PREDICTION OF TRANSMISSON TYPE OF CAR

Batuhan SAYLAM
Middle East Technical University
Ankara, Turkey
ID:2429264

ABSTRACT: This study analyses car data. The study makes use of R-Studio. Prior to data analysis, the data are preprocessed. Outlier analysis, data cleaning, and handling messy data are done. Also, some data manipulation techniques applied to prepare the data to prediction. Data manipulation is followed by the exploratory data analysis. Also, Logistic regression, Support Vector Machine, Naïve Bayes, Decision Tree, XGBoost and Neural Network are applied to predict transmission type.

1. INTRODUCTION

I have worked with the car dataset. First, I showed the first six line of my dataset and structure of it. Then, I corrected and abbreviated the column names. Then, I looked the factor columns and correct the mistakes. After that, I applied exploratory data analysis. Then, I looked the number of NA values. After that, I cleaned each column and handled with messy data. Then, I detected outliers and replace them with mean of the columns where the outliers located. After, I applied PCA to reduce dimension. Then, I applied confirmatory data analysis. Finally, I split data into train and test data and conducted prediction models and performance measure.

1.1. DATA DESCRIPTION

The data is about car sales. The data includes 13 columns. There is a name variable including names of the cars. The six of columns are categorical, and the six of them are numerical. Year is year in which car was sold, kmDriven is number of kilometers that the car is driven, seller_Type tells us whether car is sold by individual or dealer, and owner tells the number of previous owners. Source the data is https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?select=Car+details+v3.csv.

	name <chr>	year <int>	selling_price <int>	km_driven <int>	fuel <chr>	seller_type <chr>	transmission <chr>	owner <chr>
1	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner
2	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owne
3	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner
4	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner
5	Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner
6	Hyundai Xcent 1.2 VTVT E Plus	2017	440000	45000	Petrol	Individual	Manual	First Owner

mileage <chr>	engine <chr>	max_power <chr>	torque <chr>	seats <int>
23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm	5
21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm	5
17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm@ rpm)	5
23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm	5
16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500(kgm@ rpm)	5
20.14 kmpl	1197 CC	81.86 bhp	113.75nm@ 4000rpm	5

Table 1 The first six lines of the data

```
     name                year        selling_price       km_driven          fuel            seller_type        transmission
 Length:8128        Min.   :1983   Min.   :   29999   Min.   :      1   Length:8128        Length:8128        Length:8128
 Class :character   1st Qu.:2011   1st Qu.:  254999   1st Qu.:  35000   Class :character   Class :character   Class :character
 Mode  :character   Median :2015   Median :  450000   Median :  60000   Mode  :character   Mode  :character   Mode  :character
                    Mean   :2014   Mean   :  638272   Mean   :  69820
                    3rd Qu.:2017   3rd Qu.:  675000   3rd Qu.:  98000
                    Max.   :2020   Max.   :10000000   Max.   :2360457

    owner              mileage             engine            max_power            torque              seats
 Length:8128        Length:8128        Length:8128        Length:8128        Length:8128        Min.   : 2.000
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.: 5.000
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 5.000
                                                                                                Mean   : 5.417
                                                                                                3rd Qu.: 5.000
                                                                                                Max.   :14.000
                                                                                                NA's   :221
```

Table 2 The summary of the data

```
'data.frame':   8128 obs. of  13 variables:
 $ name         : chr  "Maruti Swift Dzire VDI" "Skoda Rapid 1.5 TDI Ambition" "Honda City 2017-2020 EXi" "Hyundai i20 Sportz Diesel"
...
 $ year         : int  2014 2014 2006 2010 2007 2017 2007 2001 2011 2013 ...
 $ selling_price: int  450000 370000 158000 225000 130000 440000 96000 45000 350000 200000 ...
 $ km_driven    : int  145500 120000 140000 127000 120000 45000 175000 5000 90000 169000 ...
 $ fuel         : chr  "Diesel" "Diesel" "Petrol" "Diesel" ...
 $ seller_type  : chr  "Individual" "Individual" "Individual" "Individual" ...
 $ transmission : chr  "Manual" "Manual" "Manual" "Manual" ...
 $ owner        : chr  "First Owner" "Second Owner" "Third Owner" "First Owner" ...
 $ mileage      : chr  "23.4 kmpl" "21.14 kmpl" "17.7 kmpl" "23.0 kmpl" ...
 $ engine       : chr  "1248 CC" "1498 CC" "1497 CC" "1396 CC" ...
 $ max_power    : chr  "74 bhp" "103.52 bhp" "78 bhp" "90 bhp" ...
 $ torque       : chr  "190Nm@ 2000rpm" "250Nm@ 1500-2500rpm" "12.7@ 2,700(kgm@ rpm)" "22.4 kgm at 1750-2750rpm" ...
 $ seats        : int  5 5 5 5 5 5 5 4 5 5 ...
```

Table 3 The structure of the data

1.2 RESEARCH QUESTIONS

1) What are the distributions of the numeric variables?

2) Is the mileage distributed normally?

3) In different sale types, does transmission type have an impact on sales prices?

4) Is there a statistically significant correlation among numeric variables?

5) How is transmission type distributed according to sell type?

6) Does the engine affect max power in transmission type?

1.3 AIM OF STUDY

This project analyses the car dataset using statistical applications such as exploratory data analysis and confirmatory data analysis, and models which predict the transmission type of cars.

2. METHODOLOGY/ANALYSIS

First, data preprocessing is conducted to remove duplicate values and unused columns from the data. Then, the exploratory data analysis is conducted to answer the research questions by using visual techniques. After that, I handled with missing values by using predictive mean matching method (PMM) from mice package. Then, for data manipulation, Box-Cox transformation is used to normalize data, for numerical variables, PCA is applied to reduce dimension of data, and for categorical variables, dummy variables are created by using fastdummies package. Then, the confirmatory data analysis is conducted to find out exact answers of the research questions by using hypothesis testing such as Pearson's Correlation Test, Spearman Correlation Test, Scheirer-Ray-Hare Test, Chi-Squared Test, and Kolmogorov-Smirnov test. Then, the dataset is splitted into two part which are train and test by validation set approach and by using createDataPartition function whose p is 0.8 from caret package. Finally, the models, such as Logistic regression, Support Vector Machine, Naïve Bayes, Decision Tree, XGBoost and Neural Network, are trained and compared to find out which model shows best performance.

3.RESULT AND FINDINGS

3.1 DATA PREPROCESSING

Firstly, duplicated data is checked. 1230 rows of data are duplicated value hence duplicated values are deleted. Then, the torque and seats are removed from data since there is no need the column. Then, column names are checked, and some column names are replaced with their different format. In mileage,engine and maxPower columns, there are their units; hence,the units are removed by using str_remove from stringr. Also, categorical columns are checked to find a mistake by using table function, but any mistake was not found. Then, categorical columns are converted to factor. Also, numerical columns are got from data as a new data frame to compute the standard deviations of them.

```
   sellPrice     kmDriven       mileage       engine     maxPower
519766.98599  58358.09518       4.04915    493.49328     31.77162
```

After the changes:

```
'data.frame':   6926 obs. of  11 variables:
 $ name        : chr  "Maruti Swift Dzire VDI" "Skoda Rapid 1.5 TDI Ambition" "Honda City 2017-2020 EXi" "Hyundai i20 Sportz Diesel"
 ...
 $ year        : Factor w/ 29 levels "1983","1991",..: 23 23 15 19 16 26 16 10 20 22 ...
 $ sellPrice   : int  450000 370000 158000 225000 130000 440000 96000 45000 350000 200000 ...
 $ kmDriven    : int  145500 120000 140000 127000 120000 45000 175000 5000 90000 169000 ...
 $ fuel        : Factor w/ 4 levels "CNG","Diesel",..: 2 2 4 2 4 4 3 4 2 2 ...
 $ sellType    : Factor w/ 3 levels "Dealer","Individual",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ transmission: Factor w/ 2 levels "Automatic","Manual": 2 2 2 2 2 2 2 2 2 2 ...
 $ owner       : Factor w/ 5 levels "First Owner",..: 1 3 5 1 1 1 1 3 1 1 ...
 $ mileage     : num  23.4 21.1 17.7 23 16.1 ...
 $ engine      : num  1248 1498 1497 1396 1298 ...
 $ maxPower    : num  74 103.5 78 90 88.2 ...
```

Table 4 The structure of the data

```
     name                year         sellPrice          kmDriven            fuel                   sellType      transmission
 Length:6926        2017   : 808   Min.   :   29999   Min.   :      1   CNG   :  56   Dealer            : 681   Automatic: 584
 Class :character   2016   : 693   1st Qu.:  250000   1st Qu.:  40000   Diesel:3755   Individual        :6218   Manual   :6342
 Mode  :character   2015   : 681   Median :  400000   Median :  70000   LPG   :  38   Trustmark Dealer:  27
                    2018   : 608   Mean   :  517271   Mean   :  73996   Petrol:3077
                    2012   : 604   3rd Qu.:  633500   3rd Qu.: 100000
                    2014   : 581   Max.   :10000000   Max.   :2360457
                    (Other):2951
                owner          mileage          engine        maxPower
 First Owner         :4242   Min.   : 0.00   Min.   : 624   Min.   :  0.00
 Fourth & Above Owner: 169   1st Qu.:16.80   1st Qu.:1197   1st Qu.: 67.10
 Second Owner        :1974   Median :19.44   Median :1248   Median : 81.83
 Test Drive Car      :   5   Mean   :19.47   Mean   :1431   Mean   : 87.73
 Third Owner         : 536   3rd Qu.:22.50   3rd Qu.:1498   3rd Qu.:100.00
                             Max.   :42.00   Max.   :3604   Max.   :400.00
                             NA's   :208     NA's   :208    NA's   :206
```

Table 5 The summary of the data

Finally, outliers of numeric column are replaced with their mean values except engine and seats column since the percentage of their outliers is greater than ten percent of their size.

3.2 EXPLORATORY DATA ANALYSIS

1) What are the distributions of the numeric variables?

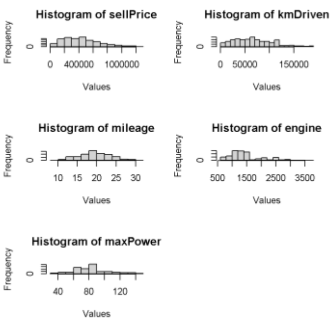


Figure 1 the histograms of numeric variables

These plots show us the mileage may be distributed approximately normal. Also, price and engine columns look like right skewed.

2) Is the mileage distributed normally?

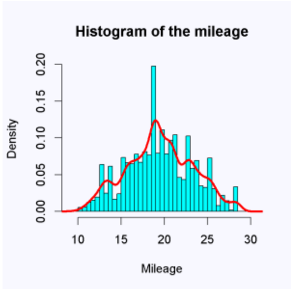


Figure 2 Histogram and density plot of mileage

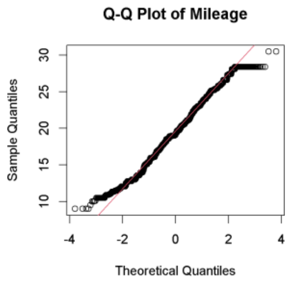


Figure 3 Q-Q plot of mileage

According to Histogram and density plot, the distribution of it is not normal. Also, when we checked the Q-Q Plot of it, there exists significant deviations from the line. Hence, according to visual evidence, the mileage does not normally distributed.

3) In different sale types, does transmission type have an impact on sales prices?

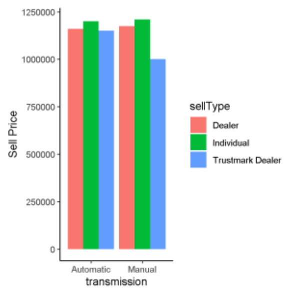


Figure 4 Bar plot of sell price according to sell type and transmission type

This plot suggests that there may be a significant difference between sell prices for each sell type depending on the transmission type. However, for each situation, the prices of cars which are sold by individual is highest and the prices of cars which are sold by Trustmark dealer are lowest. Also, the prices of cars with manual type increase for each sell type except Trustmark Dealer.

4) Is there a statistically significant correlation among numeric variables?

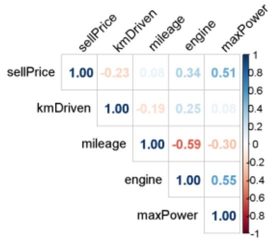


Figure 5 Correlation plot of numeric variables

According to the plot, there exists a high positive correlation between max power and sell prices. Also, there are high positive correlations between engine& max power and engine& seats. However, there are high negative correlations between engine& mileage and mileage& seats. Also, there is no multicollinearity problem.

5) How is transmission type distributed according to sell type?

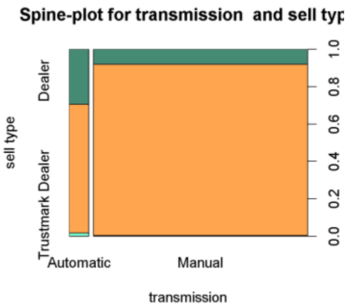


Figure 6 Spine plot of transmission type and sell type

According to plot, for each transmission type, the highest percentage is Trustmark dealer and the lowest is dealer. However, there is no individual seller for manual transmission type.

6) Does the engine affect max power in transmission type?

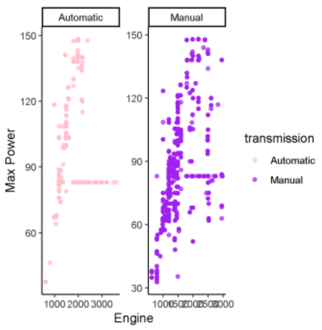
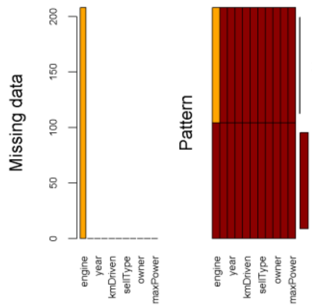


Figure 7 Scatter plot of engine and max power according to transmission type

For each plot, there is a positive correlation between engine and max power. Hence, we can say engine affects max power for each transmission type.

3.3 HANDLING MISSING VALUES

I handled with them by using mice package.



Variable	Count
engine	208
name	0
year	0
sellPrice	0
kmDriven	0
fuel	0
sellType	0
transmission	0
owner	0
mileage	0
maxPower	0

Figure 7 Missing data plot *Table 6 Missing data table*

According to the outputs, there exist 208 NA values in data. The values are in engine column. NA values are replaced with new values by using predictive mean matching method (PMM).

Then, the distribution of imputed engine column is checked to find out that the distribution do not change after imputation.

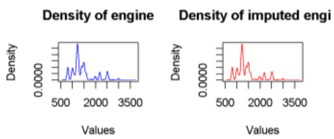
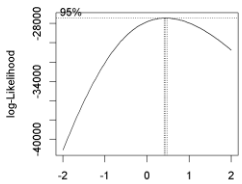


Figure 8 Density plots of engine and imputed engine

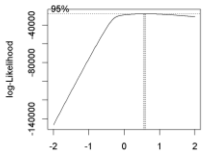
According to density plot, the distribution of it does not change.

3.4 DATA MANIPULATION

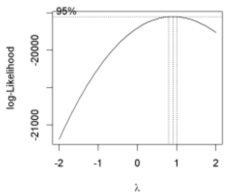
Firstly, I used Box-Cox transformation to normalize the numerical columns.



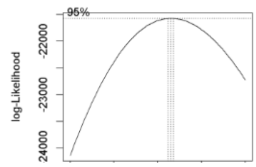
For sellPrice columns, since the interval does not include 1, I transformed sellPrice column.



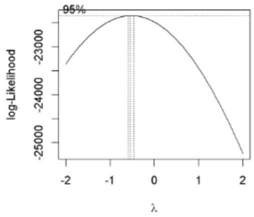
For kmDriven column, since the interval does not include 1, I transformed kmDriven column.



For mileage column, since the interval does not include 1, I transformed mileage column.



For max power column, since the interval does not include 1, I transformed max power column.



For engine column, since the interval does not include 1, I transformed engine column.

After that, in order to reduce the dimensions of the data, I used principal component analysis (PCA).

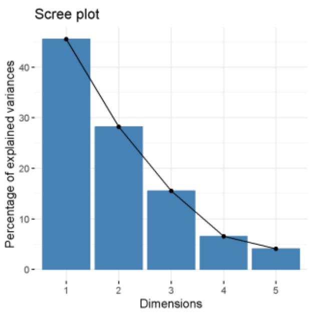


Figure 9 PCA plot

According to the plot, I chose first two principal components.

Finally, I used fastDummies package and dummy_cols function for one hot encoding of the categorical columns.

3.5 CONFIRMATORY DATA ANALYSIS

2) Is the mileage distributed normally?

Since length of the data is greater than 5000, we cannot apply Shapiro-Wilk test; hence, we will apply Kolmogorov-Smirnov test.

Null Hypothesis (H0): Sample data comes from the normal distribution.

Alternative Hypothesis (H1): Sample data does not come from the normal distribution.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  dataComp$mileage
D = 0.039173, p-value = 1.174e-09
alternative hypothesis: two-sided
```

Since p value is less than 0.05, reject null hypothesis. Hence, sample data does not come from the normal distribution.

3) In different sale types, does transmission type have an impact on sales prices?

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  dataComp$sellPrice
D = 0.036286, p-value = 2.399e-08
alternative hypothesis: two-sided
```

Since p value is less than 0.05, reject null hypothesis. Hence, sample data does not come from the normal distribution.

Since normality assumption is not provided, I used Scheirer-Ray-Hare Test which is the non-parametric version of the two way anova.

```
DV:  sellPrice
Observations:  6926
D:  0.9997079
MS total:  3998834

                       Df     Sum Sq       H
transmission            1 5.6810e+08 142.135
sellType                2 4.9883e+08 124.806
transmission:sellType   2 1.3528e+08  33.846
Residuals            6920 2.6191e+10
                         p.value
transmission          0.0000e+00
sellType              0.0000e+00
transmission:sellType 4.4723e-08
Residuals
```

Hypothesis 1 (Sell Price does not differ depending on transmission type): 0 < .05

Hypothesis 2 (Sell price levels do not differ depending on sell type): 0 < .05

Hypothesis 3 (The combination of transmission type and sell type is not impacting the sell price): 4.4723e-08 < .05

Hypothesis 1: Reject!

Hypothesis 2: Reject!

Hypothesis 3: Reject!

Hence, in different sale types, transmission type has impact on sales prices according to hypothesis 3.

4) Is there a statistically significant correlation among numeric variables?

Pearson's correlation test is conducted to test correlation between each two numeric variables. All p-values except correlation test for kmDriven and maxPower are equal to 2.2*10^(-16) and p-value of correlation test for kmDriven and maxPower is equal to 4.028*10^(-10).

According to the outputs, there is no correlation between each two numeric variables since all p-values of the Pearson's correlation tests are less than 0.05.

5) How is transmission type distributed according to sell type?

Since there exist two categorical variables, we can apply Chi-Squared Test.

```
        Pearson's Chi-squared test

data:  contingencyTable
X-squared = 309.45, df = 2, p-value < 2.2e-16
```

We fail to reject the null hypothesis, and we conclude that there is not a significant association between sell type and the situation that car is sold in 2020 or before.

6) Does the engine affect max power in transmission type?

Firstly, the normalities of maxPower and engine variables are checked for different transmission types.

For automatic type, the p-values of the Kolmogorov-Smirnov test are:

For maxPower:

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  dataAutomatic$maxPower
D = 0.30547, p-value < 2.2e-16
alternative hypothesis: two-sided
```

For engine:

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  dataAutomatic$engine
D = 0.14206, p-value = 1.159e-10
alternative hypothesis: two-sided
```

For manual type, the p-values of the Kolmogorov-Smirnov test are:

For maxPower:

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  dataManual$maxPower
D = 0.089645, p-value < 2.2e-16
alternative hypothesis: two-sided
```

For engine:

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  dataManual$engine
D = 0.14784, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Since all variables are not normally distributed, I will use Spearman correlation test which is alternative of Pearson's corr. test for nonnormal variables

transmission <fctr>	correlation <dbl>	p_value <dbl>
Automatic	0.3191162	2.72665e-15
Manual	0.6740515	0.00000e+00

Since p-values are smaller than 0.05, there is no correlation between variables according to transmission type.

3.6 MODELING

The dataset is splitted into two part which are train and test by validation set approach and by using createDataPartition function whose p is 0.8 from caret package. Logistic regression model, support vector machine (SVM), Naïve Bayes, Decision Tree, XGBoost and Neural Network are used to classify the data to predict transmission type. Tuning is applied the svm model to find the best parameters.

For Logistic Regression Model:

```
Call:
glm(formula = transmission_Automatic ~ ., family = binomial,
    data = train)

Coefficients: (3 not defined because of singularities)
                              Estimate Std. Error
(Intercept)                   -1.70219    0.52015
PC1                           -0.73890    0.04450
PC2                           -0.25165    0.05161
fuel_CNG                     -13.14582  340.13534
fuel_Diesel                   -0.99996    0.11841
fuel_LPG                     -12.95013  422.86645
fuel_Petrol                         NA         NA
sellType_Dealer                0.15824    0.46295
sellType_Individual           -1.12480    0.45580
`sellType_Trustmark Dealer`         NA         NA
`owner_First Owner`            0.38677    0.27313
`owner_Fourth & Above Owner`  -0.09232    0.57510
`owner_Second Owner`           0.24599    0.27933
`owner_Test Drive Car`        17.10646 1370.80128
`owner_Third Owner`                 NA         NA
                             z value Pr(>|z|)
(Intercept)                   -3.272  0.00107 **
PC1                          -16.605  < 2e-16 ***
PC2                           -4.876 1.08e-06 ***
fuel_CNG                      -0.039  0.96917
fuel_Diesel                   -8.445  < 2e-16 ***
fuel_LPG                      -0.031  0.97557
fuel_Petrol                       NA       NA
sellType_Dealer                0.342  0.73249
sellType_Individual           -2.468  0.01360 *
`sellType_Trustmark Dealer`       NA       NA
`owner_First Owner`            1.416  0.15676
`owner_Fourth & Above Owner`  -0.161  0.87246
`owner_Second Owner`           0.881  0.37851
`owner_Test Drive Car`         0.012  0.99004
`owner_Third Owner`               NA       NA
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3218.1  on 5540  degrees of freedom
Residual deviance: 2663.9  on 5529  degrees of freedom
AIC: 2687.9

Number of Fisher Scoring iterations: 15
```

Since p values of fuel_CNG, fuel_LPG, sellType_Dealer, `owner_First Owner` , `owner_Fourth & Above Owner`, `owner_Test Drive Car`, `owner_Second Owner` are greater than 0.05 and p values of fuel_Petrol,`sellType_Trustmark Dealer`, `sellType_Trustmark Dealer`,`owner_Third Owner` are NA, these columns do not affect the model hence we need to remove them.

```
Call:
glm(formula = transmission_Automatic ~ ., family = binomial,
    data = logmodelTrain)

Coefficients:
                    Estimate Std. Error z value
(Intercept)         -1.19986    0.12326  -9.734
PC1                 -0.74878    0.04423 -16.930
PC2                 -0.29057    0.04711  -6.168
fuel_Diesel         -1.00514    0.11711  -8.583
sellType_Individual -1.33053    0.12032 -11.059
                    Pr(>|z|)
(Intercept)          < 2e-16 ***
PC1                  < 2e-16 ***
PC2                  6.9e-10 ***
fuel_Diesel          < 2e-16 ***
sellType_Individual  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3218.1  on 5540  degrees of freedom
Residual deviance: 2676.8  on 5536  degrees of freedom
AIC: 2686.8

Number of Fisher Scoring iterations: 6
```

Since p values of all variables are smaller than 0.05, these columns affect the model. The performance measure tables of train and test are, respectively:

```
         actual                    actual
predicted    0    1      predicted   0   1
        0 3562  102              0 926  31
        1 1509  368              1 345  83
```

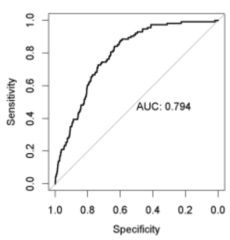


Figure 10 ROC-AUC plot of Logistic model

The ROC curve shows the model is good for the trade-off between sensitivity and specificity since the curve is close to left corner. According to AUC, degree or measure of separability is 0.794.

For SVM model:

First the model, where type is C-classification and kernel is linear, is tuned. After tuning the model, best parameters are 0.00001 for gamma, 1 for cost, and 0.01 for epsilon. The summary of the model:

```
Call:
svm(formula = transmission_Automatic ~ ., data = train,
    kernel = "linear", cost = cost, gamma = gamma,
    epsilon = epsilon, type = "C-classification")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1

Number of Support Vectors:  982

 ( 514 468 )


Number of Classes:  2

Levels:
 0 1
```

The performance measure tables of train and test are, respectively:

```
         actual                  actual
predicted    0    1 predicted    0    1
        0 5071  467         0 1270  113
        1    0    3         1    1    1
```

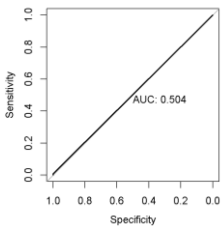


Figure 11 ROC-AUC plot of SVM model

The ROC curve shows the model is worse than logistic model for the trade-off between sensitivity and specificity since the curve is not closer to left corner. According to AUC, degree or measure of separability is 0.504.

For Naïve Bayes model:

The summary of the model:

```
          Length Class  Mode
apriori    2     table  numeric
tables    14     -none- list
levels     2     -none- character
isnumeric 14     -none- logical
call       4     -none- call
```

The performance measure tables of train and test are, respectively:

```
         actual                 actual
predicted    0    1 predicted    0    1
        0 2159   69         0 594   15
        1 2912  401         1 677   99
```

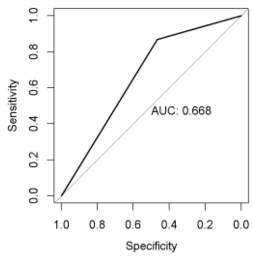
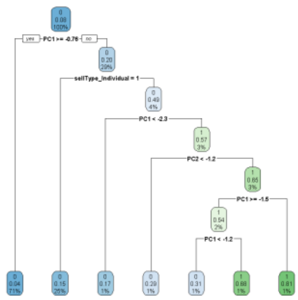


Figure 12 ROC-AUC plot of Naïve Bayes model

The ROC curve shows the model is worse than logistic model for the trade-off between sensitivity and specificity since the curve is not closer to left corner. According to AUC, degree or measure of separability is 0.668.

For Decision Tree:

The plot of the decision tree:



The performance measure tables of train and test are, respectively:

```
         actual                 actual
predicted    0    1 predicted    0    1
        0 5043  386         0 1264  101
        1   28   84         1    7   13
```

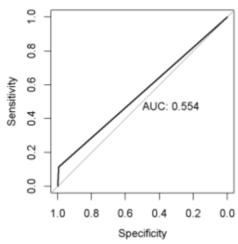


Figure 13 ROC-AUC plot of Decision Tree model

The ROC curve shows the model is worse than logistic model for the trade-off between sensitivity and specificity since the curve is not closer to left corner. According to AUC, degree or measure of separability is 0.554.

For XGBoost model:

The summary of the model:

```
                Length Class              Mode
handle              1  xgb.Booster.handle externalptr
raw             15742  -none-             raw
niter               1  -none-             numeric
evaluation_log      2  data.table         list
call               15  -none-             call
params              3  -none-             list
callbacks           1  -none-             list
feature_names      14  -none-             character
nfeatures           1  -none-             numeric
```

The performance measure tables of train and test are, respectively:

```
         actual                  actual
predicted    0    1    predicted   0   1
        0 3152   46            0 801  20
        1 1919  424            1 470  94
```

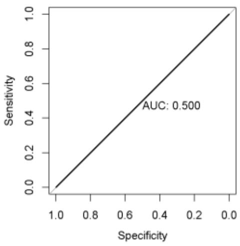


Figure 14 ROC-AUC plot of XGBoost model

The ROC curve shows the model is worse than logistic model for the trade-off between sensitivity and specificity since the curve is not closer to left corner. According to AUC, degree or measure of separability is 0.500.

For Neural Network model:

The summary of the model:

```
                    Length Class      Mode
call                    5  -none-     call
response             5541  -none-     numeric
covariate           77574  -none-     numeric
model.list              2  -none-     list
err.fct                 1  -none-     function
act.fct                 1  -none-     function
linear.output           1  -none-     logical
data                   15  data.frame list
exclude                 0  -none-     NULL
net.result              1  -none-     list
weights                 1  -none-     list
generalized.weights     1  -none-     list
startweights            1  -none-     list
result.matrix          20  -none-     numeric
```

The performance measure tables of train and test are, respectively:

```
         actual                 actual
predicted    0    1  predicted   0   1
        0 3347   84          0 872  26
        1 1724  386          1 399  88
```

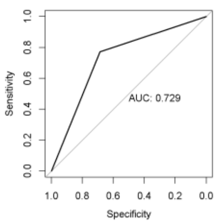


Figure 15 ROC-AUC plot of Neural Network model

The ROC curve shows the model is worse than logistic model for the trade-off between sensitivity and specificity since the curve is not closer to left corner. According to AUC, degree or measure of separability is 0.729.

The performance measure metric table for the model is:

	accuracy <dbl>	sensitivity <dbl>	specificity <dbl>
LogModel Test	0.7285199	0.9676071	0.1939252
SVM Test	0.9176895	0.9182936	0.5000000
LogModel Train	0.7092583	0.9721616	0.1960575
SVM Train	0.9157192	0.9156735	1.0000000
NBClassifier Test	0.5003610	0.9753695	0.1275773
NBClassifier Train	0.4620105	0.9690305	0.1210383
DecisionTree Test	0.9220217	0.9260073	0.6500000
DecisionTree Train	0.9252842	0.9289003	0.7500000
XGBoost Test	0.6462094	0.9756395	0.1666667
XGBoost Train	0.6453709	0.9856160	0.1809646

Performance measure metric table shows us that the best model is decision tree model for both train and test data by accuracy but to predict positive values, the best model is XGBoost since the sensitivity, which is ratio of true positive value to the actual positive values, is highest for both train and test data, and SVM has best performance for train data to predict negative values since it has highest specificity of train data, which is ratio of true negatives to the actual negative values, and Decision tree has best performance for test to predict negative values since it has highest specificity of test data.

4.CONCLUSION

As a result, according to performance measure metric table, the best model is decision tree model for both train and test data by accuracy. However, in order to predict positive values, the best model is XGBoost since the sensitivity, which is ratio of true positive value to the actual positive values, is highest for both train and test data. Also, SVM has best performance for train data to predict negative values since it has highest specificity of train data, which is ratio of true negatives to the actual negative values, and Decision tree has best performance for test to predict negative values since it has highest specificity of test data.

5. REFERENCES

[1] B. KOCA, « Logistic Regression,» *Stat412 Recitation 7*, https://rpubs.com/BurcuKG/1180677

[2] B. KOCA, «Support Vector Machines,» *Stat412 Recitation 11*, https://rpubs.com/BurcuKG/1185743

[3] KOCA, «Naïve Bayes Classifier, Decision Tree,» *Stat412 Recitation 12*, https://rpubs.com/BurcuKG/1187989

[4] B. KOCA, «XGBoost,» *Stat412 Recitation 13*, https://rpubs.com/BurcuKG/1190267

[5] Z. BOBBITT, «Kolmogorov-Smirnov Test in R (With Examples),» https://www.statology.org/kolmogorov-smirnov-test-r/

[6] G. Barabás, «13.2 The Scheirer–Ray–Hare test,» *13 Two-way ANOVA and the Scheirer–Ray–Hare test*, https://dysordys.github.io/data-with-R/Anova_two_way.html

[7] amiyaranjanrout, *Spearman Correlation Testing in R Programming*, https://www.geeksforgeeks.org/spearman-correlation-testing-in-r-programming/