Transport Prediction

Table of contents

1 Project Objective

The primary objective of this report is to find what mode of transport will the employees preferring to their office. The dataset includes employee information about their mode of transport as well as their personal and professional details like age, salary, work exp. Our aim is to predict whether or not an employee will use Car as a mode of transport. Also, which variables are a significant predictor behind this decision.

2 Exploratory data analysis

Generally, Exploratory data analysis are carried out to discover the patterns in the given dataset,missing values ,central tendency of the given dataset.

2.1 Environment set up and data Import

The working directory should be one , where the code and dataset are placed. The following R packages used for the analysis of Churn prediction dataset

- Ggplot2 package-To infer about the attributes of the loan prediction data using graphical plots
- caTools-To split the given dataset into training and testing
- corrplot-To plot the correlations between predictors
- Class-To build the K- Nearest Neighbours model
- e1071-Naïve bayes model building
- ROCR- To compute model performance measures such as KS,Area under curve, lift chart
- Ineq-To compute the gini index of data
- InformationValue-Package to calculate the concordance ratio

2.2 Variable Identification:

Some basics R functions such as mean , sd ,round are used for the statistical calculation. Here are the following functions are getting used for the better understanding of data and to make further decision **.**

Structure(Str)- To get the structure of the employee transport mode such as class category,name and the count of the fields

Summary –Generally , summary function will give the classification of attributes. Here, the transport mode prediction data having nine columns/ fields such as employee age,gender,engineer,MBA,working experience,salary etc .If the field is categorical data, summary function give the count of each sub-category If the field is integer, then it will return

the five stats notably minimum,25th percentile , median or 50th percentile , 75th percentile and maximum.

2.3 Uni-Variate Analysis:

Summary statistics

```
> summary(car_dataset)
      Age          Gender    Engineer MBA       Work.Exp       Salary        Distance     license Transport
 Min.   :18.00   Female:127   0:108    0:331   Min.   : 0.0   Min.   : 6.50   Min.   : 3.20   0:339   0:382
 1st Qu.:25.00   Male  :316   1:335    1:112   1st Qu.: 3.0   1st Qu.: 9.80   1st Qu.: 8.80   1:104   1: 61
 Median :27.00                                Median : 5.0   Median :13.60   Median :11.00
 Mean   :27.75                                Mean   : 6.3   Mean   :16.24   Mean   :11.33
 3rd Qu.:30.00                                3rd Qu.: 8.0   3rd Qu.:15.75   3rd Qu.:13.45
 Max.   :43.00                                Max.   :24.0   Max.   :57.00   Max.   :23.40
>
```

Frequency distribution of dependent variable

Category	Count
Car(0)	**382**
Non-car(1)	**61**
Total	**443**

Target variable- Transport are with three levels(Car,Two wheeler and public transport).To do binomial logistic regression,converted the three levels to two levels as car and non-car(Two wheeler and public transport). Among these 443 employees, 14 percentage are using car as mode of transport to their office.

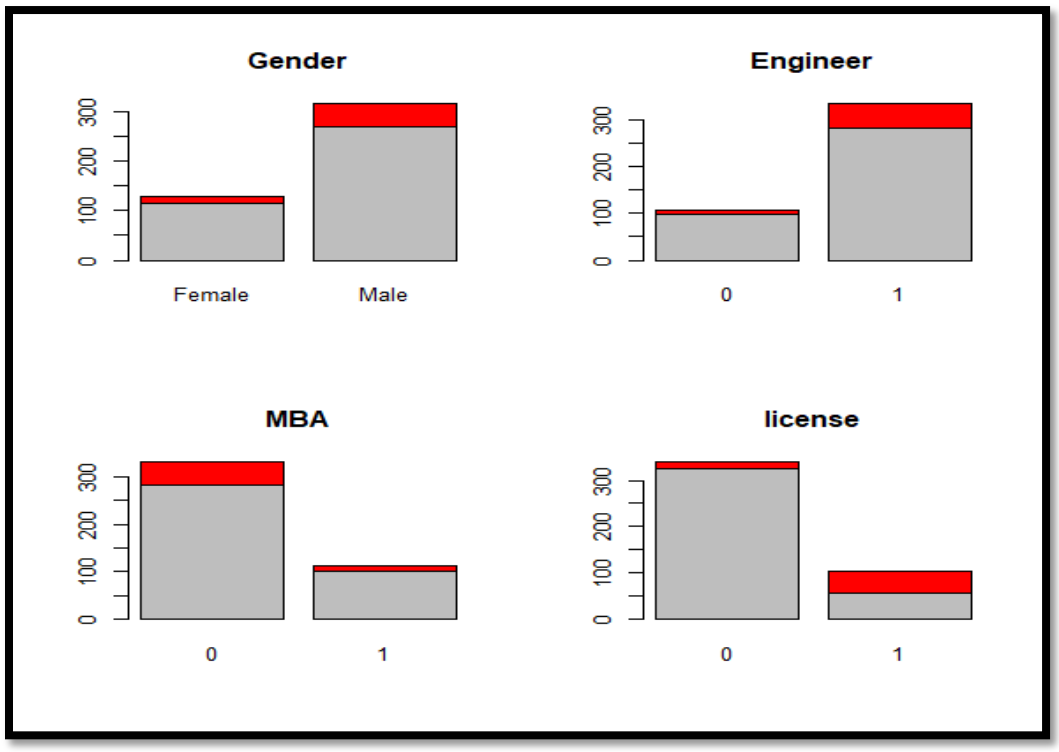
Frequency distribution of independent variables

Continous variables:

Categorical variables:

2.4 Bi-Variate Analysis:

Relationship between Categorical IV's and target variable



The above plots will give the relationship between the categorical IV(Independent Variable) and target variable

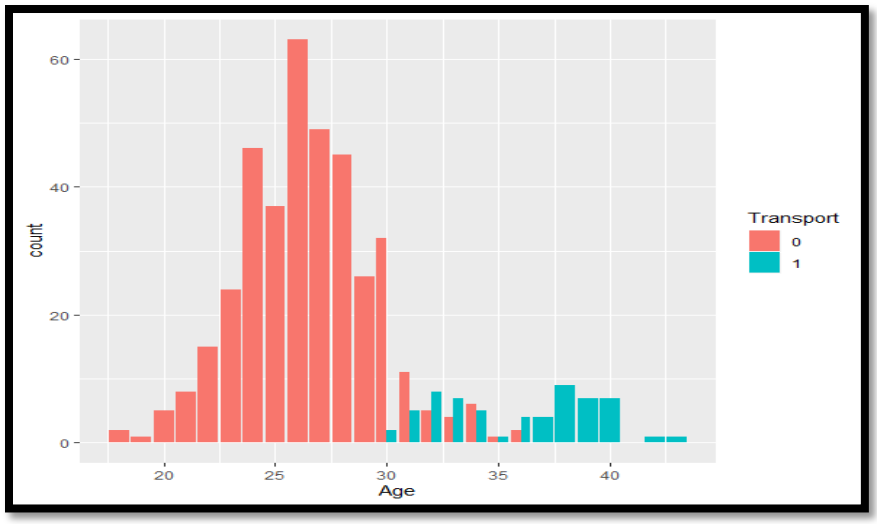
Red color – Employees who prefers car(ie car=1)

Gray color – Employees who is not preferring car(ie non-car=0)

Among the given categorical variables, license seems to be significant,because without license one can't prefer car.From the given plot , we can say that male employees preferring car's more when compared to female employee's.

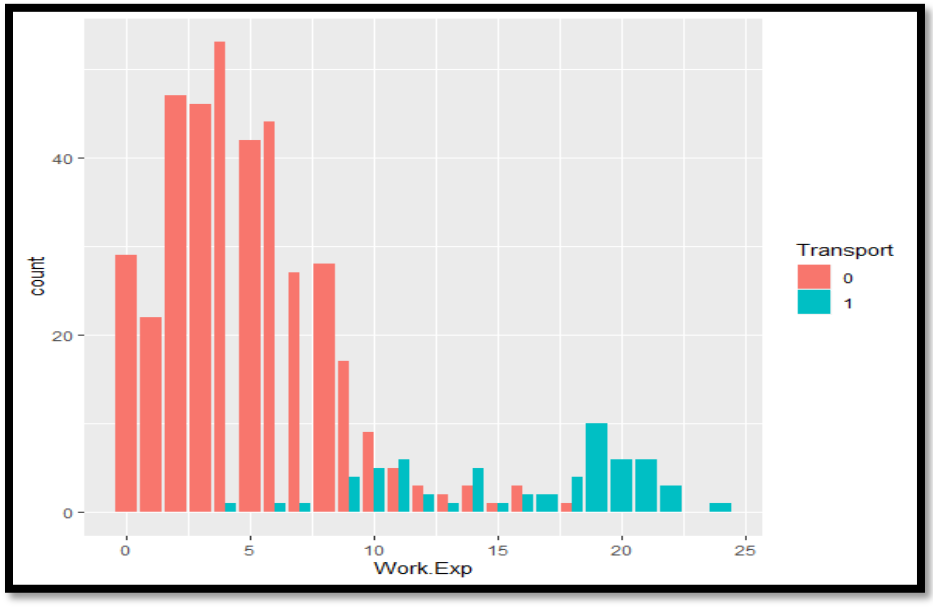
Relationship between Predictors and Target variable

Employee's Age vs target variable



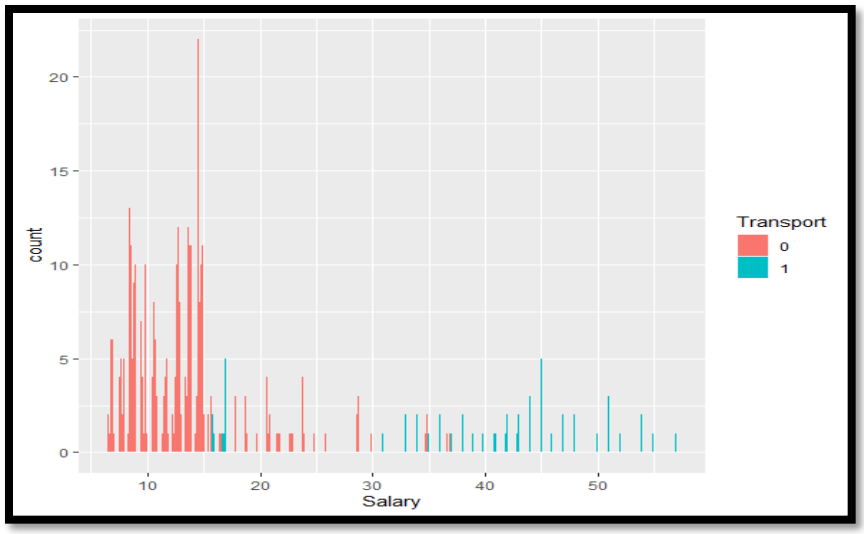
From the given bar plot , it is observed that people below 30 years are preferring two wheelers or public transport .Employee's whose age is greater than 35 are only using car as their mode of transport to the office

Work Experience vs Target variable



From the above plot ,employee with working experience more than 15 years are using car for the office.But 90% of employee's are preferring public transports and two wheelers.

Salary vs Target variable:

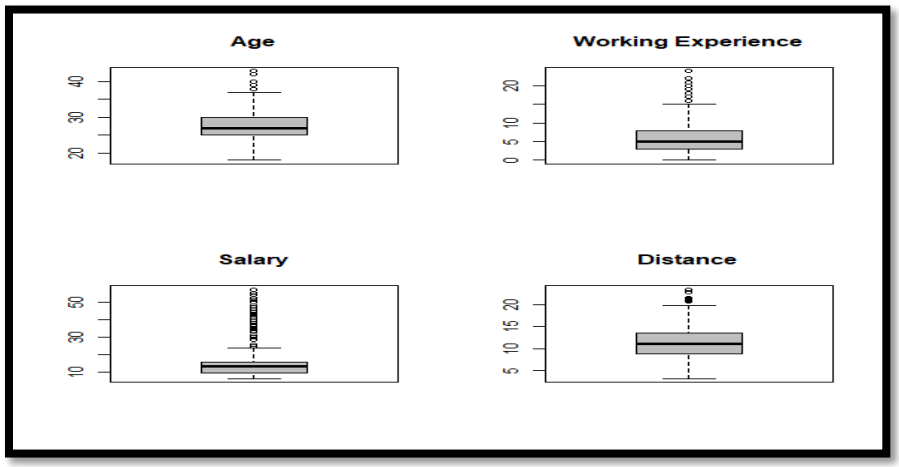


Employee who is earning more money(>30) are using car more than other employee's.

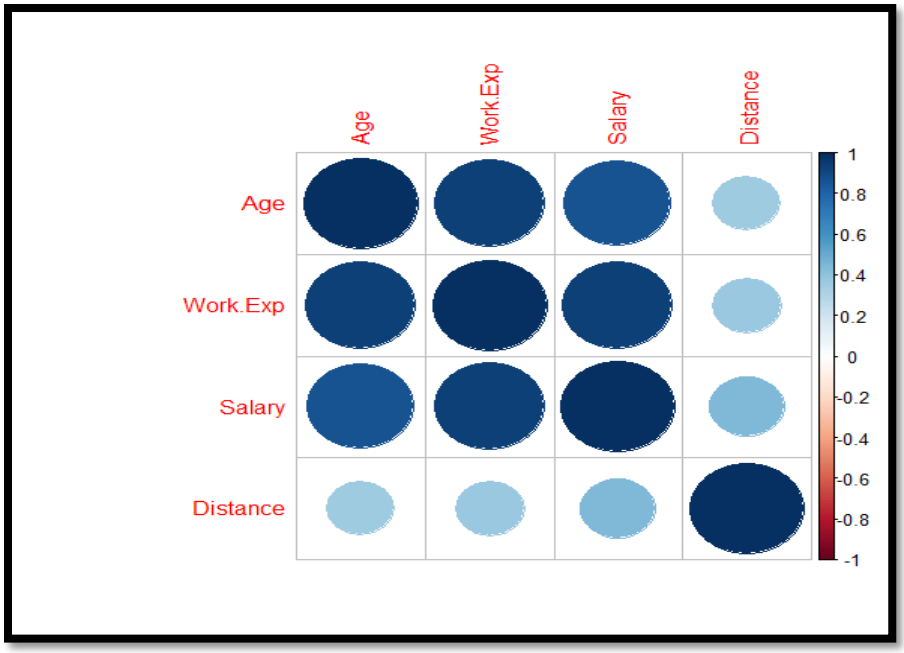
Missing values:

```
> #Checking for missing values
> sum(is.na(car_dataset))
[1] 1
> #when MARGIN=1, it applies over rows, whereas with MARGIN=2, it works over columns.
> #Note that when you use the construct MARGIN=c(1,2), it applies to both rows and columns;
> apply(car_dataset,2,function(x) sum(is.na(x)))
      Age    Gender  Engineer       MBA  Work.Exp    Salary  Distance   license Transport
        0         0         0         1         0         0         0         0         0
> car_dataset=na.omit(car_dataset)#Remove the NA value
>
```

In the given data,predictor variable –MBA is having NA value and it is removed using the omit function . From the boxplot ,it is infer that there is an outlier in the car dataset.



Multi-Collinearity:



The Independent variables which are highly correlated are refered to as multi-collinearity problem. Following methods can be used to detect multicollinearity:

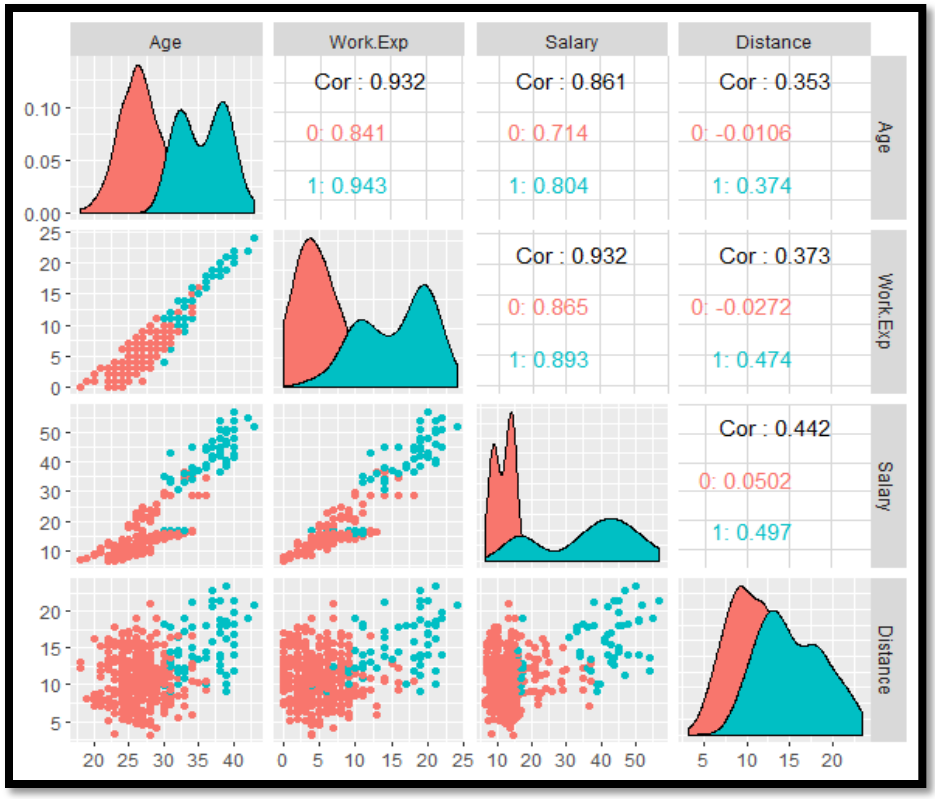
- The analysis exhibits sign of multicollinearity - such as, coefficient estimates varying from model to model
- The R-square value is large but none of the beta weights is statistically significant, i.e., F-test for overall model is significant but the t-tests for individual coefficient estimates are not.
- Correlation among pairs of variables are large.
- Variance Inflation Factor.

In the given dataset ,Age , working experience and salary are highly correlated. Based on the p-value ,salary field is more significant and removed the other two from the model .

Inferences from EDA:

From the statistical model , it is observed that salary , license and gender plays a major role in the decision making. Employee's salary ,age and working experience naturally correlated.

To Check the visual association pattern for continous predictor variables with target variable



3)SMOTE:

SMOTE(Synthetic Minority Over Sampling Technique) is the technique to balance the dataset by synthetically over sampling the minority class or it's a method to handle the imbalance problems.To generate artificial data, it uses bootstrapping and k-nearest neighbors. Steps to perform SMOTE

- Take the difference between the feature vector (sample) under consideration and its nearest neighbor.
- Multiply this difference by a random number between 0 and 1
- Add it to the feature vector under consideration
- This causes the selection of a random point along the line segment between two specific features

```
> ########################Data Preparation -SMOTE############################
> round(sum(car_dataset_train$Transport==1)/nrow(car_dataset_train),4)#14% use car as a mode of transport
[1] 0.1419
> library(DMwR)#DMwR-Data Mining with R
> table(car_dataset_train$Transport)

  0   1
254  42
> Smote_data=SMOTE(car_dataset_train$Transport~.,car_dataset_train,perc.over = 53,perc.under =1000,k=3)
> #After performing SMOTE
> round(sum(Smote_data$Transport==1)/nrow(Smote_data),4)
[1] 0.2254
> table(Smote_data$Transport)

  0   1
220  64
>
```

In the R , DMwR package is used to perform Smote.From the given car dataset , it is observed that 13% people are preferring car and 87% employee's preferring the rest of the transport.In this case,the data seems to be imbalanced, so there is need to create the data artificially.

The SMOTE function is comprised of perc over and perc under paramaters, both should be filled in such a way that its not affecting the originality of data.After performing the SMOTE,car transport mode to 22 percentage.So now the data is in the ratio of 80 and 20.

4)Supervised Learning Techniques:

4.1)Logistic Regression Model

4.1.1)Model without SMOTE:

Logistic regression can be used for both classification and regression models.Mostly, it will be used for classification problems and because of that it is known as Binomial classifier.It is the base model for the classification technique.

glm is generally used to fit generalized linear models.However, in this case, you need to make it clear that you want to fit a logistic regression model, resolve this by setting the family argument to binomial.

Building the model with all predictors:

Summary() function returns the estimate, standard errors, z-score, and p-values on each of the coefficients.Look like most of the co-efficients are not significant here while building the model with all predictors.

```
Call:
glm(formula = car_dataset_train$Transport ~ ., family = "binomial",
    data = car_dataset_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.95810  -0.03538  -0.00600  -0.00040   2.56271

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -67.4436    21.6609  -3.114  0.00185 **
Age           2.0695     0.7301   2.835  0.00459 **
GenderMale   -1.9216     1.3316  -1.443  0.14899
Engineer1    -0.0255     1.3092  -0.019  0.98446
MBA1         -1.5338     1.1168  -1.373  0.16964
Work.Exp     -1.0178     0.5259  -1.935  0.05296 .
Salary        0.2044     0.1022   2.000  0.04555 *
Distance      0.5186     0.2043   2.538  0.01114 *
license1      3.6315     1.2116   2.997  0.00272 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 241.763  on 295  degrees of freedom
Residual deviance:  33.381  on 287  degrees of freedom
AIC: 51.381

Number of Fisher Scoring iterations: 10
```

p value determines the probability of significance of predictor variables. With 95% confidence level, a variable having p < 0.05 is considered an important predictor. Employee's gender and their educational qualifications are seems to be insignificant , we can say it from the above model.

Variation inflation factor(VIF):

```
> vif(log_model_1)
      Age    Gender  Engineer       MBA  Work.Exp    Salary  Distance   license
12.999115  1.535939  1.141918  1.386080 17.100789  4.174523  1.712523  1.742478
>
```

Eliminating the insignificant variables with the help of vif values.

Interpretation:

Build the model for Significant variables:

```
Call:
glm(formula = car_dataset_train$Transport ~ Salary + Distance +
    license, family = "binomial", data = car_dataset_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.14288  -0.16885  -0.09627  -0.05040   3.04160

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.63035    1.91644  -5.547 2.91e-08 ***
Salary        0.19402    0.03671   5.286 1.25e-07 ***
Distance      0.27085    0.11593   2.336 0.019476 *
license1      2.84401    0.73179   3.886 0.000102 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 241.763  on 295  degrees of freedom
Residual deviance:  65.038  on 292  degrees of freedom
AIC: 73.038

Number of Fisher Scoring iterations: 7
```

The logistic regression coefficients give the change in the log odds of the outcome for a unit increase in the predictor variable. For every unit increase in predictors , the log odds of employee's mode of transport as car increased by predictor cofficients.

Odds Ratio:

Exponentiate the coefficients and interpret them as odds-ratio.

```
> #Odds ratio
> exp(log_model_3$coefficients)
 (Intercept)      Salary    Distance    license1
2.417123e-05 1.214122e+00 1.311079e+00 1.718462e+01
>
```

Prediction in Logistic Regression:

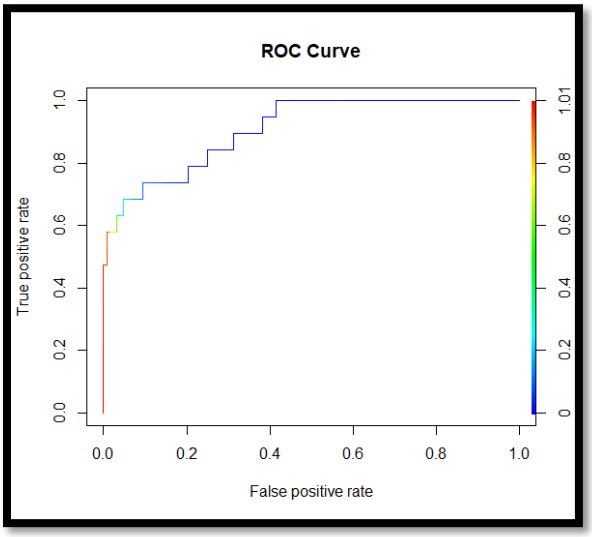
Given dataset was getting splitted in the ratio of 70 and 30.Predicting the results for the 30% of given dataset.Response's are decided based on probability threshold value.Initially the threshold value was set as 0.50.Even it shows good of amount accuracy,the given data is imbalanced data.So we can't conclude the performance of model only with accuracy.

Sensitivity of the logistic model is 0.57.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 124   8
         1   4  11

               Accuracy : 0.9184
                 95% CI : (0.8617, 0.9571)
    No Information Rate : 0.8707
    P-Value [Acc > NIR] : 0.04879

                  Kappa : 0.6016

 Mcnemar's Test P-Value : 0.38648

            Sensitivity : 0.57895
            Specificity : 0.96875
         Pos Pred Value : 0.73333
         Neg Pred Value : 0.93939
             Prevalence : 0.12925
         Detection Rate : 0.07483
   Detection Prevalence : 0.10204
      Balanced Accuracy : 0.77385

       'Positive' Class : 1
```

ROC Plot:

The ROC curve is the plot between sensitivity and (1- specificity). (1- specificity) is also known as false positive rate and sensitivity is also known as True Positive rate.



AUC represents the degree or measure of seperablity.Area under the curve is 90% in this case , ie 90% of the time the current model is predicting 0 as 0 and 1 as 1.

4.1.2)Logistic model with SMOTE data:

```
> car_dataset_test_smote$prob_test=predict(log_model_smote,newdata=car_dataset_test_smote,type="response")
> car_dataset_test_smote$Predict_result=ifelse(car_dataset_test_smote$prob_test>0.5,1,0)
> car_dataset_test_smote$Predict_result=as.factor(car_dataset_test_smote$Predict_result)
> confusionMatrix(car_dataset_test_smote$Predict_result,car_dataset_test_smote$Transport,positive="1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 123   8
         1   5  11

               Accuracy : 0.9116
                 95% CI : (0.8535, 0.9521)
    No Information Rate : 0.8707
    P-Value [Acc > NIR] : 0.0833

                  Kappa : 0.5788

 Mcnemar's Test P-Value : 0.5791

            Sensitivity : 0.57895
            Specificity : 0.96094
         Pos Pred Value : 0.68750
         Neg Pred Value : 0.93893
             Prevalence : 0.12925
```

After increasing the minority class from 14% to 22% , there is no much difference in predicting the results.Pos predicted value is reduced to 0.68 from 0.73(without smote).So for the given car dataset,we don't need to create the data artificially ie no need of SMOTE

4.2)K-Nearest Neighbours Model:

KNN is a Supervised Learning algorithm that uses labeled input data set to predict the output of the data points. It is mainly based on feature similarity. KNN checks how similar a data point is to its neighbor and classifies the data point into the class it is most similar to.

Knn() function-The mandatory paramters are

1)Training dataset

2)Testing dataset

3)Target variable

4)K-Value

Optimal value of k:

K values are calculated from the input features. One of the ways to find the optimal K value is to calculate the square root of the predictors in the data set.

Total number of input variables=9

Square root of 9 is 3.So the optimal k-value is 3

Interpreation:

car_dataset_test_KNN$Status=knn(scale(car_dataset_train_KNN[,c(1,5,6,7)]),scale(car_dataset_test_KNN[,c(1,5,6,7)]),car_dataset_train_KNN$Transport,k=3)

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 128   3
         1   0  17

               Accuracy : 0.9797
                 95% CI : (0.9419, 0.9958)
    No Information Rate : 0.8649
    P-Value [Acc > NIR] : 1.076e-06

                  Kappa : 0.9074

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 0.8500
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9771
             Prevalence : 0.1351
         Detection Rate : 0.1149
   Detection Prevalence : 0.1149
      Balanced Accuracy : 0.9250
```

Euclidean distance is used to measure the similarities between the predictors.For the classification problem,majority vote determine the response classes.

From the confusion matrix , it is inferred that KNN model predicted 82%(sensitivity) correctly from 3333 customers where the employee's are using car as mode of transport.

4.3)Naïve Baye's Model:

Naïve Baye's algorithm is basically derived from Baye's theorem,Where the classification of responses are based on conditional probability.Prior probability and evidence will be given,condition is to find the posterior probability.

Model Assumptions:

Naïve Bayes assumes all the features to be conditionally independent. So, if some of the features are in fact dependent on each other (in case of a large feature space), the prediction might be poor.

In other words, it assumes that the presence of one feature in a class is completely unrelated to the presence of all other features. If this assumption of independence holds, Naive Bayes performs extremely well and often better than other models.

Naïve Baye's for continuous variables:

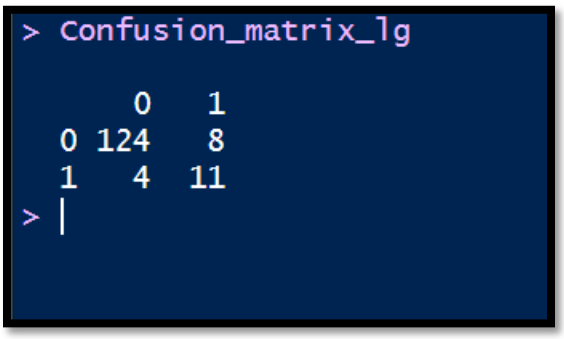
Naive Bayes can also be used with continuous features but is more suited to categorical variables. If all the input features are categorical, Naive Bayes is recommended. However, in case of numeric features, it makes another strong assumption which is that the numerical variable is normally distributed.

Interpretation of Naïve Baye's:

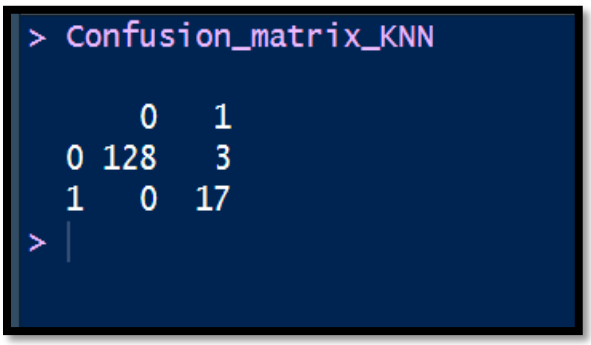
The features of the car dataset is highly dependent to each other. So we can't build the naïve bayes model.

5)Confusion matrix interpretation:

Logistic regression



KNN Model



Logistics regression:

The test data consisted of 147 observations. Out of which 124 cases have been accurately predicted (TN->True Negatives) as non-car(ie employee's who all not preferring car) in nature .Also, 11 out of 147 observations were accurately predicted (TP-> True Positives) transport mode as car

In the 4 cases,the model predicted the data wrongly (ie it predicted employess transport mode as non-car ,but it actually those employees preferred car) this is False Negatives (FN)

There were 8 cases of False Positives (FP) meaning 8 cases were actually not employee's is preferring car in nature but got predicted as preferred one.

The total accuracy of the model is 91%((TN+TP)/total test data) which shows that there may be chances to improve the model performance

KNN Model:

The test data consisted of 147 observations. Out of which 128 cases have been accurately predicted (TN->True Negatives as non-car(ie employee's who all not preferring car) in nature.Also, 17 out of 147 observations were accurately predicted (TP-> True Positives)) transport mode as car

There were no cases of False Negatives (FN) meaning no cases were recorded which actually as car in nature but got predicted as not car.

There were 3 cases of False Positives (FP) meaning 3 cases were actually not employee's is preferring car in nature but got predicted as preferred one.

The total accuracy of the model is 97%((TN+TP)/total test data) which shows that KNN model is predicting correctly the results when compared to logistic regression.

6)Model Comparison

Model	Accuracy	Sensitivity	Specificity
Logistic Regression	91%	57%	96%
KNN	97%	**85%**	100%

The given dataset is imbalanced dataset, so only with we can't conclude the best model.The other performance metrics such as sensitivity is playing major role in the decision making.The sensitivity of the knn model is very high when compared to compared to the logistic regression.so knn is the best model than logistic regression.

7)Ensemble Methods

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to **decrease variance** (bagging), **bias** (boosting), or **improve predictions** (stacking).

7.1)Bagging

Bagging stands for bootstrap aggregation. It is an sequential ensemble method ,where the base learners are generated sequentially.

```
##Bagging

library(rpart)
car_dataset_bag=car_dataset
sample_bag=sample.split(car_dataset_bag,SplitRatio = 0.7)
car_dataset_bag_train=subset(car_dataset_bag,sample_bag==TRUE)
car_dataset_bag_test=subset(car_dataset_bag,sample_bag==FALSE)
bag_model=bagging(car_dataset_bag_train$Transport~.,data=car_dataset_bag_train,
                  control=rpart.control(maxdepth = 5,minsplit = 4))

car_dataset_bag_test$Status=predict(bag_model,car_dataset_bag_test)
confusionMatrix(car_dataset_bag_test$Status,car_dataset_bag_test$Transport,positive="1")
```

This technique generally use decision tree .To built the model Rpart package is used.The rpart.control function parameters are max depth and min split.Both parameters values are such a way,the model predicts the data accurately.

Interpretation

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 127   2
         1   1  18

               Accuracy : 0.9797
                 95% CI : (0.9419, 0.9958)
    No Information Rate : 0.8649
    P-Value [Acc > NIR] : 1.076e-06

                  Kappa : 0.9114

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9000
            Specificity : 0.9922
         Pos Pred Value : 0.9474
         Neg Pred Value : 0.9845
             Prevalence : 0.1351
         Detection Rate : 0.1216
   Detection Prevalence : 0.1284
      Balanced Accuracy : 0.9461

       'Positive' Class : 1
```

From the confusion matrix , it is observed that sensitivity are relatively high when compared to rest of the models.In the comparison of logistic and KNN ,we concluded that knn is the best model.

In the knn model,the true positive rate is 85% whereas in the bagging technique , it was increased by 5%(ie 90%).The bagging method is predicting the employee's transport mode or their preference's.

7.2)Boosting

Boosting is the parallel ensemble method,there are three types of boosting.

1)Adaboost

2)Gradient boosting

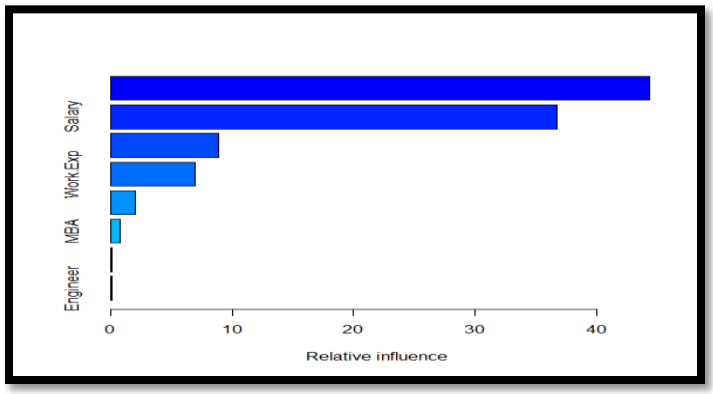
3)XGBoost

```
##Boosting
library(gbm)
car_dataset_boost=car_dataset
sample_boost=sample.split(car_dataset_boost,SplitRatio = 0.7)
car_dataset_boost_train=subset(car_dataset_boost,sample_boost==TRUE)
car_dataset_boost_test=subset(car_dataset_boost,sample_boost==FALSE)
car_dataset_boost_train$Transport=as.character(car_dataset_boost_train$Transport)
boost_model=gbm(car_dataset_boost_train$Transport~.,data=car_dataset_boost_train,distribution = 'bernoulli',n.t
summary(boost_model)
car_dataset_boost_test$prob_result=predict(boost_model,car_dataset_boost_test,type="response",n.trees = 3000)
car_dataset_boost_test$Status=ifelse(car_dataset_boost_test$prob_result>0.5,1,0)
car_dataset_boost_test$Transport=as.factor(car_dataset_boost_test$Transport)
car_dataset_boost_test$Status=as.factor(car_dataset_boost_test$Status)
confusionMatrix(car_dataset_boost_test$Status,car_dataset_boost_test$Transport,positive="1")
```

Gbm pacakage is used to develop the adaboost technique and gradient boosting.In this case, adaboost algorithm is applied , Bernoulli distribution is used because we need the results for logistic regression .Number of trees is by default 10000 , by changing the number of trees to get the best model.

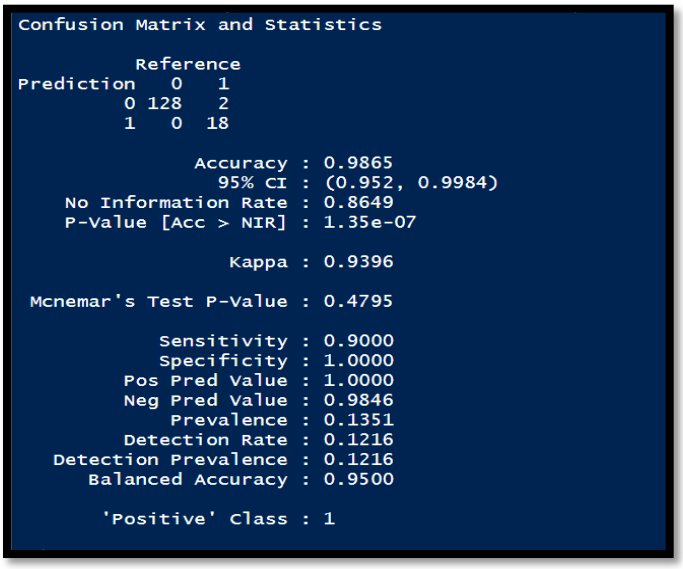
Interpretation:

Variable importance plot



The above plot gives the important predictors,so in this case the employee's salary plays a major role in determining the model performance.

Confusion matrix:



After applying the boosting technique,accuaracy of the model is 98% which is pretty good when compared to the rest of the models.The true postive rate is 90% (ie 90% of the times, boosting is predicting data as it was actual).Finally it is concluded boosting is the best technique to enhance the accurate prediction.

Conclusion:

Naïve Baye's can't be used for the given dataset,due to its dependencies between predictors. Even though Logistic regression is an binomial classifier,it's sensitivity and specificity are not fair as Knn model.By applying the ensemble technique,true positive rate and negative rates are increased.

Generally , employee's prefer public transport most of the times.Sometimes, gender also plays a major role,male employee's generally taking car a mode of transport to the office .

Employee's age , working experience and salary are mostly correlated.If the person is having more working experience ,which indirectly means that particular person is getting high salary and his/her age will be high and vice versa.

Educational qualifications of employee is not an significant predictors in this scenario.Distance should be the important predictor, because if the travelling distance is too high,most of the employee's prefer car.

License is most significant variable ,without license one can't drive car as per the rule.So based on salary license and distance we can predict the transport mode of an employee in the future.