**Mini Project – mode of transport employees prefers to commute to their office Prediction**

**13th - October – 2019**

**Submitted**

By: Shilpa pandey

Contents page No

**1. Project Objectives** .................................................................................................................................

**2. Steps and approach** ..............................................................................................................................

**3. assumption** ................................................................................................................... 4

**4 explanatory Data Analysis**.................................................................................................................................. 4

**5 Data description** ................................................................................................................................ 5

**6. Univariate analysis**............................................................... 5

7. **bivariate analysis**...............................................................

**8. outlier detection** .................................................................................................................................... 5

**9. check missing value and treatment** ................................................................................................................................. 5

**10. check multicollinearity** ........................................................................................................................... 5

**11. summarize the insights** ........................................................................................................................ 6

**12 Logistic Regression**........................................................................................... 6

**13. interpret Logistic Regression** ..................................................................................................................................... 6

**14. applying KNN Model** ............................................................................................................................................... 7

**15. Interpret KNN Model**............................................................................................................................ 7

**16. Confusion matrix interpretation for all models**...................................................................................................... 8

**17. Interpret Naive Bayes Model**......................................................................................................................................... 9

**18. Bagging** .......................................................... 9

**19. Boosting** 10

**18.SMOTE. .......................................................... 9**

**19. best model ………………………………………….. 10**

**20. Actionable Insights and Recommendations**

**1. Project Objectives**

This project requires you to understand what mode of transport employees prefers to commute to their office. The attached data '[Cars.csv](https://olympus.greatlearning.in/courses/4971/files/435759/download?wrap=1)' includes employee information about their mode of transport as well as their personal and professional details like age, salary, work exp. We need 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. Steps and approach**

We shall follow step by step approach to arrive to the conclusion as follows:

* EDA - Basic data summary, Univariate, Bivariate analysis, graphs, Check for Outliers and missing values and check the summary of the dataset
* EDA - Illustrate the insights based on EDA
* EDA - Check for Multicollinearity - Plot the graph based on Multicollinearity & treat it.
* Applying Logistic Regression & Interpret results
* Applying KNN Model & Interpret results
* Applying Naïve Bayes Model & Interpret results (is it applicable here? comment and if it is not applicable, how can you build an NB model in this case?)
* Confusion matrix interpretation
* Remarks on Model validation exercise <Which model performed the best>
* Bagging
* Boosting
* Actionable Insights and Recommendations

**3 Assumptions**

* none

**4 Exploratory Data Analysis – Step by Step approach**

* 1. Environment Set Up and Data Import
     1. Install Necessary Packages and Invoke Libraries

* + 1. Set up working Directory

* + 1. Import and read the dataset
  1. Variable Identification
     1. Variable identification – inferences

|  |  |
| --- | --- |
| **5. Data Description:** |  |

Str function indicates all the var are numerical and integer but we will change into factor as below

Machine generated alternative text:
nsole 
The following objects are masked 
cr edit. score 
from 
work. 
car 
Exp 
(pos 
Gender, work. Exp 
> head(car) 
Gender Loan. offered 
Job 
unski 11 ed 
ski 11 ed 
unski 1 1 ed 
ski 11 ed 
ski 11 ed 
EMI. Ratio 
status 
credit. History own. house 
Male 
2 Female 
Male 
4 Female 
Male 
6 Female 
> str (car) 
• data. frame': 
Gender 
L oan. offer ed 
S Job 
S work. Exp 
credit. score 
EMI. Ratio 
status 
O 
O 
poor 
poor 
poor 
poor 
poor 
cr iti cal 
1 
1 
1 
1 
1 
1 
Purpose Dependents 
personal 
personal 
personal 
personal 
personal 
personal 
1 Management 
781 obs. of 11 vari ables: 
. Factor w/ 2 levels "Female" , "Male"• 
.21 
212 
11222... 
. Factor w/ 3 levels "Management" , "skilled" , 
: int 14 15 16 13 12 12 15 12 13 13 . 
: int 86 94 86 94 85 86 86 85 87 89 
. num 3 
Factor 
Factor 
: int 1 
Factor 
: int 2 
333 3.3 3.6 3.6 3.6 3.94. 
credit. Hi story: 
own. house 
Purpose 
Dependents 
w/ 2 levels 
w/ 6 levels 
w/ 4 levels 
323244 
"Default" , "No : 
"critical " , "good" , 
111. 
.33 33311311. 
"car " , "consumer. durable" ,.. . 
220... 

Machine generated alternative text:
# binary variables need to be converted into factor variables for univariate and bivariate ana7ysis: 
as. factor ngineer) 
Engi neer = 
MBA= as. factor (MBA) 

Dimension shows it has 444 rows and 9 columns

Machine generated alternative text:
dim(car) 
[11444 9 

Variance: of the overall data

Machine generated alternative text:
> var (car) 
Gender 
Engi neer 
MBA 
work. Exp 
Sal ary 
DI stance 
li cense 
Transport 
Age 
19. 5073313 
o. 1749537 
21. 0486446 
39. 7386879 
5. 6203120 
o. 8470299 
Gender 
NA 
NA 
NA 
NA 
NA 
NA 
Engi neer 
0.174953735 
0.185645578 
o. 188828219 
o. 390795762 
o. 092163891 
o. 003457182 
work. Exp 
21. 0486446 
o. 1888282 
26. 1335438 
49. 8057511 
6. 8713649 
o. 9815956 
Sal ary 
39. 7386879 
o. 3907958 
49. 8057511 
109. 2830106 
16. 6761196 
2. 2520794 
D. stance 
5. 62031196 
o. 09216389 
6. 87136488 
16. 67611962 
13. 00431390 
o. 44353812 
1 i cense 
o. 847029874 
o. 003457182 
o. 981595591 
2. 252079393 
0.443538121 
o. 179773453 
Transport 

Summary : of the data

Machine generated alternative text:
> summary(car) 
work. Exp 
:18. oo 
:27.oo 
5.0 
:27. 75 
:16. 24 
3rd Qu 0000 
:13.43 
:43. oo 
Age 
Gender 
Femal e: 128 
Male : 316 
Engi neer 
Min. 
0000 
1st Qu 
. 0000 
Medi an . 
0000 
:o.7545 
Mean 
3rd Qu 
. 0000 
0000 
MBA 
.•o. 
0000 
. .•o. 0000 
: o. 0000 
.•o. 2528 
. .•1. 0000 
. 0000 
Min. 
1st Qu.: 3 
Median . 
Mean 
3rd Qu.: 8 
Max. 
Sal ary 
Min. 
1st Qu. : 
Medi an 
Mean 
3rd Qu 
Max. 
Di stance 
1 i cense 
Tr ansport 
Min. 
1st Qu 
Medi an 
Mean 
3rd Qu 
Min. 
1st Qu 
Medi an 
Mean 
3rd Qu 
Max. 
O 
50 
80 
. 60 
. 00 
Min. 
1st Qu. : 
Medi an 
Mean 
3rd Qu. 
Max. 
: 11 
: 11 
20 
8. 80 
. 00 
. 32 
40 
Min. 
1st Qu 
Medi an 
Mean 
Max. 
.•o. 
.•o. 
0000 
0000 
. :25. 00 
. :30. oo 
2whee1 er 
car 
public Transport: 300 
: 6.3 
: 24 
g. 
.:15.72 
: 57 
0000 
.•o. 2342 

**6. Univariate analysis**

Histogram for numerical variables and box plot for categorical variables

Histogram: for numeric variables

Machine generated alternative text:
Age 
Salary 
Work. Exp 
MBA 

Boxplot for categorical variables:

Machine generated alternative text:
Male 
2Wheeler 
MBA 
Car 
Engineer 
Female 
Gender 
Public Transpon 

**7. Bivariate Analysis: outliers seems to be work ex, salary, age and distance**

**Can do box plot for categorical variable vs numeric variable:**

Machine generated alternative text:
Gender 
Engineer 
MBA 
Work. Exp 
Salary 
Distance 
license 
Transport 
Female 
Two wheeler 
Public Transport 
Carl 

Analysis using qplot

Machine generated alternative text:
Carl 
Carl 
15 
075 
0 75 
Salary 
Salary 
Carl 
15 
075 

Machine generated alternative text:
075 
OSO 
Salary 
075 
OSO 
2Wheeler 
075 
OSO 
075 
OSO 
Public Transpon 
075 
MBA 
Car 
Transport 
075 
075 
OSO 
Female 
Gender 
075 
OSO 
Distance 
075 
OSO 
Male 
075 
OSO 
Age 
075 
Engineer 

Inferences:

1. More age, more work experience.
2. Car users are more between age 30-40.
3. If distance is more than 10 km then car users will be more
4. If salary is more than 30k then chances of using car is more

If average daytime minutes per month is high the customer will cancel the service.

Bivariate analysis for numeric variables can be done using correlation plot

Machine generated alternative text:
Engineer 
MBA 
Work. Exp 
Salary 
Distance 
license 
0.8 
0.6 
0.2 
-0.2 
-0.6 
-0.8 

As per graph :

**1. "MonthlyCharge" and "DayMins" seems to be correlated**

**2."MonthlyCharge" and "Datausage" seems to be highly correlated**

**8. outlier detection**

Machine generated alternative text:
Gender 
Engineer 
MBA 
Work. Exp 
Salary 
Distance 
license 
Transport 
Female 
Two wheeler 
Public Transport 
Carl 

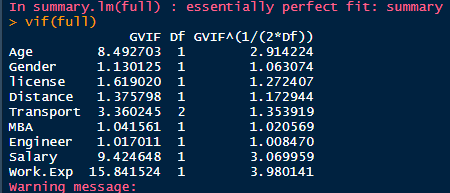
**9. Missing values and treatment**

There is one missing values, filled it with median

Machine generated alternative text:
> 
> 
> 
sapply(car , function(x) sum(is. na(x))) 
Gender Eng i nee r 
MBA work 
Age 
. Exp 
#checking unique val ue 
sapply(car , function(x) length(unique(x))) 
Gender Eng i nee r 
MBA work. 
Age 
Exp 
Sal ary 
Sal ary 
122 
Dlstance 
Dlstance 
137 
1 i cense 
1 i cense 
Tr ansport 
Tr ansport 
carl 
carl 

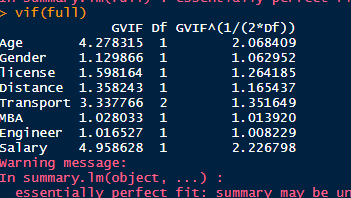
Machine generated alternative text:
> 
> 
> 
> 
> 
> 
> 
> 
car Scarl<- i fel se (carSTransport= 
"car" 
#checking null value 
sapply(car , function(x) sum(is. na(x))) 
Gender Eng i nee r 
MBA work. Exp 
Age 
#treating missing values 
Sal ary 
#the missing values is filled using the mean of the total 
license Transport 
Dlstance 
values under the columns "MBA" 
car SMSA [i s . na (car SMSA) SMSA, na. rm=TRUE) 
#checking null value 
sapply(car , function(x) sum(is. na(x))) 
Gender Eng i nee r 
MBA work. Exp 
Age 
Sal ary 
Dlstance 
license Transport 
carl 
carl 

1. **Check for multicollinearity & treat it**



Checked that age, work.exp and salary are multicollinear

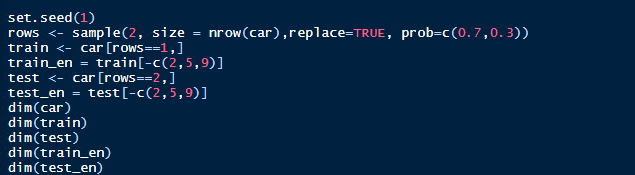
After removing work.experience we don’t see multicollinearity



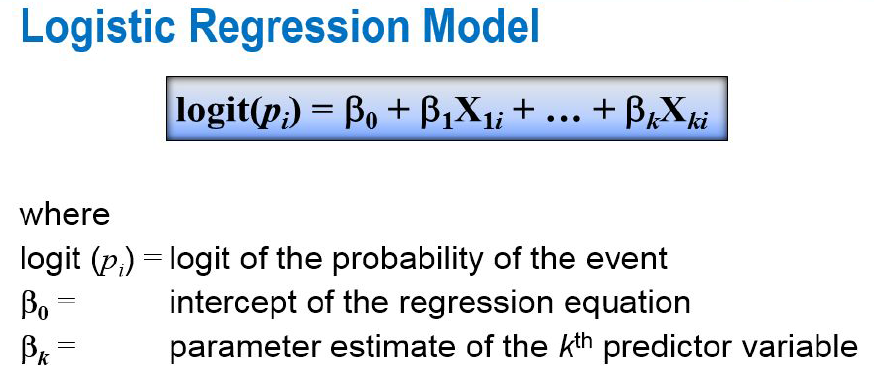
1. **Summarize the insights you get from EDA**
2. **Added dummy variable:**



1. **Split into training and test data ( removing column gender, and transport**)



* **12. Logistic Regression**

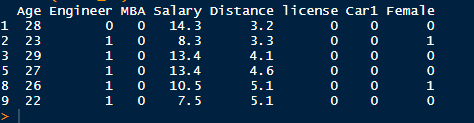


What logistic regression predicts

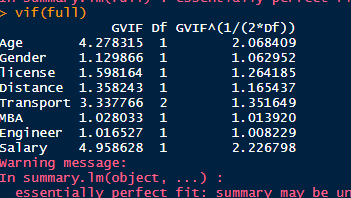
The variate or value produced by logistic regression is a probability value

between 0.0 and 1.0.

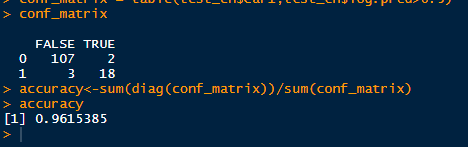
removed column Work experience to remove multicollinearity.



No collinearity between significant data:



* Confusion matrix:

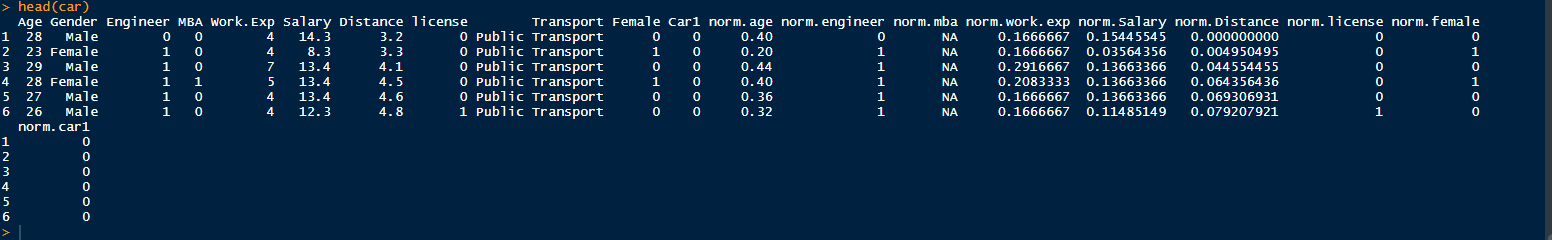


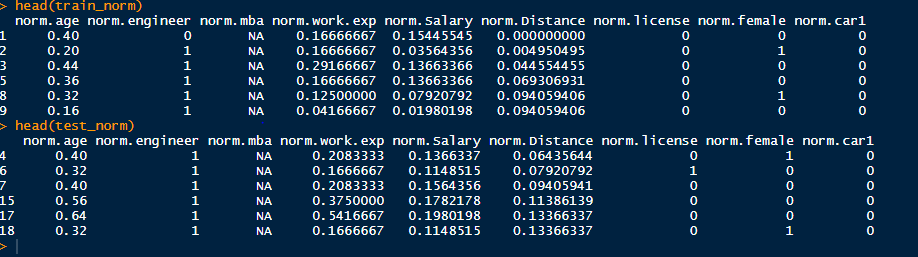
The accuracy is showing up 96.1 percent here

* **12. Interpretation of Logistic Regression**
* **1. Variables work exper was not that significant hence removed**
* **2. The accuracy of model is 96.1 per which is good**

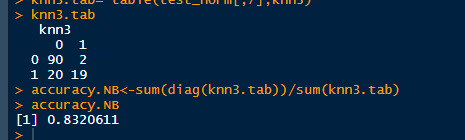
* **13. KNN model**
* **What is kNN Algorithm?**
* Let’s assume we have several groups of labeled samples. The items present in the groups are homogeneous in nature. Now, suppose we have an unlabeled example which needs to be classified into one of the several labeled groups. How do you do that? Unhesitatingly, using kNN Algorithm.
* k nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. This algorithms segregates unlabeled data points into well defined groups.
* **Pros:** The algorithm is highly unbiased in nature and makes no prior assumption of the underlying data. Being simple and effective in nature, it is easy to implement and has gained good popularity.
* **Cons:** Indeed it is simple but kNN algorithm has drawn a lot of flake for being extremely simple! If we take a deeper look, this doesn’t create a model since there’s no abstraction process involved. Yes, the training process is really fast as the data is stored verbatim (hence lazy learner) but the prediction time is pretty high with useful insights missing at times. Therefore, building this algorithm requires time to be invested in data preparation (especially treating the missing data and categorical features) to obtain a robust model.

Normalized data





**Output of knn model for K=3**



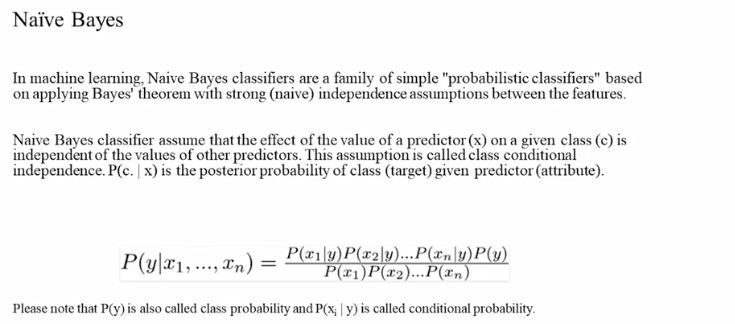
* 1. **Interpretation of KNN model** 
     1. **k=3 is best as ( loss function is also very imp along with accuracy and we see there is not much difference in accuracy when k=5 but loss is increased.**

**Also for higher k values the accuracy is dropping and loss is increasing hence we will chose k =3**

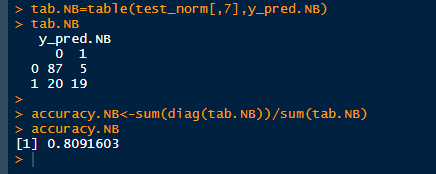
* + 1. **Normalized data is must for knn model as distance is the measure for this model**

**The accuracy is 83.2 percent which is not very good good**

* 1. **Analysis of Naive Bayes**



Naïve based: for categorical x, but continuous variable has been accommodated into naïve based.





**16. Interpretation of Naive Bayes**

is it applicable here? comment and if it is not applicable, how can you build an NB model in this case?)

* + 1. **It is initially based on categorical variable, when both x and y are categorical but here we have continuous variable**

**Hence we have a version of Naïve bayes algorithm where it uses some methods where x need not to be discrete.**

**There are two primary ways to incorporate continuous features into the Naive Bayes model:**

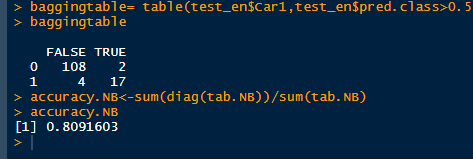
1. **Discretization**: One can transform continuous features into discrete features by categorizing different values into discrete buckets. For instance, a continuous feature can be binarized by treating all values that exceed a threshold as "Large" and all values the don't "Small". Over course, more fine grained discretization that categorizes values into any arbitrary number of buckets is possible as well.
2. **Distribution modeling**: To incorporate continuous features into the Naive Bayes model without discretization, one can replace the conditional probability of that feature given each class label in the Naive Bayes
   * 1. **It also performs the analysis on normalized data.**
     2. **The accuracy is 80.91 percent which is not good and worst among naïve baiyes, knn and logistic**
   1. **Confusion matrix interpretation for all models**

**The model which performed best is logistic with 96.1 percent accuracy ( among logistic, knn and naïve baiyes)**

**Confusion matrix is shown above**

**17. Bagging**

* It is a way to decrease the variance of your prediction by generating additional data for training from your original dataset using combinations with repetitions to produce multisets of the same cardinality/size as your original data.

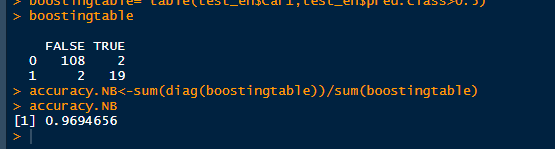


This is model performance is equivalent to naïve baiyes and not very good model

Bagging can help us only so much when we are using a data set that is such imbalanced.

**18. Boosting :** The idea of training the weak learners sequentially. Three types of boosting are:

* AdaBoosting (Adaptive Boosting) – Building on weak learners combining decision stumps and weighting incorrect observations.
* Gradient Boosting – builds on each model, trying to fit the next model to the residuals of the previous model.
* XGBoost (Extreme Gradient Boosting) – a specialized implementation of gradient boosting decision trees designed for performance. Three main types are: gradient boosting, stochastic gradient boosting and regularized gradient boosting



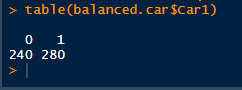
**The result of accuracy is 96.9 percent which is best among all the models so far. Also it reduces variance and bias both.**

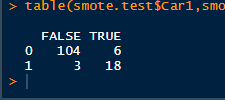
**19.SMOTE.**

**Synthetic Minority Over-sampling technique :**

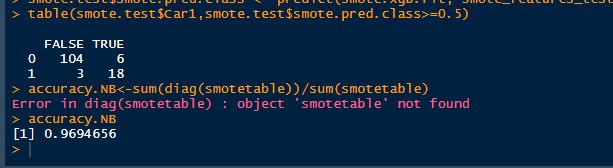
**A methodology to handle class imbalance problems.**

* **Synthesis of new minority class instances**
* **Over – sampling of minority class**
* **Under – sampling of majority class**
* **Tweaking the cost function to make misclassification of minority instances more important than misclassification of majority instances.**





**Accuracy is 96.94 percent**



**19. best model**

**Boosting and smote are best among all as it works similarly in test and training data and reduces bias and variance both and accuracy is also 96.9 percent which is fairly good.**

**20. Actionable Insights and Recommendations**

* + 1. Need to normalize data for knn
    2. Need to figure out multicollinearity.