Universidad de Navarra

IMPLEMENTING THE NAIVE BAYES ALGORITHM FOR GENDER PREDICTION

MONTERO MUÑOZ, GONZALO
RICCI ZARAGOZA, DANIEL A.
VILLANUEVA JOVE, JAVIER E.

MACHINE LEARNING

Introduction

In this project, we will analyze a data set containing statistics of Great Britain's road accidents between 1979 -2015, using this information we will apply a prediction algorithm that will let us determine the driver's gender based in their accident's description. We believe that there could be a direct correlation between the number and severity of car accidents with the driver's gender.

The Data Set

The data set “Road accidents data Great Britain 1979-2015” was taken from *kaggle*. Primarily, the data captures fatal road accidents in the UK between 1979 and 2015 and have 70 features/columns and about 250K rows. Nevertheless, for our project we will only be using the variables *“Drivers Age”, “Vehicles Age”, “Number Of Casualties” and “Number Of Vehicles Involved”* to see if a gender prediction with a high accuracy can be made.

In the chart below, it’s possible to appreciate the range of the 4 quantitative variables that we select from the data set to develop our project and the number of male and female drivers that were involve in car accidents during the period we are studding.

https://www.kaggle.com/akshay4/road-accidents-incidence

Range	DriversAge	VehiclesAge	NumberOfCasualties	NumberOfVehicles
min	7	0	1	1
avg	40,35963079	7,0073131	1,646134745	2,119902609
max	97	105	38	37

DriversGender	Number
Male	81052
Female	129381

Objective

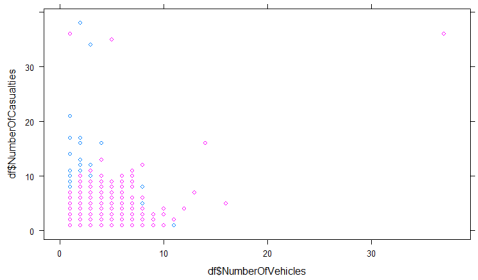
The goal of this project is to determine if there is correlation between the 4 quantitative variables (*“Drivers Age”, “Vehicles Age”, “Number Of Casualties” and “Number Of Vehicles Involved”*) and our qualitative variable (“Drivers Gender). We believe that there are differences in the accidents that will let us know if the driver is male or female. For example, the 61,5% of the subjects are females meaning they have a higher probability to be involve in a car accident, so we will apply the ***Naïve Bayes*** algorithm to find if there is indeed this correlation.

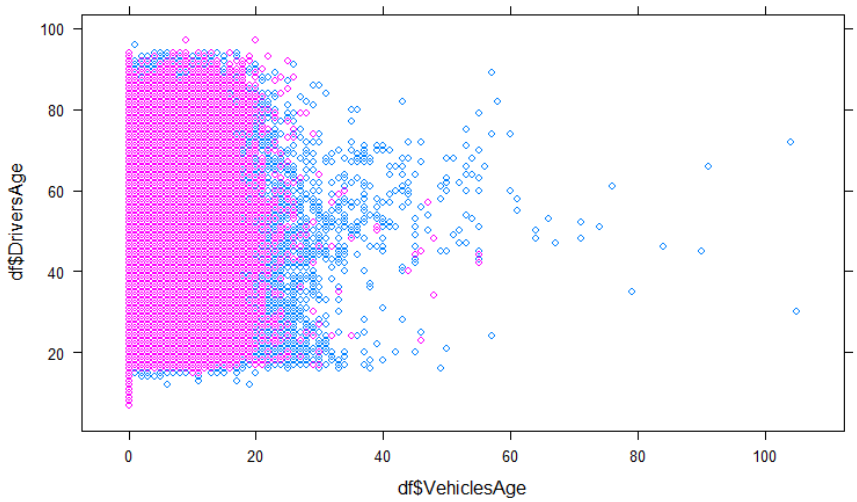
Analysis

Just by analyzing the data set, we found correlations among the four numerical variables that started to validate our assumptions. As we displayed the gender of the drivers, we came to acknowledge that the majority of accidents were indeed caused by women behind the wheel. Also, by comparing the ages of the drivers and vehicles, and the number of casualties and vehicles involved, other clear distinctions among both genders were discovered.

```
> table(df$DriversGender)
> xyplot(df$DriversAge ~ df$VehiclesAge, group=DriversGender, data=df)
> xyplot(df$NumberOfCasualties ~ df$NumberOfVehicles, group=DriversGender, data=df)
```

```
Female    Male
122835   77982
```





As shown above, women tend to have accidents in newest cars than men, accidents caused by men tend to be more deadlier, and accidents with more vehicles involved tend to be caused by women.

Analysis

To fulfill our objective, we proceeded with the first step which consisted in creating a copy of our data set omitting any NAs, and uploading the libraries necessary to implement the algorithms:

```
> df <- na.omit(Accidents)
> Library (caTools)
> Library (e1071)
> Library (Lattice)
```

Next, we divide the data into a training set consisting of the 80% of the data and a test set with the remaining 20%:

```
> split <- sample.split(df$DriversGender, SplitRatio = 0.8)
> train <- subset(df, split == TRUE)
> test <- subset(df, split == FALSE)
```

We then proceed to implement the ***Naïve Bayes*** algorithm and display the results:

```
> m <- naiveBayes(DriversGender ~., data = train, subset)
> p <- predict(m, test, threshold = 0.05)
> t <- ftable(Predicted=p, Correct=test$DriversGender)

> ftable(Predicted=p, Correct=test$DriversGender)
```

Correct	Female	Male
Predicted		
Female	18112	9065
Male	6455	6531

To finalize our investigation, we display the algorithm's accuracy based on the results:

```
> accuracy <- function(x){sum(diag(x))/sum(x)}
> accuracy(t)
  [1] 0.6135747
```

Conclusion

In conclusion, we successfully implemented an algorithm into our data set being capable of identifying drivers' genders based on the accident's characteristics they were involved. After we apply the algorithm we determine that it is possible to predict the gender but with a margin error between 35% - 40%. With the ***Naïve Bayes*** algorithm, we demonstrated that is possible but the accuracy is low; therefore, more variables are needed to actual determine a driver's gender with only the details of the accident.

We can also conclude that based on the statistics, women have a higher rate of accidents and bigger consequences than men in Great Britain. From the 27.177 females we studied, we predict that 18.112 are female and from the 12.986 men we studied, we predict 6.455, meaning that the algorithm had higher accuracy predicting females than men.

Predicted	Female	Male
Female	18112	9065
Male	6455	6531