***Project Report***

***Topic: Safe Driver Prediction***

* **Purpose :** Build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year.
* **Methodology Involved :** Many aspects were kept in mind for the sole purpose of predicting with better accuracy.

1. **Knowing the dataset:**

I studied the whole dataset and the variable meanings whatever was given as an introduction over there, but many variables were not well explained, so was not able to know what they really signify in real aspects.

#Commands involved

str(data) , head(data) as well as tail(data)

All the basic information was with me that how many variables are of numeric types and of the integer type.

**2. Missing Value Imputations:**

As my dataset didn’t had any missing value so this much of pre processing was saved and I went ahead to process the further steps in order to gain some predictions

# Commands involved:

**MissingData = data.frame(varaibles = colnames(test), MissingInfo = apply(test,2,function(x)sum(is.na(x))))**

A data set was created which was named missing data for both train as well as test data set.

**3. Outlier Analysis:**

The dataset I had, it contained many outlier values for which I had some methods either to replace them by their mean or with some other processes.

Even removing the values which had outliers would have pushed my test dataset into error prone area as I would have got some rows deleted which would have lead to No submission for the accuracy check in kaggle.

The best method I found was Outlier Capping where I capped all the upper outliers with the 95% quantile and lower outliers with 5% quantiles.

**# function for outlier capping :**

fun <- function(x){

quantiles <- quantile( x, c(.05, .95 ) )

x[ x < quantiles[1] ] <- quantiles[1]

x[ x > quantiles[2] ] <- quantiles[2]

x

}

**4. Exhaustive EDA:**

Plotted n number of graphs which had both univariate as well as bivariate analysis. For Univariates I plotted histograms of each independent variable to know the nature of those variables like checking which of them had skewness present with them. For Bivariates I compared all the numeric variables with the target variable in train dataset

**# libraries used as well as function**

hist(train$V) # where V is the name of the variable

Library(ggplot2) # used this library for bivariate analysis and for checking of outliers as well

**5. Feature Engineering and Variable Importance:**

Using functions varImp(xyz) I tried finding the important variables but this wasn’t working so was not able to detect the Importance of each variable. Also as data variables were bit hard to understand so was not able to perform any sort of feature engineering as well but as variables were not much correlated with each other I went ahead with all the variables present in my dataset.

**6. Model Building:**

As a data science enthusiast I keep reading all the new things that I can follow on my data for achieving better results in data science competitions be it Analytics Vidhya or Kaggle. So while reading I came across a library called as H2O which I personally feel increased the speed of the outputs I had. For that I installed java 8 as this library doesn’t works on java 9 currently so I faced many problems while running java 8 in present of java 9. Finally I was through with that problem as well.

**Random Forest:**

In h2o library I ran random forest for my train dataset then predicted my values for the test dataset and I was disappointed with the score and rank I had on kaggle.

Got around 4180 rank with a score 0.24180

**GBM:**

After running random forest I ran gradient Boosting algorithm and as expected my score got improved and obviously the rankings as well.

This time I got a score of 0.27190 and rank would have been around 3200 if I had participated in the competition.

**Conclusion:**

If I would have understood the meaning of variables more accurately present in my dataset I would have feature engineered some new variables which would have increased the accuracy by some margins.

Currently I am in the learning stage so I do no know how to use cross validation techniques more properly otherwise results would have been much better and accurate.

Also I would like to mention I tried studying about taking the mean of the models I ran and it would have fetched me better results, would really do that in coming days after learning a bit more about them as I was not sure how to proceed with that.