Auto Insurance Claim Prediction

Shilpa Kolekar

## Overall, write a coherent narrative that tells a story with the data as you complete this section.

I have picked below link for my research. Dataset is on Kaggle site and 2 years old. <https://www.kaggle.com/xiaomengsun/car-insurance-claim-data>. Dataset contains .CSV file which has 10,303 total records and 27 columns. Majority of columns hold String, Integer and Boolean values.I have chosen this dataset because it has relevant variables that I can use in my research.

# Read the file `car\_insurance\_claim.CSV`   
ins\_df <- read.csv("car\_insurance\_claim.CSV")  
str(ins\_df)

## 'data.frame': 10302 obs. of 27 variables:  
## $ ï..ID : int 63581743 132761049 921317019 727598473 450221861 743146596 871024631 792300541 7945239 3577610 ...  
## $ KIDSDRIV : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ BIRTH : chr "16-Mar-39" "21-Jan-56" "18-Nov-51" "5-Mar-64" ...  
## $ AGE : int 60 43 48 35 51 50 34 54 40 44 ...  
## $ HOMEKIDS : int 0 0 0 1 0 0 1 0 1 2 ...  
## $ YOJ : int 11 11 11 10 14 NA 12 NA 11 12 ...  
## $ INCOME : chr "$67,349 " "$91,449 " "$52,881 " "$16,039 " ...  
## $ PARENT1 : chr "No" "No" "No" "No" ...  
## $ HOME\_VAL : chr "$0 " "$257,252 " "$0 " "$124,191 " ...  
## $ MSTATUS : chr "z\_No" "z\_No" "z\_No" "Yes" ...  
## $ GENDER : chr "M" "M" "M" "z\_F" ...  
## $ EDUCATION : chr "PhD" "z\_High School" "Bachelors" "z\_High School" ...  
## $ OCCUPATION: chr "Professional" "z\_Blue Collar" "Manager" "Clerical" ...  
## $ TRAVTIME : int 14 22 26 5 32 36 46 33 21 30 ...  
## $ CAR\_USE : chr "Private" "Commercial" "Private" "Private" ...  
## $ BLUEBOOK : chr "$14,230 " "$14,940 " "$21,970 " "$4,010 " ...  
## $ TIF : int 11 1 1 4 7 1 1 1 6 10 ...  
## $ CAR\_TYPE : chr "Minivan" "Minivan" "Van" "z\_SUV" ...  
## $ RED\_CAR : chr "yes" "yes" "yes" "no" ...  
## $ OLDCLAIM : chr "$4,461 " "$0 " "$0 " "$38,690 " ...  
## $ CLM\_FREQ : int 2 0 0 2 0 2 0 0 1 0 ...  
## $ REVOKED : chr "No" "No" "No" "No" ...  
## $ MVR\_PTS : int 3 0 2 3 0 3 0 0 2 0 ...  
## $ CLM\_AMT : chr "$0 " "$0 " "$0 " "$0 " ...  
## $ CAR\_AGE : int 18 1 10 10 6 17 7 1 1 10 ...  
## $ CLAIM\_FLAG: int 0 0 0 0 0 0 1 0 1 0 ...  
## $ URBANICITY: chr "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/ Urban" ...

I have following type of data which will be used in this project.

1. KIDSDRIV – Number of kid drivers per household. It has integer value.
2. AGE – Age of the driver between age of 16 years - 99 years. It has integer value.
3. MSTATUS - Material status of the driver. It has integer value. ‘Yes’ has value of 1 and ‘No’ has value of 0.
4. GENDER - Gender of the driver. It has integer value. ‘Male’ has value of 1 and ‘Female’ has value of 0.
5. TRAVTIME – Driver’s travel time per day. It is given in minutes and has integer value.
6. CLM\_FREQ – Claim frequency of the driver. Variable will give you how many claims customer had in the past few years. Value is integer.
7. REVOKED – If the driver has revoked license then the value is 1, if not the value is 0.
8. CLAIM\_FLAG - Claim Flag on the drivers policy. If there is flag, value is 1 and if no flag it is 0.

CLAIM\_FLAG variable is the outcome variable while all other variables are independent variables.

The car\_insurance\_claim file has 10302 rows. After reviewing the file, I found that there are seven rows where the age is blank. These rows were removed from the dataset. GENDER variable has values “z\_F” and “M” which were converted to integer values, 0 for z\_F and 1 for M. MSTATUS variable has values “z\_No” and “Yes” which were converted to integer values, 0 for z\_No and 1 for Yes. REVOKED variable has values “No” and “Yes” which were converted to integer values, 0 for No and 1 for Yes. Any variables not being used were removed. After doing data cleaning process, the number of rows of the dataset were reduced from 10302 to 10295.

new\_ins\_df <- ins\_df[!is.na(ins\_df$AGE),]  
  
new\_ins\_df$REVOKED[new\_ins\_df$REVOKED == 'Yes'] <- 1  
new\_ins\_df$REVOKED[new\_ins\_df$REVOKED == 'No'] <- 0  
new\_ins\_df$REVOKED <- as.integer(new\_ins\_df$REVOKED)  
  
new\_ins\_df$MSTATUS[new\_ins\_df$MSTATUS == 'Yes'] <- 1  
new\_ins\_df$MSTATUS[new\_ins\_df$MSTATUS == 'z\_No'] <- 0  
new\_ins\_df$MSTATUS <- as.integer(new\_ins\_df$MSTATUS)  
  
new\_ins\_df$GENDER[new\_ins\_df$GENDER == 'M'] <- 1  
new\_ins\_df$GENDER[new\_ins\_df$GENDER == 'z\_F'] <- 0  
new\_ins\_df$GENDER <- as.integer(new\_ins\_df$GENDER)  
  
new\_ins\_df$ï..ID <- NULL  
new\_ins\_df$BIRTH <- NULL  
new\_ins\_df$HOMEKIDS <- NULL  
new\_ins\_df$YOJ <- NULL  
new\_ins\_df$INCOME <- NULL  
new\_ins\_df$PARENT1 <- NULL  
new\_ins\_df$HOME\_VAL <- NULL  
new\_ins\_df$EDUCATION <- NULL  
new\_ins\_df$OCCUPATION <- NULL  
new\_ins\_df$CAR\_USE <- NULL  
new\_ins\_df$BLUEBOOK <- NULL  
new\_ins\_df$TIF <- NULL  
new\_ins\_df$RED\_CAR <- NULL  
new\_ins\_df$MVR\_PTS <- NULL  
new\_ins\_df$CAR\_AGE <- NULL  
new\_ins\_df$CAR\_TYPE <- NULL  
new\_ins\_df$URBANICITY <- NULL  
new\_ins\_df$OLDCLAIM <- NULL  
new\_ins\_df$CLM\_AMT <- NULL

I have first few rows of data as below which are from clean and final dataset. I used head() function to display top few records.

head(new\_ins\_df)

## KIDSDRIV AGE MSTATUS GENDER TRAVTIME CLM\_FREQ REVOKED CLAIM\_FLAG  
## 1 0 60 0 1 14 2 0 0  
## 2 0 43 0 1 22 0 0 0  
## 3 0 48 0 1 26 0 0 0  
## 4 0 35 1 0 5 2 0 0  
## 5 0 51 1 1 32 0 0 0  
## 6 0 50 1 0 36 2 1 0

## Summarize the problem statement you addressed.

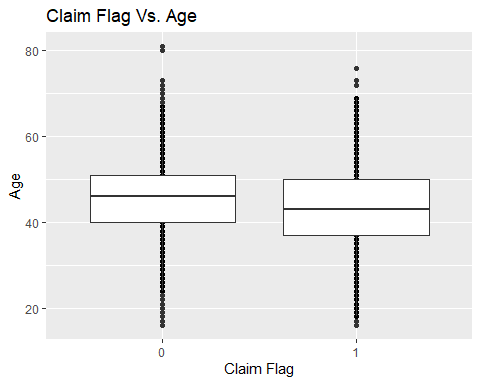
Problem statement for this project is, how insurance company predicts the risk of each individual and offer them appropriate auto rate, or in other words, how actuaries use to predict customers’ risk factors and come up with the auto rate. Following questions will focus on my problem statement.

1. Does travel time put you in high risk driver category?
2. Does number of kids driver in the household play any role?
3. Does gender play any role when actuaries decide rate?
4. Does age of the driver play any role?
5. Does past high claims frequency put individual in high risk driver category?
6. Does traffic violations/tickets impact auto rate?
7. Does history of suspended license impact auto rate?

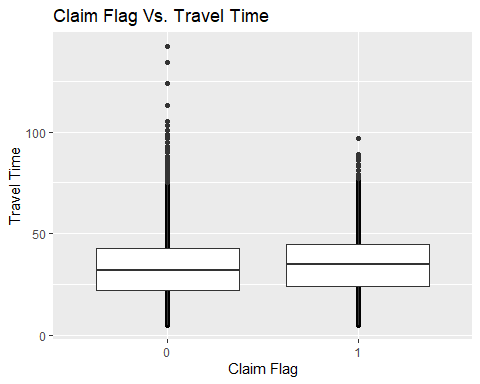
## Summarize how you addressed this problem statement (the data used and the methodology employed).

I started my analysis by plotting my variables. Based on the type of the variable, I determined which graph to use.between I used box plot for age and travel time variables, and bar grapg for remaining variables. I compared each of my variable with “claim Flag” variable to check the relationship of them.

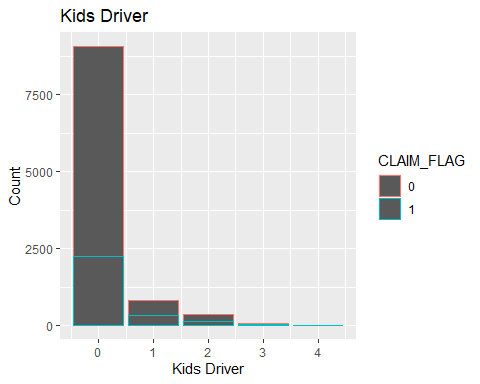
library('ggplot2')  
ins\_df\_factor <- new\_ins\_df  
ins\_df\_factor$CLAIM\_FLAG <- as.factor(ins\_df\_factor$CLAIM\_FLAG)  
ins\_df\_factor$MSTATUS <- as.factor(ins\_df\_factor$MSTATUS)  
ins\_df\_factor$GENDER <- as.factor(ins\_df\_factor$GENDER)  
ins\_df\_factor$REVOKED <- as.factor(ins\_df\_factor$REVOKED)  
  
ggplot(ins\_df\_factor, aes(x=CLAIM\_FLAG, y=AGE)) + geom\_point() + geom\_boxplot() + ggtitle("Claim Flag Vs. Age") + xlab("Claim Flag") + ylab("Age")



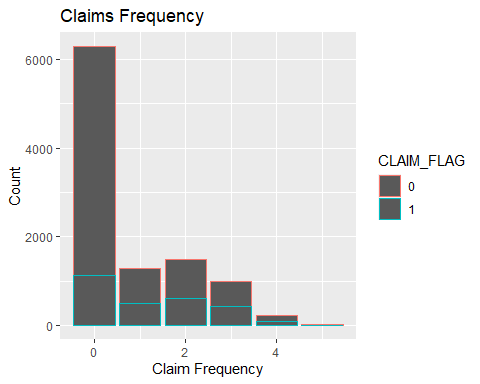
ggplot(ins\_df\_factor, aes(x=CLAIM\_FLAG, y=TRAVTIME)) + geom\_point() + geom\_boxplot() + ggtitle("Claim Flag Vs. Travel Time") + xlab("Claim Flag") + ylab("Travel Time")



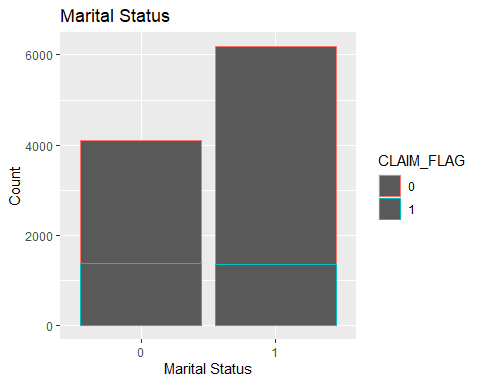
ggplot(ins\_df\_factor, aes(KIDSDRIV, color=CLAIM\_FLAG)) + geom\_bar() + ggtitle("Kids Driver") + xlab("Kids Driver") + ylab("Count")



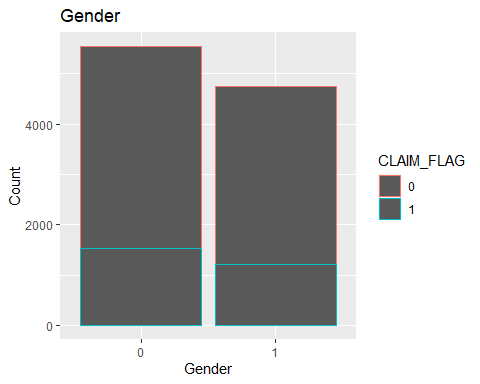
ggplot(ins\_df\_factor, aes(CLM\_FREQ, color=CLAIM\_FLAG)) + geom\_bar() + ggtitle("Claims Frequency") + xlab("Claim Frequency") + ylab("Count")



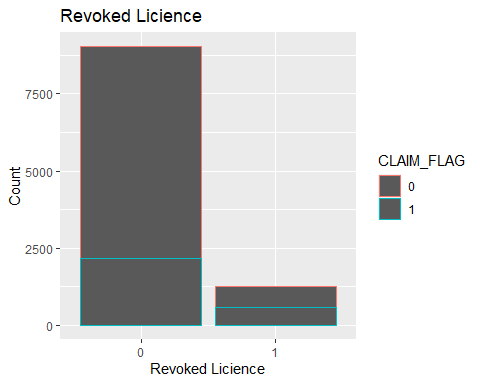
ggplot(ins\_df\_factor, aes(MSTATUS, color=CLAIM\_FLAG)) + geom\_bar() + ggtitle("Marital Status") + xlab("Marital Status") + ylab("Count")



ggplot(ins\_df\_factor, aes(GENDER, color=CLAIM\_FLAG)) + geom\_bar() + ggtitle("Gender") + xlab("Gender") + ylab("Count")



ggplot(ins\_df\_factor, aes(REVOKED, color=CLAIM\_FLAG)) + geom\_bar() + ggtitle("Revoked Licience") + xlab("Revoked Licience") + ylab("Count")



Based on visual of my graphs, I came up with following assumptions about my variables.

1. Claim Flag Vs. Age - This box plot clearly indicates that when age goes up claims are going down, which means younger person has higher chances of getting claims.
2. Claim Flag Vs. Travel Time - This box plot is almost the same for shorter travel time and longer travel time. It means travel time does not have much impact on claim.
3. Kid Driver - Kids driver is 1 to 4 in range, and I decided to use histogram. BY looking at the graph, I can tell that those who have 0 kids have low claims but when number of kids grows in household, claims are growing respectively. 4.Claims frequency - claims frequency is in 1 to 4 range and histogram was the better option to visualize this variable. Graph suggests that those who have more claim frequency in the past have higher chances of having future claims. 5.Marital Status - I plotted this variable using Histogram and it indicates that married drivers are safer than single drivers. 6.Gender - Through histogram, I was able to determine that this variable does not show any significant difference wheather you are male or female. 7.Revoked license - Through histogram, I was able to determine that driver with revoked license has more claims that driver with non-revoked license.

After visualizing each variables using graph, I found out Correlation Coefficient for each variable with CLAIM\_FLAG variable by using cor() function.

cor(new\_ins\_df$CLAIM\_FLAG, new\_ins\_df$KIDSDRIV)

## [1] 0.1088177

cor(new\_ins\_df$CLAIM\_FLAG, new\_ins\_df$AGE)

## [1] -0.1069695

cor(new\_ins\_df$CLAIM\_FLAG, new\_ins\_df$MSTATUS)

## [1] -0.1289854

cor(new\_ins\_df$CLAIM\_FLAG, new\_ins\_df$GENDER)

## [1] -0.02183889

cor(new\_ins\_df$CLAIM\_FLAG, new\_ins\_df$TRAVTIME)

## [1] 0.05400096

cor(new\_ins\_df$CLAIM\_FLAG, new\_ins\_df$CLM\_FREQ)

## [1] 0.2231861

cor(new\_ins\_df$CLAIM\_FLAG, new\_ins\_df$REVOKED)

## [1] 0.1560245

I compared correlation coefficient and my above analysis to see if they truly match. Below is my comparison.

KIDDRIVER, low positive correlation, when kids per household increases, claims increase. AGE, low negative correlation, when age goes up, claims are going down. MSTATUS, low negative correlation, when individual is single, claims are high and married has low claims. GENDER, no correlation, gender has no impact on claims. TRAVTIME, no correlation, travel time has no impact on claims. CLAM\_FREQ, low positive correlation, more past claims has higher chances of having future claims. REVOKED, low positive correlation, those with revoked license have more chances of having claims.

The outcomes of graphs and correlation coefficient match.

After getting satisfactory results, I created a logistic regression model and checked p-value of each variable.

newModel <- glm(CLAIM\_FLAG ~ KIDSDRIV + AGE + MSTATUS + GENDER + TRAVTIME + CLM\_FREQ + REVOKED, data = new\_ins\_df, family = binomial)  
  
summary(newModel)

##   
## Call:  
## glm(formula = CLAIM\_FLAG ~ KIDSDRIV + AGE + MSTATUS + GENDER +   
## TRAVTIME + CLM\_FREQ + REVOKED, family = binomial, data = new\_ins\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.9939 -0.7652 -0.5895 0.9679 2.2594   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.533091 0.136558 -3.904 9.47e-05 \*\*\*  
## KIDSDRIV 0.409328 0.042479 9.636 < 2e-16 \*\*\*  
## AGE -0.022308 0.002761 -8.080 6.48e-16 \*\*\*  
## MSTATUS -0.531935 0.047527 -11.192 < 2e-16 \*\*\*  
## GENDER -0.069921 0.047473 -1.473 0.141   
## TRAVTIME 0.008304 0.001475 5.631 1.79e-08 \*\*\*  
## CLM\_FREQ 0.389193 0.019133 20.341 < 2e-16 \*\*\*  
## REVOKED 0.881206 0.064668 13.627 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11930 on 10294 degrees of freedom  
## Residual deviance: 10909 on 10287 degrees of freedom  
## AIC: 10925  
##   
## Number of Fisher Scoring iterations: 4

The summary of model indicates that all variables are statistically significant based on the p values, except for the variable gender. I then created confusion matrix and calculated accuracy of the model after splitting the dataset in 80-20 proportion for training and testing the data.

data\_split <- sample(1:nrow(new\_ins\_df), 0.8 \* nrow(new\_ins\_df))  
train <- new\_ins\_df[data\_split,]  
test <- new\_ins\_df[-data\_split,]  
  
res <- predict(newModel, test, type="response")  
res <- predict(newModel, train, type="response")  
  
confmatrix <- table(Actual\_Value = train$CLAIM\_FLAG, Predicted\_Value = res > 0.5)  
confmatrix

## Predicted\_Value  
## Actual\_Value FALSE TRUE  
## 0 5755 301  
## 1 1798 382

accuracy <- (confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)  
accuracy

## [1] 0.7451433

The accuracy of the logistic regression model was 74.70%.

In the last step of the project, I created K-nearest neighbors algorithm. I split the data in 80-20 proportion and prepared train and test datasets. Data was run through train and test data set and confusion matrix was created and accuracy was calculated.

##extract 8th column of train dataset because it will be used as 'cl' argument in knn function.  
target\_category <- new\_ins\_df[data\_split,8]  
  
##extract 8th column if test dataset to measure the accuracy  
test\_category <- new\_ins\_df[-data\_split,8]  
  
##load the package class  
library("class")  
  
##run knn function, k=sqrt(10295)  
test\_pred <- knn(train,test,cl=target\_category,k=101)  
  
##create confusion matrix  
table <- table(test\_category, test\_pred)  
table

## test\_pred  
## test\_category 0 1  
## 0 1498 1  
## 1 549 11

# Accuracy  
(table[[1,1]]+ table[[2,2]]) / sum(table)

## [1] 0.73288

Accuracy rate was calculated which was 74.21%.

## Summarize the interesting insights that your analysis provided.

After analyzing graphs, coefficient correlation and logistic model, I concluded that auto rates vary for person to person and there are many factors that are involved in rate making. Factors that I chose to predict auto rate model gave me good idea on how insurance company predicts the rate. According to my model, individuals who are teenagers or in their 20’s have higher rate compared to individuals in their 30-50’s and older generation. Also, married couples are more safe drivers and therefore rate is lower for this category. Surprisingly, travel time and gender do not impact on auto rate much, but if you have revoked license and more claim frequency in past, that might cost you more.

## Summarize the implications to the consumer (target audience) of your analysis.

My target audience is insurance companies and general public. Based on my analysis, insurance companies can use these factors to determine which individual should be charged more. Younger people, families with high number of kid drivers, single people, people with revoked license, and people with past claim history should be charged more compared to the opposite.

I can suggest to general public, that having a revoked license can cost more money on your rate. If you are young driver, your rate is going to be high. Keep your driving record clean and find out ways on how to save on auto rate. If you have more past claim, learn from past and see how you can avoid some driving mistakes. Past claims only stay for certain years on your records. If you have kids driver in the house, you are expected to pay more rate because risk is higher with new drivers.

## Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

I was expecting to get at least one strong variable that has significant impact onrate making. I was open for positive or negative impact of that variable. Unfortunately all my variables are making some impact in positive and negative way, but I didn’t have any variable in my model which stood out. If anybody is building on top of this model, I would suggest them to research and add more variables that could have significant impact on possibility of claims.