**Machine Learning – Predict the type of transport**

*Read the train and test data set*

cars = read.csv("cars.csv")

Given sample data set containing 444 rows

carsTest = read.csv("test.csv")

Sample of two tests for which prediction must be done

*Data exploration and analysis*

str(cars)

'data.frame': 444 obs. of 9 variables:

$ Age : int 28 23 29 28 27 26 28 26 22 27 ...

$ Gender : Factor w/ 2 levels "Female","Male": 2 1 2 1 2 2 2 1 2 2 ...

$ Engineer : int 0 1 1 1 1 1 1 1 1 1 ...

$ MBA : int 0 0 0 1 0 0 0 0 0 0 ...

$ Work.Exp : int 4 4 7 5 4 4 5 3 1 4 ...

$ Salary : num 14.3 8.3 13.4 13.4 13.4 12.3 14.4 10.5 7.5 13.5 ...

$ Distance : num 3.2 3.3 4.1 4.5 4.6 4.8 5.1 5.1 5.1 5.2 ...

$ license : int 0 0 0 0 0 1 0 0 0 0 ...

$ Transport: Factor w/ 3 levels "2Wheeler","Car",..: 3 3 3 3 3 3 1 3 3 3 ...

Variables like Engineer, MBA and license has been read as numeric so should be converted to factors first.

cars$Engineer = as.factor(cars$Engineer)

cars$MBA = as.factor(cars$MBA)

cars$license = as.factor(cars$license)

**Descriptive Analysis**

summary(cars)

Age Gender Engineer MBA Work.Exp Salary Distance

Min. :18.00 Female:128 0:109 0 :331 Min. : 0.0 Min. : 6.50 Min. : 3.20

1st Qu.:25.00 Male :316 1:335 1 :112 1st Qu.: 3.0 1st Qu.: 9.80 1st Qu.: 8.80

Median :27.00 NA's: 1 Median : 5.0 Median :13.60 Median :11.00

Mean :27.75 Mean : 6.3 Mean :16.24 Mean :11.32

3rd Qu.:30.00 3rd Qu.: 8.0 3rd Qu.:15.72 3rd Qu.:13.43

Max. :43.00 Max. :24.0 Max. :57.00 Max. :23.40

license Transport

0:340 2Wheeler : 83

1:104 Car : 61

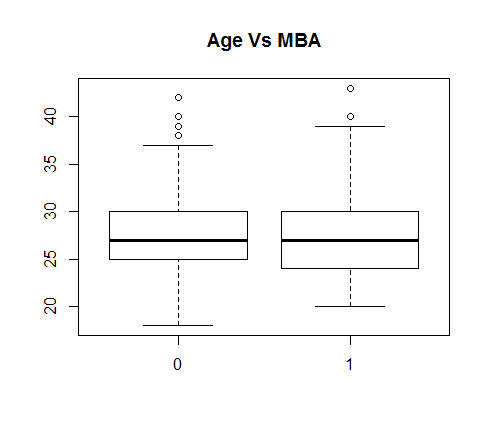
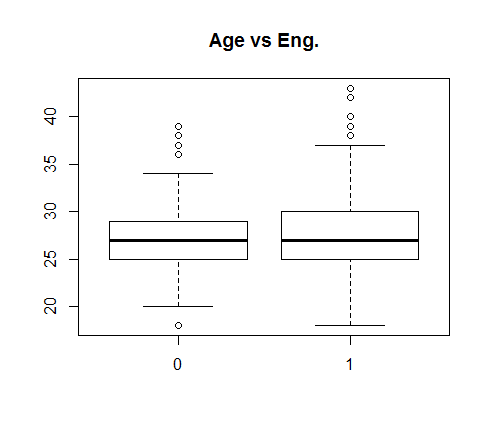
Public Transport:300

* We can conclude that we have majority of Males approx.. 75%
* Similarly Engineers outnumber MBA’s
* Total number of engineers and MBA’s is greater then 444, hence possibly some of candidates have dual degree
* One of data point for MBA is missing
* Salary might have skewed distribution
* Again, public transport is most common mode of transportation

**Visual Analysis**

boxplot(cars$Age ~cars$Engineer, main = "Age vs Eng.")

boxplot(cars$Age ~cars$MBA, main ="Age Vs MBA")



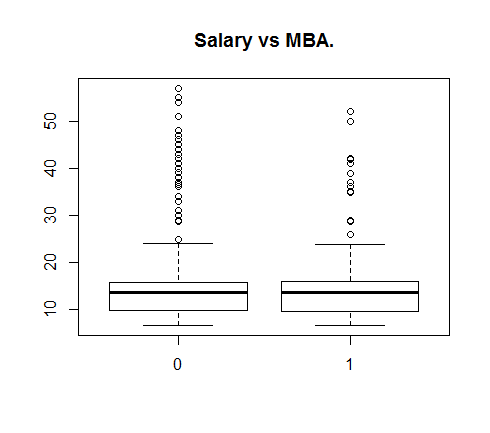
|  |
| --- |
|  |
|  |
| |  | | --- | |  | |

As expected not much of difference here, people for all qulaifications and all work exp would be employed in firm

Let us see the avg difference in salary for two profession

boxplot(cars$Salary ~cars$Engineer, main = "Salary vs Eng.")

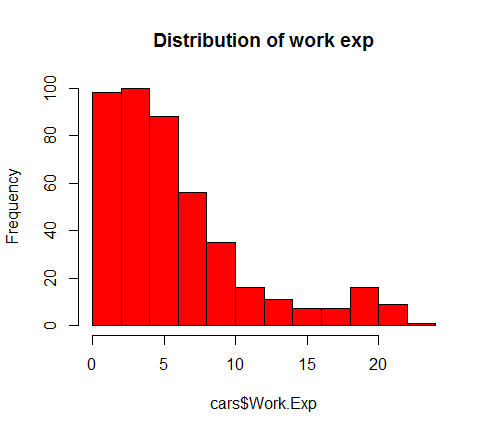
boxplot(cars$Salary ~cars$MBA, main = "Salary vs MBA.")



We do not see any appreciable difference in salary of Engs Vs Non-Engs or Mba vs Non-MBA’s

Also, mean salary for both MBA’s and Eng is around 16

|  |
| --- |
| hist(cars$Work.Exp, col = "red", main = "Distribution of work exp") |
|  |
| |  | | --- | |  | |



This is skewed towards right, again this would be on expected lines as there would be more juniors then seniors in any firm

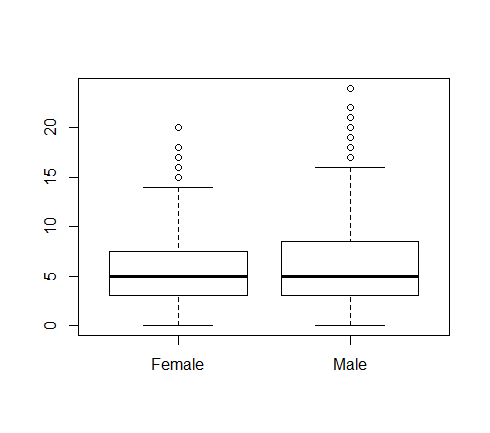
table(cars$license,cars$Transport)

2Wheeler Car Public Transport

0 60 13 267

1 23 48 33

boxplot(cars$Work.Exp ~ cars$Gender)

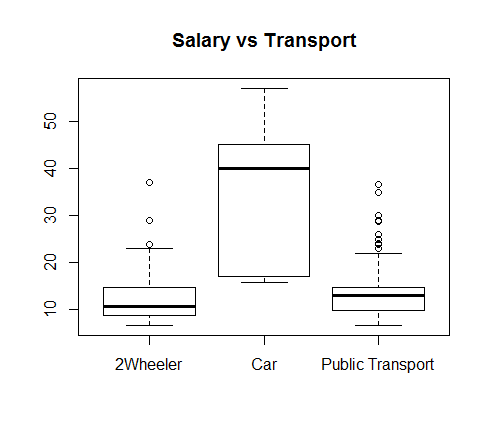


Not much of difference between mean work experience in two genders, so population is equally distributed for both male and females.

**Hypothesis Testing**

1. Higher the salary more the chances of using car for commute.

boxplot(cars$Salary~cars$Transport, main="Salary vs Transport")



Plot clearly shows as salary increase, inclination of commuting by car is higher.

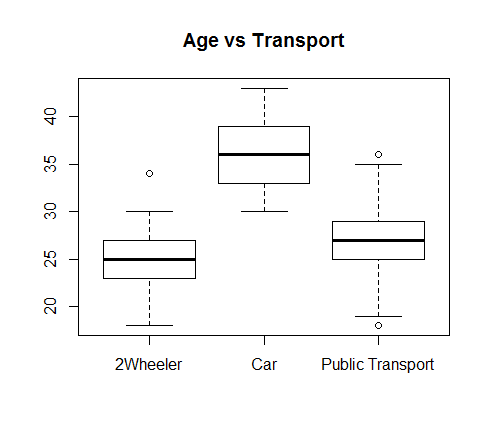
1. Again with age or work. Exp (Age and work exp would be collinear), propensity of using car

Increases

cor(cars$Age, cars$Work.Exp)

[1] 0.8408335

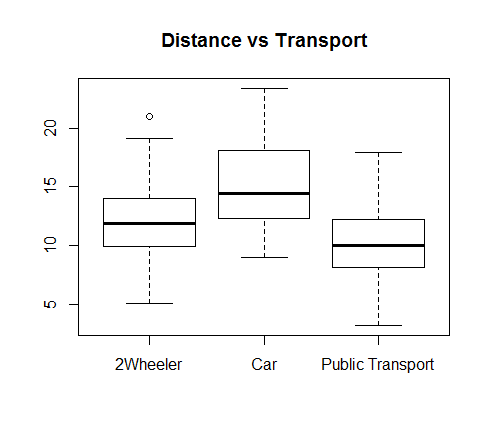
boxplot(cars$Age~cars$Transport, main="Age vs Transport")



As was the case with salary, we could see clear demarcation in usage of transport. With lower age group 2-wheeler is preferable and with higher work exp car is preferred.

1. As distance increase employee, would prefer car for comfort and ease.

boxplot(cars$Distance~cars$Transport, main="Distance vs Transport")



There is a slight pattern that could be observed here. For greater distance car is preferred followed by 2-wheeler and then public transport.

1. Females would prefer more of private transfer then public transport

table(cars$Gender,cars$Transport)

2Wheeler Car Public Transport

Female 38 13 77

Male 45 48 223

We could see that around 40 % of females use private transport and 10% use car compared to males where 15% prefers car and total of 30% uses private transport. Thus, even though percentage of car usage

is high but they are also high on public transport.

**Data cleaning**

Missing values

anyNA(cars)

[1] TRUE

Finding out where the missing value is

cars[!complete.cases(cars), ]

Age Gender Engineer MBA Work.Exp Salary Distance license Transport

145 28 Female 0 <NA> 6 13.7 9.4 0 PublicTransport

Use KNN means method to impute the missing value

library(DMwR)

cars = knnImputation(cars, 5)

Normalize continuous variables

cars$Salary = log(cars$Salary)

Perform similar transformation on test data

carsTest$Salary = log(carsTest$Salary)

carsTest$Engineer = as.factor(carsTest$Engineer)

carsTest$MBA = as.factor(carsTest$MBA)

carsTest$license = as.factor(carsTest$license)

Create test and train data from sample data

library(caret)

random <- createDataPartition(cars$Transport, p=0.70, list=FALSE)

cars\_train <- cars[ random,]

cars\_test <- cars[-random,]

This sample has all the three categories representation above 10% so we can go ahead without any oversampling

*Model Building and Predictions*

Naïve Bayes

library(e1071)

Naive\_Bayes\_Model=naiveBayes(cars\_train$Transport ~., data=cars\_train)

Naive\_Bayes\_Model

Naive Bayes Classifier for Discrete Predictors

Call:

naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

Y

2Wheeler Car Public Transport

0.1891026 0.1378205 0.6730769

Conditional probabilities:

Age

Y [,1] [,2]

2Wheeler 25.42373 2.620893

Car 35.72093 3.340413

Public Transport 26.73333 2.924134

Gender

Y Female Male

2Wheeler 0.4915254 0.5084746

Car 0.2558140 0.7441860

Public Transport 0.2761905 0.7238095

Engineer

Y 0 1

2Wheeler 0.2542373 0.7457627

Car 0.1395349 0.8604651

Public Transport 0.2714286 0.7285714

MBA

Y 0 1

2Wheeler 0.7966102 0.2033898

Car 0.7674419 0.2325581

Public Transport 0.7333333 0.2666667

Work.Exp

Y [,1] [,2]

2Wheeler 4.084746 3.114417

Car 15.674419 4.921870

Public Transport 4.866667 3.062559

Salary

Y [,1] [,2]

2Wheeler 2.452621 0.3659353

Car 3.514029 0.4321709

Public Transport 2.508357 0.3066213

Distance

Y [,1] [,2]

2Wheeler 11.92881 3.524009

Car 15.85581 3.864263

Public Transport 10.27286 3.090404

license

Y 0 1

2Wheeler 0.7288136 0.2711864

Car 0.2558140 0.7441860

Public Transport 0.8857143 0.1142857

This gives us the rule or factors which can help us employees decision to use car or not.

(These are summarized at the end)

General way to interpret this output is that for any factor variable say license we can say that 72% of

people without license use 2-wheeler and 27% with license.

For continuous variables for example distance we can say 2-wheeler is used by people for whom

commute distance is 11.9 with sd of 3.5

#Prediction on the test dataset

NB\_Predictions=predict(Naive\_Bayes\_Model,cars\_test)

table(NB\_Predictions,cars\_test$Transport)

NB\_Predictions 2Wheeler Car Public Transport

2Wheeler 8 0 6

Car 3 14 3

Public Transport 13 4 81

# prediction for test sample

NB\_Predictions=predict(Naive\_Bayes\_Model,carsTest)

NB\_Predictions

**[1] Public Transport Public Transport**

Levels: 2Wheeler Car Public Transport

**LDA**

We would once again import the two files and do data cleaning as required by LDA. LDA works best with continuous variables hence convert factors as 1 and 0.

cars = read.csv("cars.csv")

carsTest = read.csv("test.csv")

cars[145,4] = 0

Normalize continuous variables

cars$Salary = log(cars$Salary)

carsTest$Salary = log(carsTest$Salary)

cars$Gender<-ifelse(cars$Gender=="Male",1,0)

carsTest$Gender<-ifelse(carsTest$Gender=="Male",1,0)

random <- createDataPartition(cars$Transport, p=0.70, list=FALSE)

cars\_train <- cars[ random,]

cars\_test <- cars[-random,]

library(MASS)

fit.ld=lda(Transport~., data=cars\_train, cv=TRUE)

fit.ld

Call:

lda(Transport ~ ., data = cars\_train, cv = TRUE)

Prior probabilities of groups:

2Wheeler Car Public Transport

0.1891026 0.1378205 0.6730769

Group means:

Age Gender Engineer MBA Work.Exp Salary Distance license

2Wheeler 25.42373 0.5593220 0.7288136 0.1694915 4.186441 2.450022 11.56102 0.2372881

Car 35.67442 0.7441860 0.8139535 0.1860465 15.790698 3.536208 15.50000 0.7906977

Public Transport 26.76190 0.7666667 0.7285714 0.2857143 4.980952 2.515765 10.35238 0.1190476

Coefficients of linear discriminants:

LD1 LD2

Age -0.11042612 -0.3860466

Gender 0.25706348 -1.3517327

Engineer -0.14185048 0.2586975

MBA 0.18988407 -0.7316381

Work.Exp -0.07413621 0.2145325

Salary -0.58477768 -0.5036353

Distance -0.10677304 0.1340226

license -1.11223223 1.5268154

Proportion of trace:

LD1 LD2

0.9029 0.0971

Almost similar output as in Naïve Bayes

Predictions and accuracy

LDA\_predictions = predict(fit.ld,cars\_train)

table(LDA\_predictions$class, cars\_train$Transport)

2Wheeler Car Public Transport

2Wheeler 18 0 11

Car 3 36 3

Public Transport 38 7 196

LDA\_predictions = predict(fit.ld,cars\_test)

table(LDA\_predictions$class, cars\_test$Transport)

2Wheeler Car Public Transport

2Wheeler 11 0 6

Car 1 14 1

Public Transport 12 4 83

predict(fit.ld,carsTest)

$class

[**1] Public Transport Public Transport**

Levels: 2Wheeler Car Public Transport

$posterior

2Wheeler Car Public Transport

1 0.2036210 7.228535e-05 0.7963068

2 0.2078997 5.165238e-06 0.7920952

$x

LD1 LD2

1 0.7702525 0.2470294

2 1.4835708 0.3306443

**KNN**

cars = read.csv("cars.csv")

carsTest = read.csv("test.csv")

cars[145,4] = 0

Normalize continuous variables

cars$Salary = log(cars$Salary)

carsTest$Salary = log(carsTest$Salary)

cars$Gender<-ifelse(cars$Gender=="Male",1,0)

carsTest$Gender<-ifelse(carsTest$Gender=="Male",1,0)

random <- createDataPartition(cars$Transport, p=0.70, list=FALSE)

cars\_train <- cars[ random,]

cars\_test <- cars[-random,]

library(class)

trControl <- trainControl(method = "cv", number = 10)

fit.knn <- train(Transport ~ .,

+ method = "knn",

+ tuneGrid = expand.grid(k = 2:20),

+ trControl = trControl,

+ metric = "Accuracy",

+ preProcess = c("center","scale"),

+ data = cars\_train)

fit.knn

k-Nearest Neighbors

312 samples

8 predictor

3 classes: '2Wheeler', 'Car', 'Public Transport'

Pre-processing: centered (8), scaled (8)

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 281, 281, 280, 281, 280, 282, ...

Resampling results across tuning parameters:

k Accuracy Kappa

2 0.7365457 0.4543489

3 0.7855712 0.5248631

4 0.7629839 0.4800127

5 0.7828562 0.5081854

6 0.7734812 0.4905393

7 0.7634005 0.4624704

8 0.7408065 0.4118105

9 0.7534005 0.4199273

10 0.7536022 0.4116860

11 0.7598454 0.4168749

12 0.7662970 0.4266860

13 0.7662970 0.4213708

14 0.7566129 0.3930122

15 0.7661895 0.4135919

16 0.7660887 0.4090611

17 0.7566129 0.3862387

18 0.7629637 0.3926229

19 0.7661895 0.4026549

20 0.7661895 0.3942178

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was k = 3.

KNN\_predictions = predict(fit.knn,cars\_train)

table(KNN\_predictions, cars\_train$Transport)

KNN\_predictions 2Wheeler Car Public Transport

2Wheeler 37 0 8

Car 0 35 2

Public Transport 22 8 200

KNN\_predictions = predict(fit.knn,cars\_test)

table(KNN\_predictions, cars\_test$Transport)

KNN\_predictions 2Wheeler Car Public Transport

2Wheeler 9 0 11

Car 1 15 3

Public Transport 14 3 76

predict(fit.knn,carsTest)

**[1] Public Transport Public Transport**

Levels: 2Wheeler Car Public Transport

We see that all three models predict **Public Transport** for the two test samples

Let us summarize the conclusions from analysis and models for employee’s decision whether to use car

Or not:

* Important variables are Age, Work.Exp, Distance and License
* Age and Work.Exp are correlated hence we could use any one (prefer Work.Exp) here
* Hence employees with work exp of 10 and above are likely to use car
* Employees who must commute for distance greater than 12 are more likely to prefer car
* With license, we do see that 74% who commute through car have license and 89% who commute through bus don’t have. But surprisingly 72% without license use 2-wheeler.
* Again, people with higher salaries (>20) are likely to use cars