---

title: "Recommendations for Insurance Policy Purchases"

author: "Team-6"

date: "December 16, 2016"

output: pdf\_document

---

Prepared by Davin Kaing, Nikhita Kandikonda, and Praneeth Tamvada

#Abstract

The objective of our project is to develop business recommendations for insurance companies. More specifically, we are interested in exploring factors that are most associated with insurance policies purchases. With this information, we can then pinpoint recommendations and further studies that can increase revenues. To investigate this, we collected Allstate data from Kaggle. The dataset contains attributes about the automotive vehicle, the consumers, and the location of purchase. We began the analysis by preprocessing and cleaning the data. Afterwards, we implemented three machine learning algorithms: Logistic Regression, Naive Bayes, and Decision Tree. Our motivation for selecting various models is to capture the relationships between the attributes (for example, if Decision Tree provides the highest accuracy, we can conclude that the attributes are conditionally related). Once these models were implemented, we evaluated the models with the ROC and AUC and based on these criteria, Logistic Regression is the best model for this dataset. Given that Logistic Regression yields the most accuracy, we examined the important variables in this model, which are State, Car Age, and Shopping Point. Our results show that Tennessee, Alabama, and Utah are the best states, whereas, the worst states are Maine, Connecticut, and North Dakota. As of car age, we saw that if the car is older than 70 months, the purchase rate is greater than 50%. In regards to Shopping Point, 8-11 Shopping Point ID yield the highest subscription rates. From these results, we suggest the following recommendations: conduct comparative studies of the top 3 best and worst states and examine other specific factors that may lead the top states to outperform the worst states, conduct a survey study to pinpoint regions that have cars older than 70 months and target those regions, and lastly, analyze incentive systems for Shopping Point's 8-11 and incorporate those incentive system into the other Shopping Points. Our analysis is reliable because we implemented different algorithms to capture the different relationships of the attributes. In addition, we only selected the most accurate model which support the credibility of our analysis. For future study, our team would like to examine the recommendations we provided and then generate more specific recommendations that have tangible impact in the business revenue.

# Background Research and Question Development

When shopping for car insurance policies, certain factors may influence consumers' subscriptions. For example, the car age may be one of those factors. If the car is older, the insurance price is cheaper than new cars, hence, individuals may be more likely to purchase the insurance policies. Our team is interested in understanding similar factors, such as car age, that can influence consumer subscriptions. We narrowed down our project objective to work on business recommendations for insurance companies based on different attributes of a customer (age, income, assets etc.) as well as product information. We came up with the plan to analyze different attributes of an insurance customer and then build different models like random forest, logistic regression and Na?ve Bayes and compare all these models for accuracy. Using the best model, we looked at the most important variables that influence the subscription rate. From these variables, we can then generate business recommendations to potentially increase the revenue.

As of the dataset used for this project, we searched different sources such as healthcare.gov, census.gov and kaggle.com. We found a dataset in Kaggle from 'Allstate Purchase Prediction Challenge', the dataset had many policies as variables but, we chose to take just one claim and build our models from other attributes of a customer, if he/she is going to purchase the policy We have modified our data according to our question and removed some irrelevant variables to optimize our dataset.

# S.M.A.R.T Question

Our project aims to investigate attributes that can influence the subscription or cost of insurance and generate recommendations based on these findings.

# Tools Used

This project is completed in R. We apply logistic regression, Decision tree and Naive Bayes to build our models. To visualize our results, we use ggplot.

#Code

```{r packages, include=FALSE}

library(rpart)

library(caret)

library(e1071)

library(randomForest)

library(rattle)

library(dplyr)

library(pROC)

library(ROCR)

library(reshape2)

library(gridExtra)

library(grid)

```

## Data Import and Preprocessing

```{r train data}

# read file

train <- read.csv2("~/Downloads/train.csv",header=T,sep =",")

# dimension of data

dim(train)

```

We studied the variables to remove unrelated variables and

removed the colums with more than 10% NA values

```{r cleaning}

# removed unwanted columns

train <- train[!names(train) %in%

c("A", "C", "D", "E", "F", "G", "cost")]

# converted time to two separate rows: hours and minutes

train$time <- as.character(train$time)

train$hr <- as.numeric(substr(train$time, 1,2))

train$min <- as.numeric(substr(train$time, 4,5))

# remove irrelevant variables

train$time <- NULL

train$risk\_factor <- NULL

train$C\_previous <- NULL

# dimension of data after cleaning

dim(train)

```

After initial analysis and transformation, we reduced our variables from 25 to 16

```{r test data}

# testing dataset

test <- read.csv2("~/Downloads/test\_v2.csv",header=T,sep =",")

# remove irrelevant variables

test <- test[!names(test) %in%

c("A", "C", "D", "E", "F", "G", "cost")]

# create two variables for time as hour and minute

test$time <- as.character(test$time)

test$hr <- as.numeric(substr(test$time, 1,2))

test$min <- as.numeric(substr(test$time, 4,5))

# remove irrelevant variables

test$time <- NULL

test$risk\_factor <- NULL

test$C\_previous <- NULL

```

## Building models

### Decision Tree

```{r }

# Building decison tree model

dt <- rpart(B~., method = "class", data = train)

# Predicting values using decision tree model

dt\_pred <- predict(dt, test, type = "prob")

dt\_prediction <- prediction(dt\_pred[,2], test$B)

# Store performance measure data

dt\_perf <- performance(dt\_prediction, measure = "tpr", x.measure = "fpr")

dt\_roc <- data.frame(dt\_perf@x.values, dt\_perf@y.values)

colnames(dt\_roc) <- c("fpr", "tpr")

dt\_roc$type <- "Decision Tree"

```

### Naive Bayes

```{r }

# Building naive bayes model

nb <- naiveBayes(B~., data = train)

# Predicting outcome using the model

nb\_pred <- predict(nb, test, type = "raw")

nb\_prediction <- prediction(nb\_pred[,2], test$B)

# Store performance measure data

nb\_perf <- performance(nb\_prediction, measure = "tpr", x.measure = "fpr")

nb\_roc <- data.frame(nb\_perf@x.values, nb\_perf@y.values)

colnames(nb\_roc) <- c("fpr", "tpr")

nb\_roc$type <- "Naive Bayes"

```

### Logistic Regression

```{r }

# Building logistic regression model

logit <- glm(B~., binomial(link = "logit"), data = train)

# Predicting using the model built

logit\_pred <- predict(logit, test, type = "response")

logit\_prediction <- prediction(logit\_pred, test$B)

# Store performance measure data

logit\_perf <- performance(logit\_prediction, measure = "tpr", x.measure = "fpr")

logit\_roc <- data.frame(logit\_perf@x.values, logit\_perf@y.values)

colnames(logit\_roc) <- c("fpr", "tpr")

logit\_roc$type <- "Logistic Regression"

```

## Plots for ROC (Receiver Operating Characteristic)

```{r roc}

# Combining the performance metrics of 3 models

roc <- rbind(dt\_roc, nb\_roc, logit\_roc)

# Plot of ROC of three models

ggplot(roc, aes(fpr, tpr, color = type)) +geom\_line() +

theme(legend.title = element\_blank(),

text = element\_text(size = 15))+

xlab("False Positive Rate") +

ylab("True Positive Rate") +

ggtitle("Figure 1: ROC Plots")

```

## Confusion Matrix

```{r Confusionmatrix}

# Confusion matrix of three models

dt\_acc <- confusionMatrix(ifelse(dt\_pred[,2]>0.5, 1,0),

test$B)$overall[1]

nb\_acc <- confusionMatrix(ifelse(nb\_pred[,2]>0.5, 1,0),

test$B)$overall[1]

logit\_acc <- confusionMatrix(ifelse(logit\_pred>0.5,1,0),

test$B)$overall[1]

# Building a data frame with the accuracy measures of all three models

acc <- data.frame(Accuracy = c(dt\_acc, nb\_acc, logit\_acc),

Model = c("Decision Tree", "Naive Bayes",

"Logistic Regression"))

acc <- acc[,c(2,1)]

plot.new()

grid.table(acc, rows = NULL)

```

## Variable Importance and Visualization of Most Important Variables

```{r VariableImportance}

# Variable importance of Logistic Regression

varimp\_logit <- varImp(logit)

varimp\_logit <- data.frame(Variable = rownames(varimp\_logit),

Importance = varimp\_logit$Overall)

varimp\_logit <- varimp\_logit[order(varimp\_logit$Importance, decreasing = T),]

varimp\_logit <- varimp\_logit[1:24,]

ggplot(varimp\_logit, aes(reorder(Variable, Importance), Importance)) +

geom\_bar(stat = "identity", fill = "darkgreen") +

coord\_flip() +xlab("Variable") +

ggtitle("Figure 2: Variable Importance of Logistic Regression")

# Variable importance of decision tree

varimp\_dt <- varImp(dt)

varimp\_dt <- data.frame(Variable = rownames(varimp\_dt),

Importance = varimp\_dt$Overall)

varimp\_dt <- varimp\_dt[order(varimp\_dt$Importance, decreasing = T),]

varimp\_dt <- varimp\_dt[1:5,]

ggplot(varimp\_dt, aes(reorder(Variable, Importance), Importance)) +

geom\_bar(stat = "identity", fill = "darkred") +

coord\_flip() +xlab("Variable") +

theme(text = element\_text(size = 15)) +

ggtitle("Figure 3: Variable Importance of Decision Tree Model")

# Visualization of most important variable - State

state\_B <- train %>% group\_by(state) %>% summarise(mean = mean(B))

ggplot(state\_B, aes(reorder(state, -mean), mean)) +

geom\_bar(stat = "identity", fill = "goldenrod3") + xlab("State") +

ylab("Percent of Purchases") +

ggtitle("Figure 4: Purchase Rate by State")

# Visualization of most important variable - Car Age

car\_age\_B <- train %>% group\_by(car\_age) %>% summarize(mean = mean(B))

ggplot(car\_age\_B, aes(car\_age, mean)) + geom\_point() +

geom\_smooth() + xlab("Car Age") +

ylab("Percent of Purchases") +

ggtitle("Figure 5: Purchase Rate by Car Age")

# Visualization of most important variable - Shopping Pt

shopping\_pt\_B <- train %>% group\_by(shopping\_pt) %>% summarize(mean = mean(B))

ggplot(shopping\_pt\_B, aes(shopping\_pt, mean)) + geom\_bar(stat ="identity", fill = "navyblue") + xlab("Shopping Point") +

ylab("Percent of Purchases")+

ggtitle("Figure 6: Purchase Rate by Shopping Point")

```

# Summary

The objective of our project is to generate business recommendations that can potentially increase the subscription rate of car insurance policies. To do this, we downloaded Allstate data from Kaggle and applied four machine learning algorithms: Decision Tree, NaÃ¯ve Bayes, and Logistic Regression. These models output accuracies of 0.5889, 0.5910, and 0.5925 respectively. We then selected the most accurate model, Logistic Regression, and explored the variable importance of this model. The most important variables are: state, car age, and shopping points. Visualizing these important variables provide us key insights into possible business recommendations that can potentially increase revenue. Based on our findings, our recommendations are as follow: conduct comparative studies of the top 3 states and the worst 3 states, conduct a survey to map out regions with cars older than 70 months, and analyze incentive systems for shopping points ID 8 through 11. These recommendations can be used to conduct future studies which can further narrow down specific actions items to potentially increase revenue.

# References

"Allstate Purchase Prediction Challenge." Allstate Purchase Prediction Challenge. Kaggle, n.d. Web. 16 Dec. 2016.

https://en.wikipedia.org/wiki/Key\_person\_insurance

https://www.kaggle.com/c/allstate-purchase-prediction-challenge/data