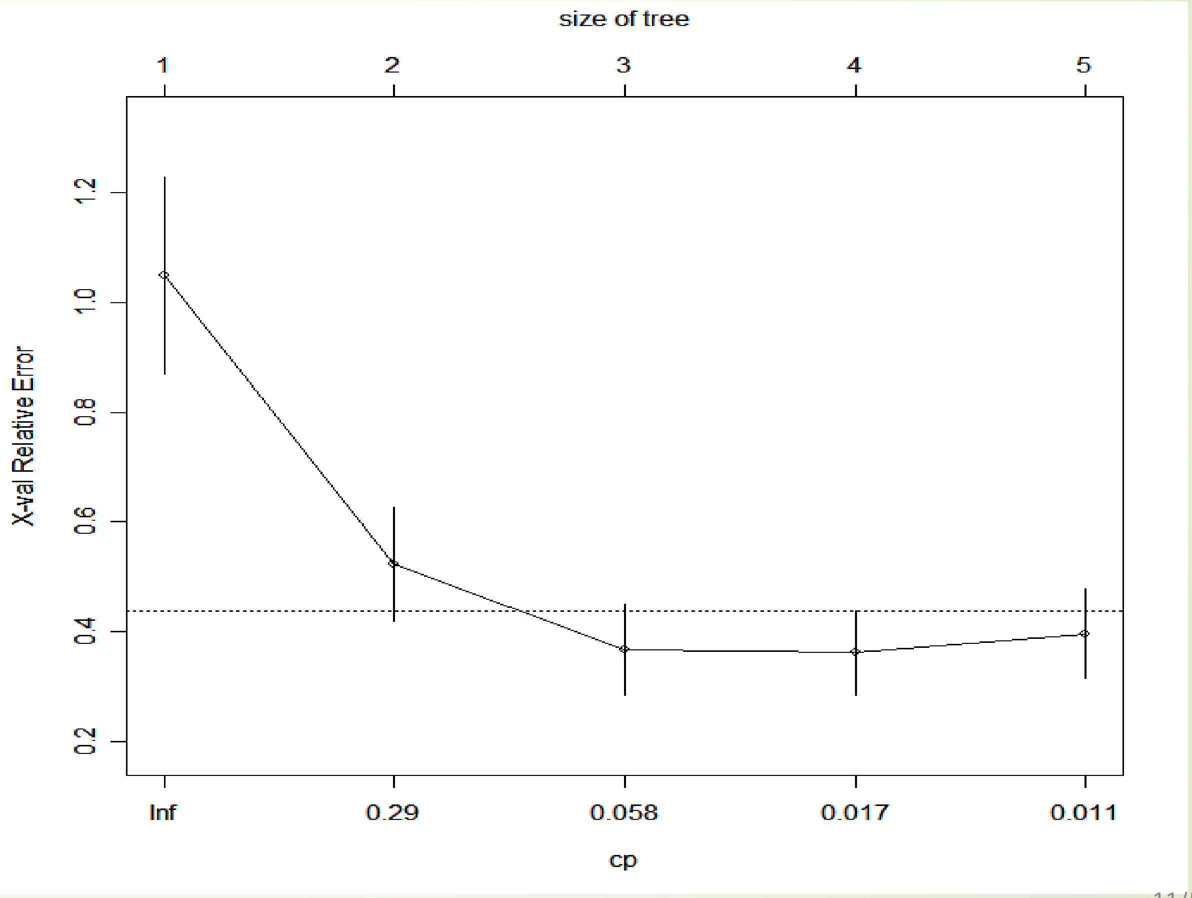
- No need for pruning trees
- Accuracy and variable importance generated automatically
- Over fitting is not a problem
- Not very sensitive to outliers in training data
- Easy to set parameters
- **Limitation:**
  - Regression can't predict beyond range in the training data
  - In regression extreme values are often not predicted accurately – underestimate highs and overestimate lows

Regression Trees

- A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility
- Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees
- Regression trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal

Results

```
# Regression Tree Example
library(rpart)

# grow tree
fit <- rpart(Mileage~Price + Country + Reliability + Type,
   method="anova", data=cu.summary)

printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits

# create additional plots
par(mfrow=c(1,2)) # two plots on one page
rsq.rpart(fit) # visualize cross-validation results

# plot tree
plot(fit, uniform=TRUE,
   main="Regression Tree for Mileage ")
text(fit, use.n=TRUE, all=TRUE, cex=.8)

# create attractive postcript plot of tree
post(fit, file = "c:/tree2.ps",
   title = "Regression Tree for Mileage ")
```

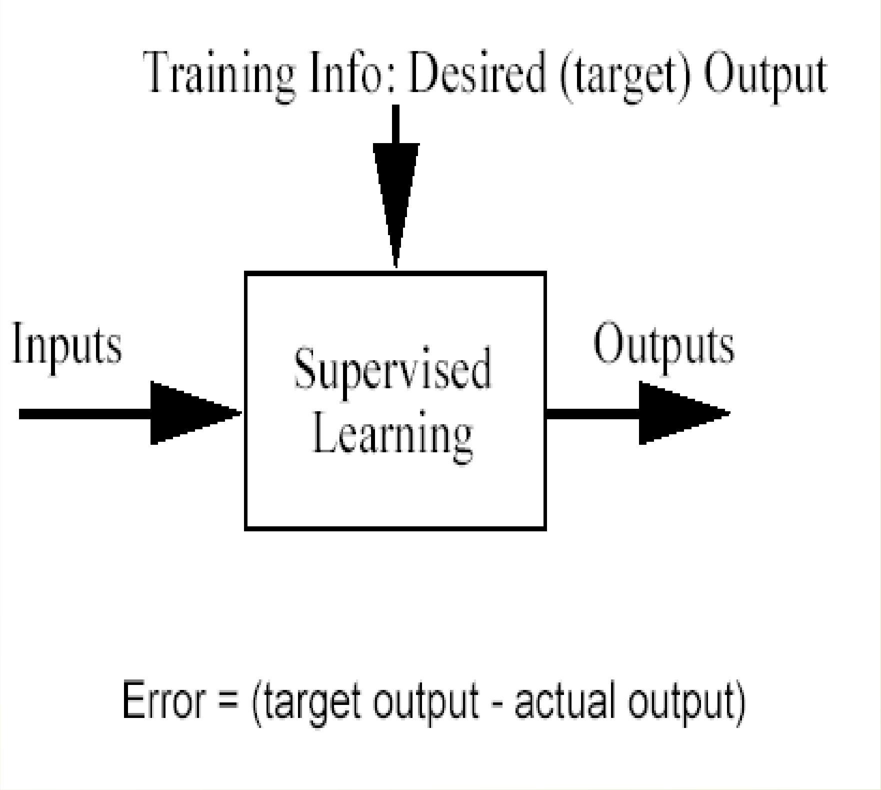


Advantage of Regression Trees

- Simple to understand and interpret
- Requires little data preparation
- Able to handle both numerical and categorical data
- Possible to validate a model using statistical tests
- Performs well with large datasets
- **Limitations**
  - For data including categorical variables with different number of levels, information gain in decision trees are biased in favour of those attributes with more levels
  - Calculations can get very complex particularly if many values are uncertain and/or if many outcomes are linked

Naïve Bayes

- Simple probabilistic classifier based on applying Bayes' theorem
- Assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature



Results

- Tested in diabetics data with 768 rows
- Split the data set randomly into train and datasets with a ratio of 67% train and 33% test
- Split 768 rows into train=514 and test=254 rows
- Accuracy: 76.3779527559%

Advantage of Naïve Bayes

- Can be trained very efficiently in a supervised learning setting
- Have worked quite well in many complex real-world situations
- Requires a small amount of training data to estimate the parameters
- **Limitations**
  - Assumption: class conditional independence , therefore loss of accuracy
  - Practically, dependencies exist among variables
  - Dependencies among these cannot be modeled by Naïve Bayesian Classifier

Conclusion and Future Work

- Random Forest is fast to build, even faster to predict
- Regression trees are easy to interpret and explain
- Naive Bayes is Not So Naïve since it handles real and discrete data

Future Work

- Choose the best model function for our dataset
- Modify the model function to more closely meet the business needs
- Make the solution scalable to distributed platform
- Make the solution generic enough to transform other non-structured information

Thank You ☺