Re-Purchase Prediction Model

Name: Wenying Wu
Date: 18/04/2020

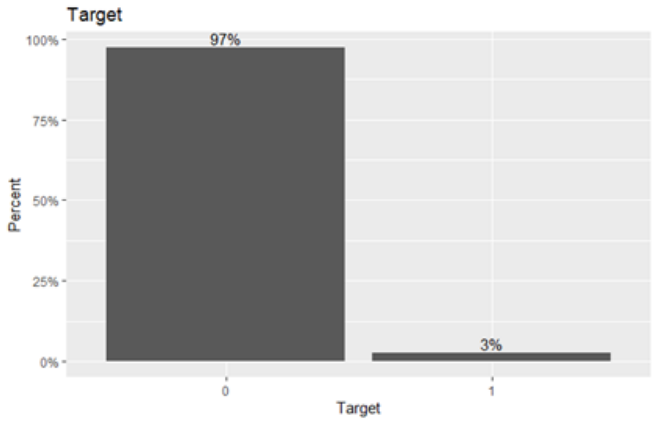
INTRO

This report is aiming at providing suggestions on targeting existing customers for a re-purchase campaign for an automotive manufacturer. Exploratory Data Analysis (EDA) is performed on data set, both linear model and tree-based model are trained for selection and customers that are likely to re-purchase are predicted based on the random forest model, which performs the best under this circumstance.

EXPLORATORY DATA ANALYSIS

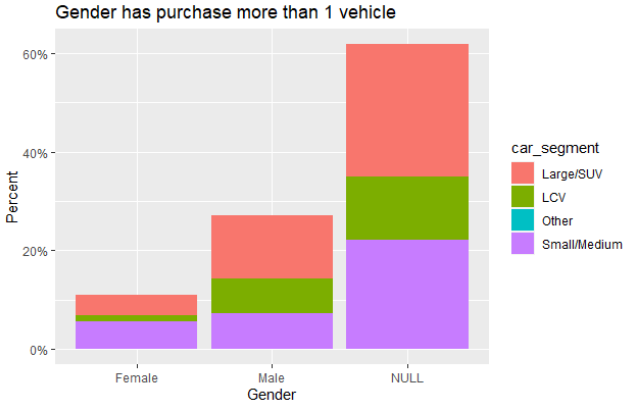
The original data 'repurchase_training.csv' contains 131,337 observations of 17 variables, including ID, Target, age_band etc. Key findings are listed below:

There are 127,816 (97%) customers who only purchased 1 vehicle, which means this data set is unbalanced regarding Target.

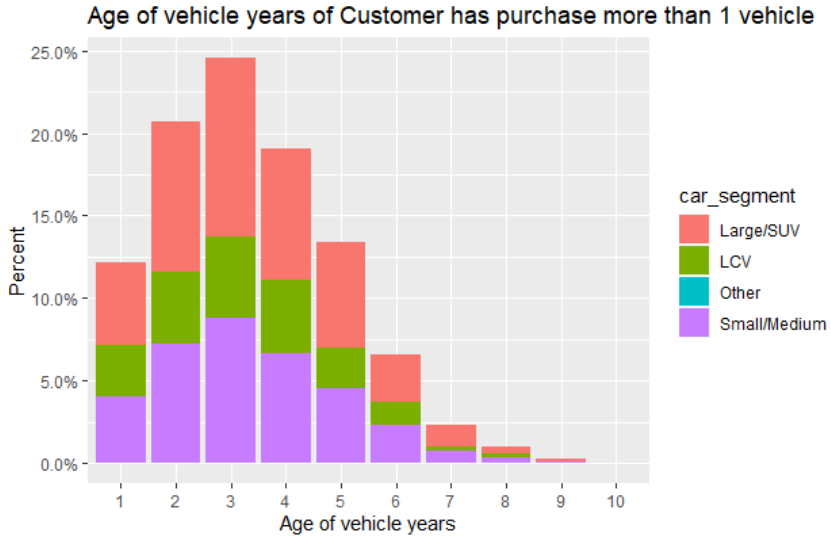


There are 112,375 (86%) customers selected NULL for age_band and 69,308 (53%) customers selected NULL for gender. This could be an indication of customers' data privacy awareness or these customers are potentially introverts and prefer not to be annoyed by others.

Customers with NULL gender takes up about 60% of customers who purchased more than 1 vehicle.



Customers with Age_of_Vehicle = 2, 3 and 4 make up about 60% of customers who purchased more than 1 vehicle.



ID is a unique value not relating to Target, to be deleted in models.

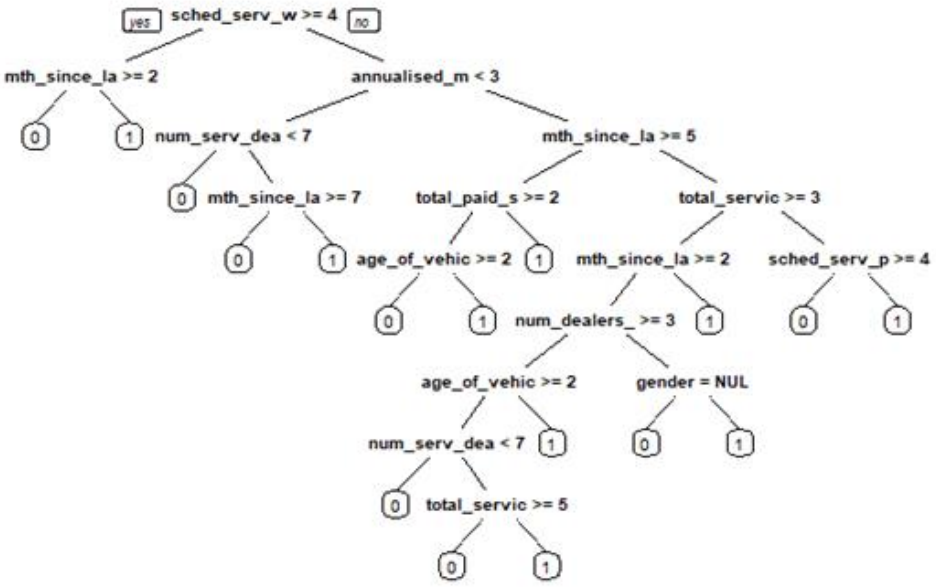
Considering the abovementioned and this is targeting for a re-purchase campaign, car manufacturer would prefer to include more customer to not miss out on a potential re-purchasing customer. The possibility threshold in all models in this report is set to 0.4 instead of 0.5 default threshold to include more customers in this campaign while not annoying customers with lower repurchasing possibilities. The possibility threshold will be illustrated in a later section. Other EDA graphs are shown in Appendix 1.

CLASSIFICATION MODELS

Classification models are appropriate in this case because they can draw a conclusion from the given data ('repurchase_training.csv') and predict the category (Target in 'repurchase_validation.csv'). Predicted '1' in Target means the customer is predicted to be likely to re-purchase a vehicle, while '0' means the opposite. Classification models output prediction with default possibility threshold = 0.5, which means a customer with predicted probability < 0.5 indicates he is not a re-purchaser, while a customer with predicted probability > 0.5 indicates he is a re-purchaser. However, given the abovementioned, all possibility thresholds are set to 0.4 in this report.

There are four models trained for comparison, namely logistic model, LASSO regression model, decision tree model and random forest model. Logistic and LASSO model are linear models, logistic model considers all provided variables while the LASSO model eliminates the effect of some variables that are not significantly related to Target. The decision tree model and random forest are tree-based models, decision tree model predict each observation based on the most commonly occurring class of training observations in its region while random forest can be considered as producing multiple decision trees then combine to yield a single consensus prediction (James et al., 2013).

Below is the decision tree from the decision tree model in this case.



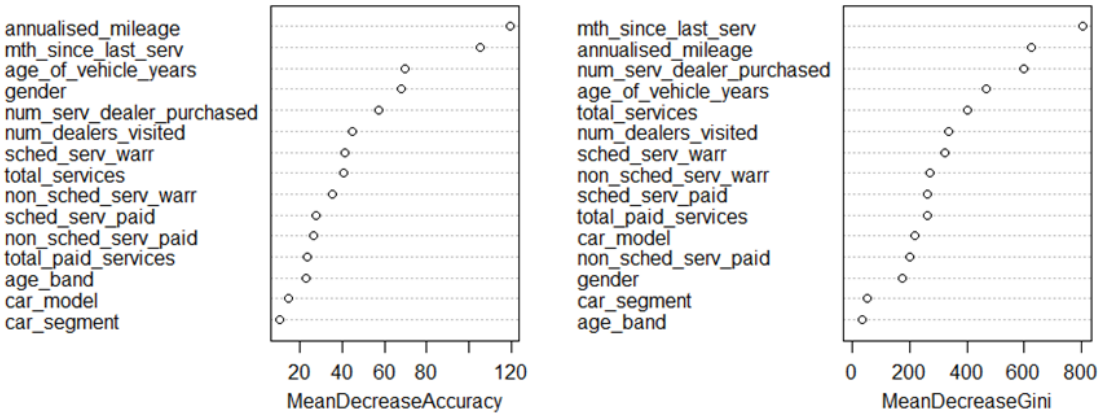
The summarised result, including the confusion matrix and ROC curve of each model are in Appendix 2.

Variable Importance:

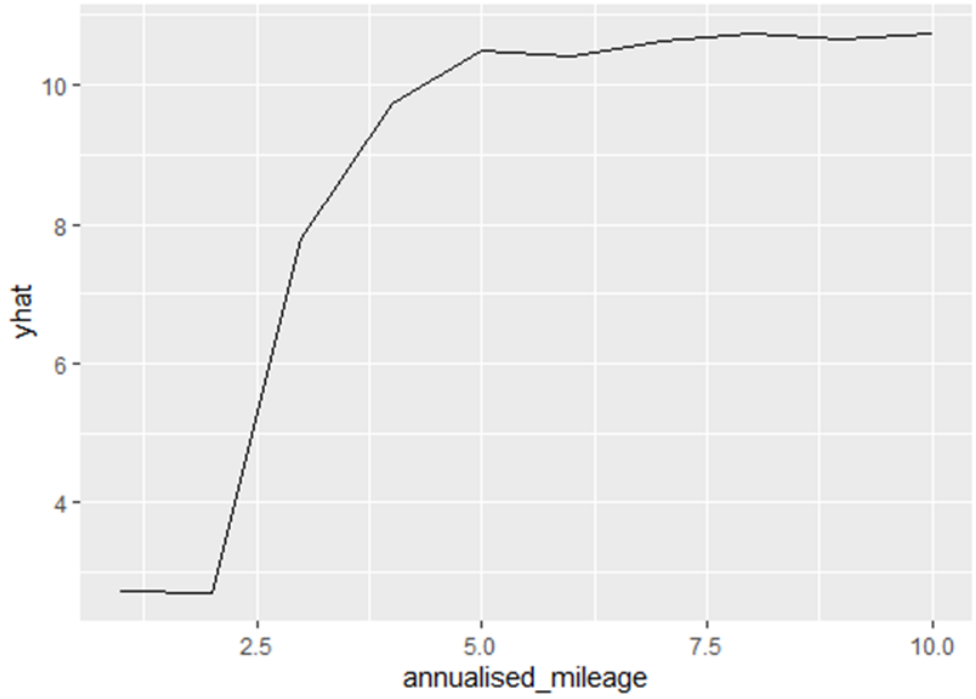
Variable importance in the linear model is simply the coefficient of each variable, a larger coefficient means a larger increase in the log odds (possibility). For example, the below screenshot is part of the coefficient of the logistic model, the 2.04855 coefficient of age_band 7.75+ is the largest among all age_bands, which means customers in age_band 7.75+ is the customers with the largest possibility to re-purchase a car given all other variables are the same.

```
Coefficients: (2 not defined because of singularities)
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.48256    0.61694  -7.266 3.71e-13 ***
age_band2. 25 to 34       0.95156    0.65389   1.455  0.14561
age_band3. 35 to 44       1.09099    0.64525   1.691  0.09087 .
age_band4. 45 to 54       1.71863    0.62968   2.729  0.00635 **
age_band5. 55 to 64       1.78579    0.63114   2.829  0.00466 **
age_band6. 65 to 74       1.57819    0.66020   2.390  0.01683 *
age_band7. 75+            2.04855    0.69568   2.945  0.00323 **
```

Meanwhile, variable importance measures for the random forest model are named Mean Decrease Accuracy (MDA) and Mean Decrease Ginni (MDG), meaning how much accuracy the model losses without that variable and how much that variable contributes to the homogeneity of the resulting random forest. In other words, a variable having higher MDA and MDG means it is of higher importance.



Partial dependency plots provide a graphic presentation of random forest model coefficients, Appendix 3 shows Partial dependency plots of the top 5 most important features. Below is an example of annualised _mileage:



This plot indicates that the possibility of re-purchasing depends on annualised _mileage a lot from 2 to 5, but not depending on annualised _mileage in the range 0 to 2 and 6 to 10 given that the partial dependency plot is to be interpreted on trend instead of value.

Model Selection:

There are a few evaluation scores for regression models, for example, Precision, Recall, F1 and AUC. Though these metrics focus on different aspects of models, AUC is particularly useful for comparing different models because AUC takes all possibility thresholds into account (James et al., 2013).

Therefore, in this case, AUC is the preferred model selection metric among these models. Below is the screenshot of the selection metric comparison among these four models.

	cm_rf.byClass	cm_lr.byClass	cm_dt.byClass	cm_las.byClass
Sensitivity	0.83554084	0.288079470	0.56181015	0.221854305
Specificity	0.99815215	0.997525760	0.99686805	0.998715901
Pos Pred Value	0.92769608	0.767647059	0.83579639	0.830578512
Neg Pred Value	0.99534651	0.980150792	0.98768075	0.978369589
Precision	0.92769608	0.767647059	0.83579639	0.830578512
Recall	0.83554084	0.288079470	0.56181015	0.221854305
F1	0.87921022	0.418940610	0.67194719	0.350174216
Prevalence	0.02759251	0.027592508	0.02759251	0.027592508
Detection Rate	0.02305467	0.007948835	0.01550175	0.006121517
Detection Prevalence	0.02485153	0.010354804	0.01854728	0.007370184
Balanced Accuracy	0.91684649	0.642802615	0.77933910	0.610285103
AUC	0.99634295	0.911962538	0.87858687	0.912386146

It suggests that the random forest model is the most preferred model with the largest AUC 0.9168. Also, the random forest model is not only holding the largest AUC but also the largest Precision, Recall and F1.

RESULT AND RECOMMENDATION

Random forest prediction shows that 1,246 (2.5%) customers in ‘repurchase_validation.csv’ are with Target = 1 which means they predicted to be a re-purchaser. However, to ensure not a single re-purchaser is missed, the car manufacturer can prepare two versions of re-purchase communication. The brief version can be sent to all customers with Target = 0 and a detailed version can be sent to those with Target = 1.

REFERENCE:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R.*

APPENDIX

Appendix 1

a. Explore the dataset

Structure of Raw Dataset

```
'data.frame':   131337 obs. of  17 variables:
 $ ID                        : int  1 2 3 5 6 7 8 9 10 11 ...
 $ Target                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ age_band                  : chr  "3. 35 to 44" "NULL" "NULL" "NULL" ...
 $ gender                    : chr  "Male" "NULL" "Male" "NULL" ...
 $ car_model                 : chr  "model_1" "model_2" "model_3" "model_3" ...
 $ car_segment               : chr  "LCV" "Small/Medium" "Large/SUV" "Large/SUV" ...
 $ age_of_vehicle_years      : int  9 6 9 5 8 7 8 7 1 3 ...
 $ sched_serv_warr           : int  2 10 10 8 9 4 2 4 2 1 ...
 $ non_sched_serv_warr       : int  10 3 9 5 4 10 8 9 1 1 ...
 $ sched_serv_paid           : int  3 10 10 8 10 5 2 6 1 2 ...
 $ non_sched_serv_paid       : int  7 4 9 4 7 7 9 9 3 1 ...
 $ total_paid_services       : int  5 9 10 5 9 6 9 8 1 2 ...
 $ total_services            : int  6 10 10 6 8 8 4 6 2 1 ...
 $ mth_since_last_serv       : int  9 6 7 4 5 8 7 9 1 1 ...
 $ annualised_mileage        : int  8 10 10 10 4 5 6 5 1 1 ...
 $ num_dealers_visited       : int  10 7 6 9 4 10 10 5 2 1 ...
 $ num_serv_dealer_purchased: int  4 10 10 7 9 4 4 8 3 1 ...
```

Summary of Dataset

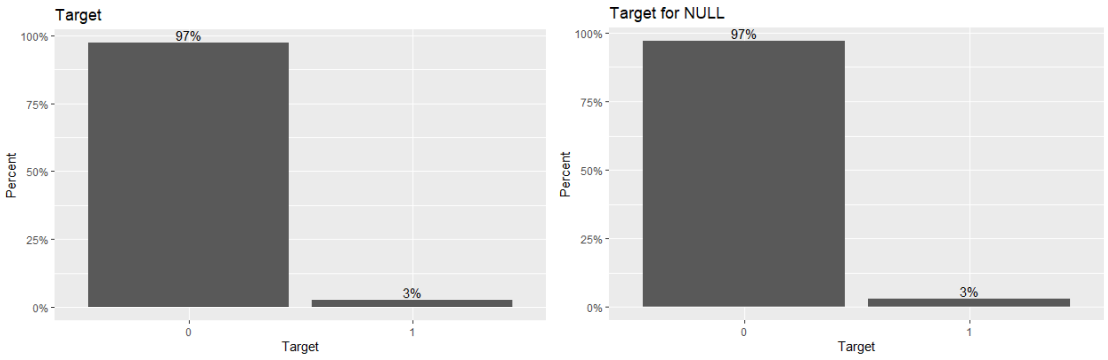
```
       ID              Target              age_band         gender        car_model
 Min.   :     1   Min.   :0.00000   NULL       :112375   Female:25957   2      :34491
 1st Qu.: 38563   1st Qu.:0.00000   4. 45 to 54:  4058   Male  :36072   5      :24674
 Median : 77132   Median :0.00000   3. 35 to 44:  3833   NULL  :69308   3      :17074
 Mean   : 77097   Mean   :0.02681   2. 25 to 34:  3548                  1      :15331
 3rd Qu.:115668   3rd Qu.:0.00000   5. 55 to 64:  3397                  4      :15155
 Max.   :154139   Max.   :1.00000   6. 65 to 74:  2140                  7      : 8167
                                    (Other)    :  1986                  (Other):16445
      car_segment    age_of_vehicle_years sched_serv_warr  non_sched_serv_warr sched_serv_paid
 Large/SUV   :52120   Min.   : 1.000       Min.   : 1.000   Min.   : 1.000      Min.   : 1.000
 LCV         :24606   1st Qu.: 3.000       1st Qu.: 3.000   1st Qu.: 3.000      1st Qu.: 3.000
 Other       :   58   Median : 5.000       Median : 5.000   Median : 5.000      Median : 5.000
 Small/Medium:54553   Mean   : 5.493       Mean   : 5.452   Mean   : 5.473      Mean   : 5.452
                      3rd Qu.: 8.000       3rd Qu.: 8.000   3rd Qu.: 8.000      3rd Qu.: 8.000
                      Max.   :10.000       Max.   :10.000   Max.   :10.000      Max.   :10.000

 non_sched_serv_paid total_paid_services total_services   mth_since_last_serv annualised_mileage
 Min.   : 1.000      Min.   : 1.000      Min.   : 1.000   Min.   : 1.00       Min.   : 1.000
 1st Qu.: 3.000      1st Qu.: 3.000      1st Qu.: 3.000   1st Qu.: 3.00       1st Qu.: 3.000
 Median : 5.000      Median : 5.000      Median : 5.000   Median : 5.00       Median : 5.000
 Mean   : 5.497      Mean   : 5.482      Mean   : 5.455   Mean   : 5.47       Mean   : 5.503
 3rd Qu.: 8.000      3rd Qu.: 8.000      3rd Qu.: 8.000   3rd Qu.: 8.00       3rd Qu.: 8.000
 Max.   :10.000      Max.   :10.000      Max.   :10.000   Max.   :10.00       Max.   :10.000

 num_dealers_visited num_serv_dealer_purchased
 Min.   : 1.000      Min.   : 1.000
 1st Qu.: 3.000      1st Qu.: 3.000
 Median : 5.000      Median : 5.000
 Mean   : 5.485      Mean   : 5.481
 3rd Qu.: 8.000      3rd Qu.: 8.000
 Max.   :10.000      Max.   :10.000
```

Levels of variable

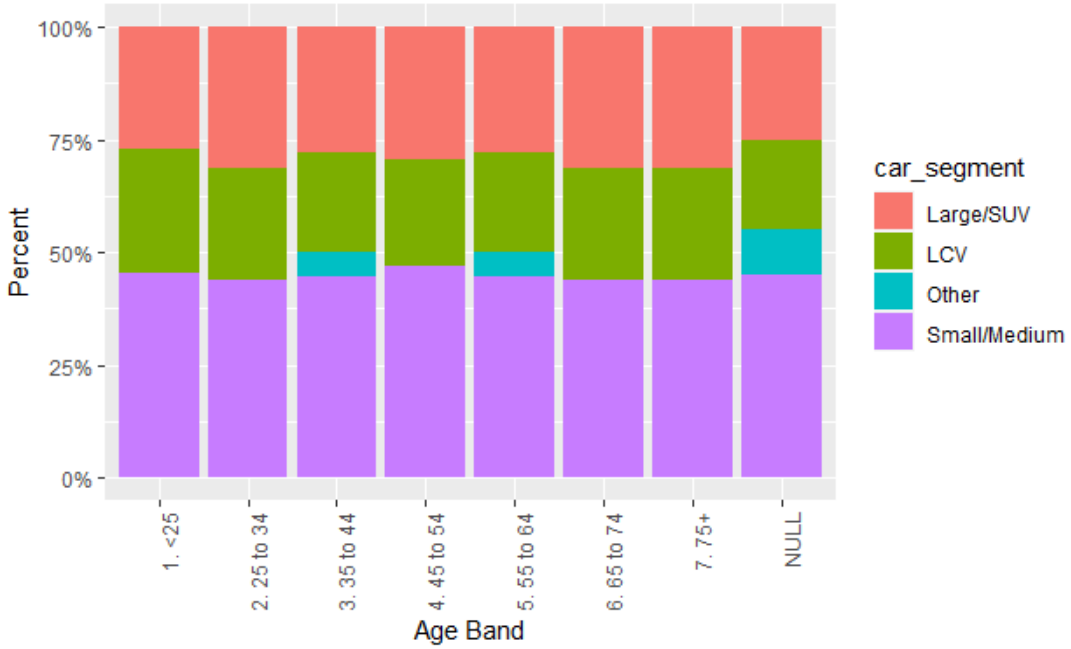
```
$Target

     0      1
127816   3521

$age_band

     1. <25 2. 25 to 34 3. 35 to 44 4. 45 to 54 5. 55 to 64 6. 65 to 74      7. 75+        NULL
        967        3548        3833        4058        3397        2140        1019      112375

$gender

Female   Male   NULL
 25957  36072  69308

$car_model

    1    10    11    12    13    14    15    16    17    18    19     2     3     4     5     6     7
15331  3215   612   614   714    78   334   114   153   101     2 34491 17074 15155 24674  3071  8167
    8     9
 6443   994

$car_segment

   Large/SUV          LCV        Other Small/Medium
       52120        24606           58        54553

$age_of_vehicle_years

    1     2     3     4     5     6     7     8     9    10
11893 13633 13825 13537 13438 13114 13227 12834 12840 12996

$sched_serv_warr

    1     2     3     4     5     6     7     8     9    10
13484 13788 13305 13129 12859 12972 12836 12938 13119 12907


$non_sched_serv_warr

    1     2     3     4     5     6     7     8     9    10
13246 13296 13210 13327 13311 13074 12974 12983 12873 13043

$sched_serv_paid

    1     2     3     4     5     6     7     8     9    10
13587 13718 13273 13122 12911 12837 12934 12910 13088 12957

$non_sched_serv_paid

    1     2     3     4     5     6     7     8     9    10
13145 13285 13195 13136 13020 13110 12921 13056 13223 13246

$total_paid_services

    1     2     3     4     5     6     7     8     9    10
13252 13451 13245 13127 12970 13081 12950 12964 13116 13181

$total_services

    1     2     3     4     5     6     7     8     9    10
13201 13626 13655 13205 12984 13089 12809 12825 12946 12997

$mth_since_last_serv

    1     2     3     4     5     6     7     8     9    10
13281 12434 13991 13548 13342 13152 12900 12916 12906 12867


$annualised_mileage

    1     2     3     4     5     6     7     8     9    10
12984 12669 13152 13459 13474 13372 13256 13039 12914 13018

$num_dealers_visited

    1     2     3     4     5     6     7     8     9    10
13265 13358 13207 12982 13064 13170 12965 13203 13038 13085

$num_serv_dealer_purchased

    1     2     3     4     5     6     7     8     9    10
13074 13379 13241 13132 13233 13247 13056 13072 12899 13004
```

b. correlation

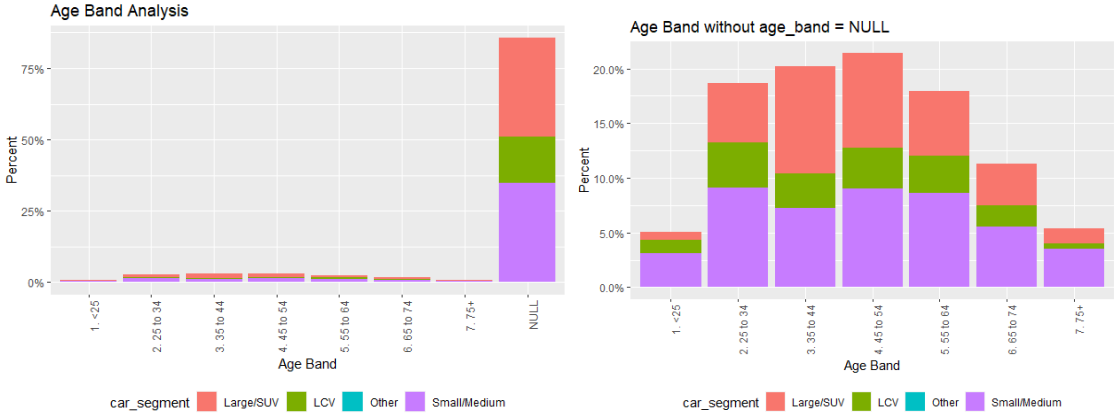
Variable: Target

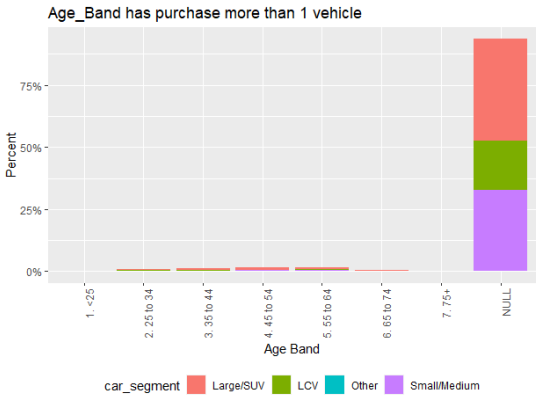


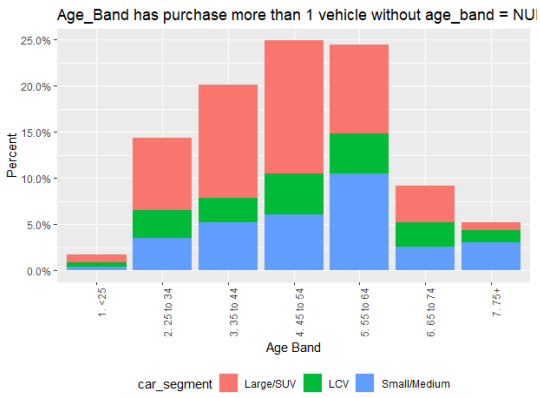
Variable: Car Segment



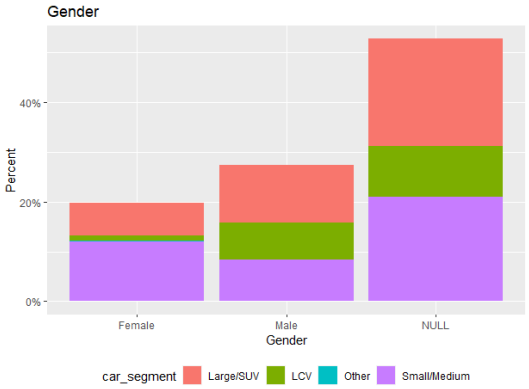
Variable: Age Band

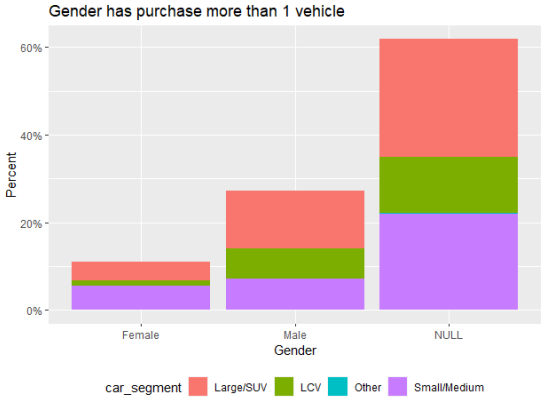




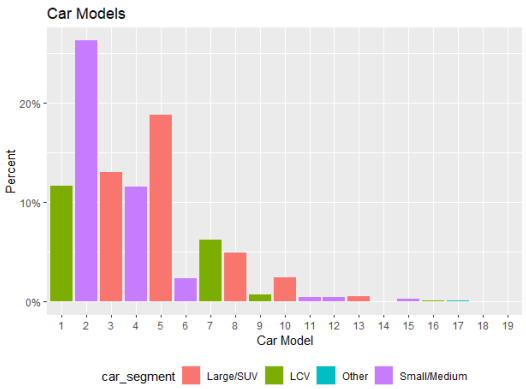


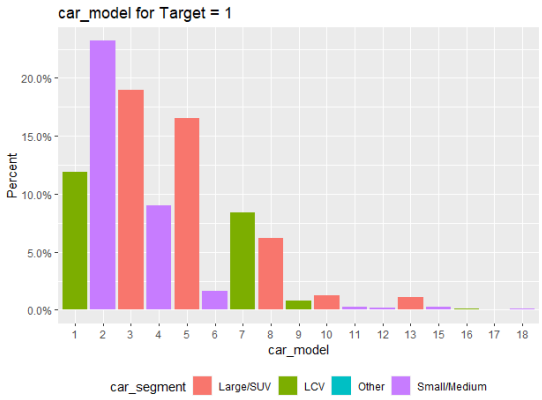
Variable: Gender

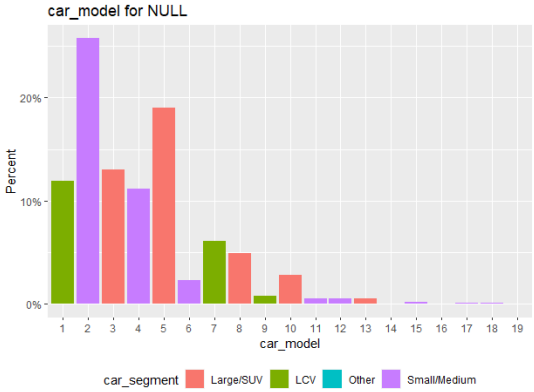


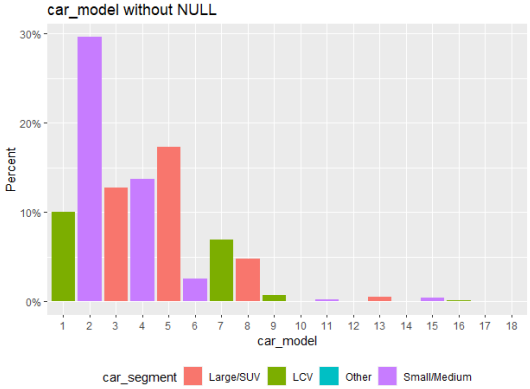


Variable: Car Models

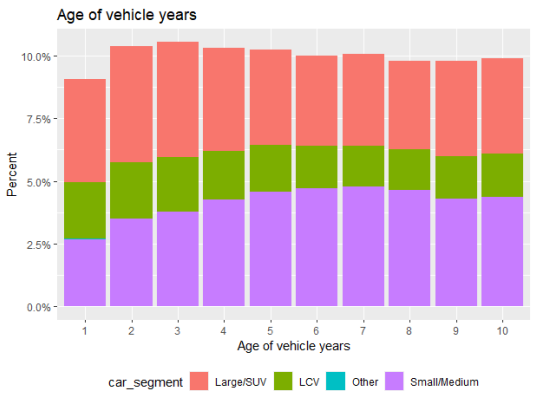


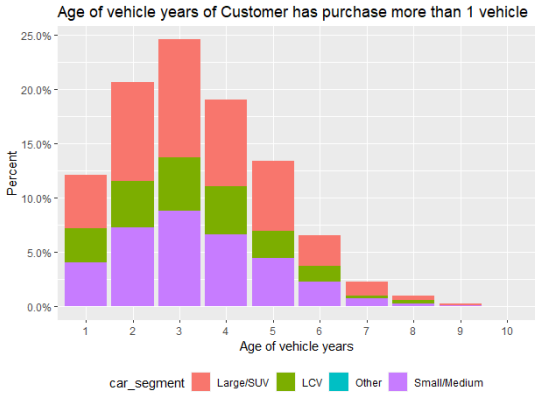


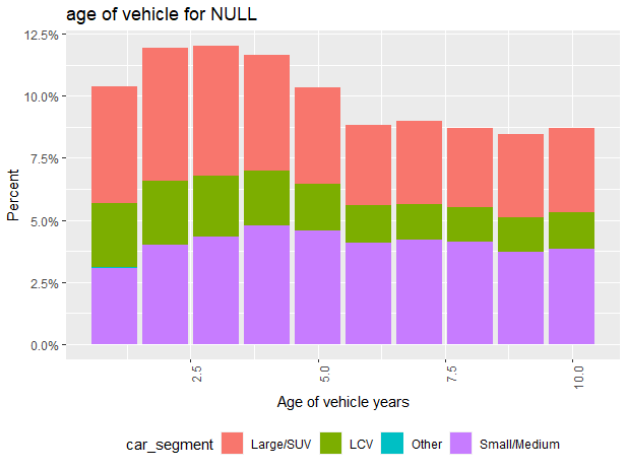




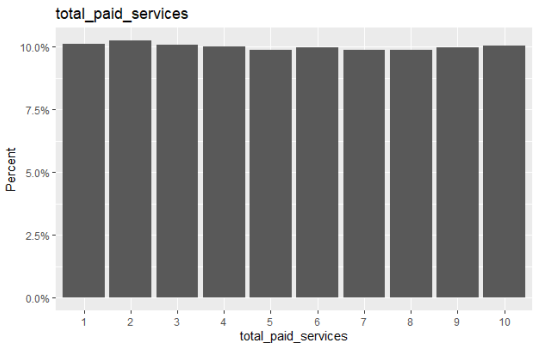
Variable: Age of Vehicle Years

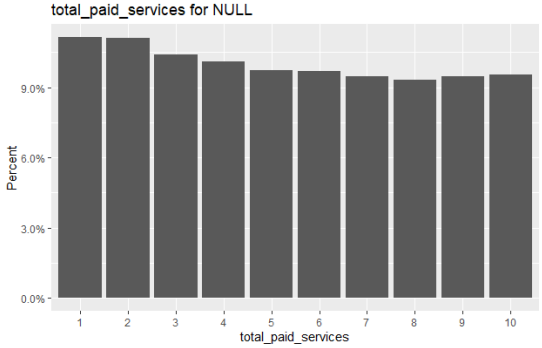


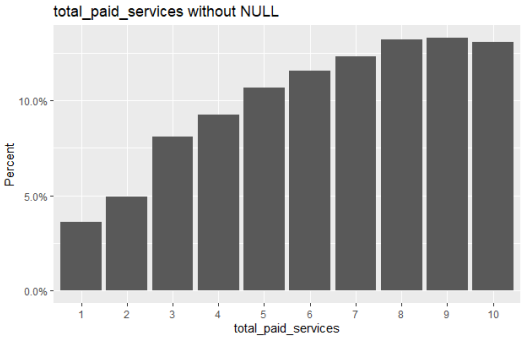




Variable: total paid services







Appendix 2

a. correlation

	ID	Target	age_of_vehicle_years	sched_serv_warr
ID	1.00000000	0.27971573	-0.0344684	-0.05119044
Target	0.27971573	1.00000000	-0.1240913	-0.17742357
age_of_vehicle_years	-0.03446840	-0.12409128	1.0000000	0.43148378
sched_serv_warr	-0.05119044	-0.17742357	0.4314838	1.00000000
non_sched_serv_warr	-0.02399252	-0.08132769	0.4275670	0.45247173
sched_serv_paid	-0.05103514	-0.17903979	0.5339139	0.86284251
non_sched_serv_paid	-0.01236023	-0.02839598	0.3305404	0.38545102
total_paid_services	-0.02731252	-0.08442882	0.4726686	0.64796246
total_services	-0.04671736	-0.15990503	0.5055771	0.82491698
mth_since_last_serv	-0.03983597	-0.14169452	0.6224721	0.27436211
annualised_mileage	-0.02254414	-0.07305477	0.4325934	0.74581124
num_dealers_visited	-0.01341630	-0.04724432	0.3671224	0.40034819
num_serv_dealer_purchased	-0.01672422	-0.05261752	0.3185578	0.59587740

	non_sched_serv_warr	sched_serv_paid	non_sched_serv_paid	total_paid_services
ID	-0.02399252	-0.05103514	-0.01236023	-0.02731252
Target	-0.08132769	-0.17903979	-0.02839598	-0.08442882
age_of_vehicle_years	0.42756703	0.53391391	0.33054041	0.47266857
sched_serv_warr	0.45247173	0.86284251	0.38545102	0.64796246
non_sched_serv_warr	1.00000000	0.46155020	0.77812570	0.71151055
sched_serv_paid	0.46155020	1.00000000	0.42144779	0.73400201
non_sched_serv_paid	0.77812570	0.42144779	1.00000000	0.85804534
total_paid_services	0.71151055	0.73400201	0.85804534	1.00000000
total_services	0.80616296	0.76856992	0.65347473	0.79122976
mth_since_last_serv	0.32803613	0.34099182	0.29164672	0.33737983
annualised_mileage	0.58543868	0.71378969	0.50579818	0.67593903
num_dealers_visited	0.46041560	0.42123198	0.41815665	0.47038832
num_serv_dealer_purchased	0.55610392	0.55243840	0.48004066	0.56786394

	total_services	mth_since_last_serv	annualised_mileage	num_dealers_visited
ID	-0.04671736	-0.03983597	-0.02254414	-0.01341630
Target	-0.15990503	-0.14169452	-0.07305477	-0.04724432
age_of_vehicle_years	0.50557707	0.62247212	0.43259336	0.36712236
sched_serv_warr	0.82491698	0.27436211	0.74581124	0.40034819
non_sched_serv_warr	0.80616296	0.32803613	0.58543868	0.46041560
sched_serv_paid	0.76856992	0.34099182	0.71378969	0.42123198
non_sched_serv_paid	0.65347473	0.29164672	0.50579818	0.41815665
total_paid_services	0.79122976	0.33737983	0.67593903	0.47038832
total_services	1.00000000	0.31090882	0.78364139	0.48050266
mth_since_last_serv	0.31090882	1.00000000	0.24410844	0.39675203
annualised_mileage	0.78364139	0.24410844	1.00000000	0.47883320
num_dealers_visited	0.48050266	0.39675203	0.47883320	1.00000000
num_serv_dealer_purchased	0.66830282	0.28607114	0.52838866	0.41088694

	num_serv_dealer_purchased
ID	-0.01672422
Target	-0.05261752
age_of_vehicle_years	0.31855784
sched_serv_warr	0.59587740
non_sched_serv_warr	0.55610392
sched_serv_paid	0.55243840
non_sched_serv_paid	0.48004066
total_paid_services	0.56786394
total_services	0.66830282
mth_since_last_serv	0.28607114
annualised_mileage	0.52838866
num_dealers_visited	0.41088694
num_serv_dealer_purchased	1.00000000

b. Logistic Regression

Model

```
Call:
glm(formula = Target ~ . - ID, family = "binomial", data = trainset_lr_all)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3163  -0.1868  -0.0617  -0.0238   4.8797

Coefficients: (2 not defined because of singularities)
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -4.48256    0.61694  -7.266 3.71e-13 ***
age_band2. 25 to 34        0.95156    0.65389   1.455  0.14561
age_band3. 35 to 44        1.09099    0.64525   1.691  0.09087 .
age_band4. 45 to 54        1.71863    0.62968   2.729  0.00635 **
age_band5. 55 to 64        1.78579    0.63114   2.829  0.00466 **
age_band6. 65 to 74        1.57819    0.66020   2.390  0.01683 *
age_band7. 75+             2.04855    0.69568   2.945  0.00323 **
age_bandNULL               1.69319    0.60845   2.783  0.00539 **
genderMale                 0.35088    0.07965   4.405 1.06e-05 ***
genderNULL                -0.15159    0.07460  -2.032  0.04215 *
car_model2                 0.66786    0.08201   8.144 3.82e-16 ***
car_model3                 0.81775    0.08519   9.599  < 2e-16 ***
car_model4                 0.57719    0.10348   5.578 2.44e-08 ***
car_model5                 0.28884    0.08632   3.346  0.00082 ***
car_model6                 0.48438    0.18484   2.621  0.00878 **
car_model7                 0.83343    0.10351   8.052 8.16e-16 ***
car_model8                 0.88342    0.11610   7.609 2.76e-14 ***
car_model9                 0.13278    0.26110   0.509  0.61106
car_model10               -0.93385    0.20157  -4.633 3.60e-06 ***
car_model11               -1.26736    0.40280  -3.146  0.00165 **
car_model12               -1.00624    0.52928  -1.901  0.05728 .
car_model13                1.57699    0.22176   7.111 1.15e-12 ***
car_model14              -12.13848  173.02142  -0.070  0.94407
car_model15                1.81857    0.43733   4.158 3.21e-05 ***
car_model16                0.52082    0.78754   0.661  0.50841
car_model17               -0.83404    1.03164  -0.808  0.41883
car_model18               -0.45998    0.58196  -0.790  0.42930
car_model19               -0.62752 1037.41597  -0.001  0.99952
car_segmentLCV                NA         NA      NA       NA
car_segmentOther          -11.67651  219.43004  -0.053  0.95756
car_segmentSmall/Medium       NA         NA      NA       NA
age_of_vehicle_years      -0.03165    0.01181  -2.680  0.00736 **
sched_serv_warr           -0.32095    0.02505 -12.812  < 2e-16 ***
non_sched_serv_warr        0.02255    0.02258   0.998  0.31812
sched_serv_paid           -0.29046    0.02203 -13.183  < 2e-16 ***
non_sched_serv_paid        0.31279    0.02671  11.711  < 2e-16 ***
total_paid_services       -0.04824    0.02750  -1.754  0.07937 .
total_services            -0.94027    0.03339 -28.159  < 2e-16 ***
mth_since_last_serv       -0.35037    0.01384 -25.307  < 2e-16 ***
annualised_mileage         0.44170    0.01296  34.069  < 2e-16 ***
num_dealers_visited        0.04780    0.01169   4.088 4.35e-05 ***
num_serv_dealer_purchased  0.47048    0.01673  28.122  < 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: 24139  on 98501  degrees of freedom
Residual deviance: 15389  on 98462  degrees of freedom
AIC: 15469

Number of Fisher Scoring iterations: 14
```

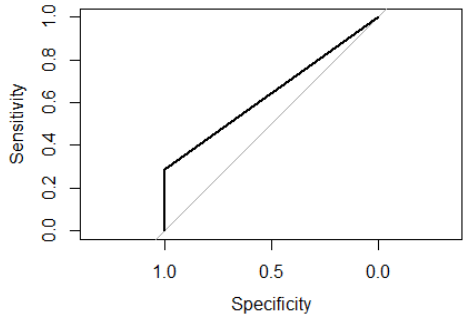
Confusion Matrix

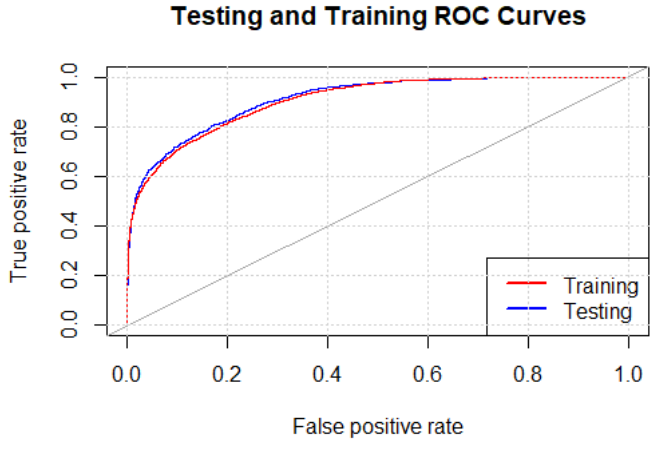
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 31850   645
         1    79   261

               Accuracy : 0.978
                 95% CI : (0.9763, 0.9795)
    No Information Rate : 0.9724
    P-Value [Acc > NIR] : 1.254e-10

                  Kappa : 0.4101

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.288079
            Specificity : 0.997526
         Pos Pred Value : 0.767647
         Neg Pred Value : 0.980151
             Prevalence : 0.027593
         Detection Rate : 0.007949
   Detection Prevalence : 0.010355
      Balanced Accuracy : 0.642803

       'Positive' Class : 1
```

```
> cm_lr_all$byClass
         Sensitivity          Specificity       Pos Pred Value       Neg Pred Value
         0.288079470          0.997525760          0.767647059          0.980150792
           Precision               Recall                   F1           Prevalence
         0.767647059          0.288079470          0.418940610          0.027592508
   Detection Rate Detection Prevalence    Balanced Accuracy
         0.007948835          0.010354804          0.642802615
```

AUC

```
Call:
roc.default(response = testset_lr_all$Target, predictor = testset_lr_all$prediction)

Data: testset_lr_all$prediction in 31929 controls (testset_lr_all$Target 0) < 906 cases (testset
_lr_all$Target 1).
Area under the curve: 0.6428
```





c. Decision Tree

Model

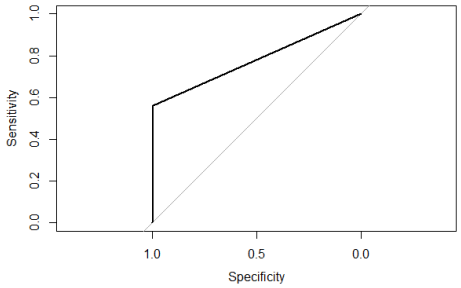
```
n= 98502

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 98502 2615 0 (0.973452316 0.026547684)
    2) sched_serv_warr>=3.5 68040  410 0 (0.993974133 0.006025867)
      4) mth_since_last_serv>=1.5 67974  344 0 (0.994939241 0.005060759) *
      5) mth_since_last_serv< 1.5 66    0 1 (0.000000000 1.000000000) *
    3) sched_serv_warr< 3.5 30462 2205 0 (0.927614733 0.072385267)
      6) annualised_mileage< 2.5 19168  516 0 (0.973080134 0.026919866)
       12) num_serv_dealer_purchased< 6.5 18963  403 0 (0.978748088 0.021251912) *
       13) num_serv_dealer_purchased>=6.5 205   92 1 (0.448780488 0.551219512)
         26) mth_since_last_serv>=6.5 51    4 0 (0.921568627 0.078431373) *
         27) mth_since_last_serv< 6.5 154   45 1 (0.292207792 0.707792208) *
      7) annualised_mileage>=2.5 11294 1689 0 (0.850451567 0.149548433)
       14) mth_since_last_serv>=4.5 7669  260 0 (0.966097275 0.033902725)
         28) total_paid_services>=1.5 7595  200 0 (0.973666886 0.026333114)
           56) age_of_vehicle_years>=1.5 7558  167 0 (0.977904207 0.022095793) *
           57) age_of_vehicle_years< 1.5 37    4 1 (0.108108108 0.891891892) *
         29) total_paid_services< 1.5 74   14 1 (0.189189189 0.810810811) *
       15) mth_since_last_serv< 4.5 3625 1429 0 (0.605793103 0.394206897)
         30) total_services>=2.5 2565  624 0 (0.756725146 0.243274854)
           60) mth_since_last_serv>=1.5 2448  507 0 (0.792892157 0.207107843)
            120) num_dealers_visited>=2.5 2108  335 0 (0.841081594 0.158918406)
              240) age_of_vehicle_years>=1.5 2070  297 0 (0.856521739 0.143478261)
                480) num_serv_dealer_purchased< 6.5 1583  136 0 (0.914087176 0.085912824) *
                481) num_serv_dealer_purchased>=6.5 487  161 0 (0.669404517 0.330595483)
                  962) total_services>=4.5 332   24 0 (0.927710843 0.072289157) *
                  963) total_services< 4.5 155   18 1 (0.116129032 0.883870968) *
              241) age_of_vehicle_years< 1.5 38    0 1 (0.000000000 1.000000000) *
            121) num_dealers_visited< 2.5 340  168 1 (0.494117647 0.505882353)
              242) gender=NULL 217   68 0 (0.686635945 0.313364055) *
              243) gender=Female,Male 123   19 1 (0.154471545 0.845528455) *
           61) mth_since_last_serv< 1.5 117    0 1 (0.000000000 1.000000000) *
         31) total_services< 2.5 1060  255 1 (0.240566038 0.759433962)
           62) sched_serv_paid>=3.5 100   16 0 (0.840000000 0.160000000) *
           63) sched_serv_paid< 3.5 960  171 1 (0.178125000 0.821875000) *
```

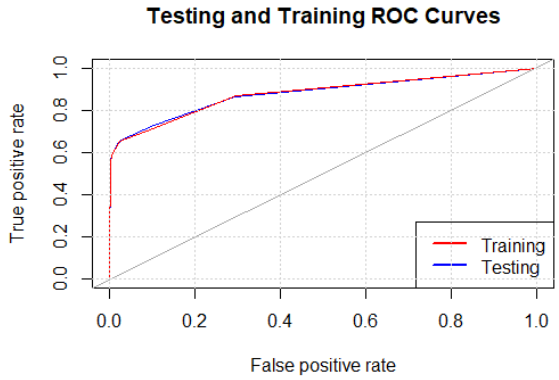
Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 31829   397
         1   100   509

               Accuracy : 0.9849
                 95% CI : (0.9835, 0.9862)
    No Information Rate : 0.9724
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6645

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.56181
            Specificity : 0.99687
         Pos Pred Value : 0.83580
         Neg Pred Value : 0.98768
             Prevalence : 0.02759
         Detection Rate : 0.01550
   Detection Prevalence : 0.01855
      Balanced Accuracy : 0.77934

       'Positive' Class : 1


   Sensitivity          Specificity       Pos Pred Value       Neg Pred Value
    0.56181015           0.99686805           0.83579639           0.98768075
     Precision               Recall                   F1           Prevalence
    0.83579639           0.56181015           0.67194719           0.02759251
Detection Rate Detection Prevalence    Balanced Accuracy
    0.01550175           0.01854728           0.77933910
```

AUC

```
Call:
roc.default(response = testset_dt_all$Target, predictor = testset_dt_all$prediction)

Data: testset_dt_all$prediction in 31929 controls (testset_dt_all$Target 0) < 906 cases (testset_dt_all$Tar
get 1).
Area under the curve: 0.7793
```





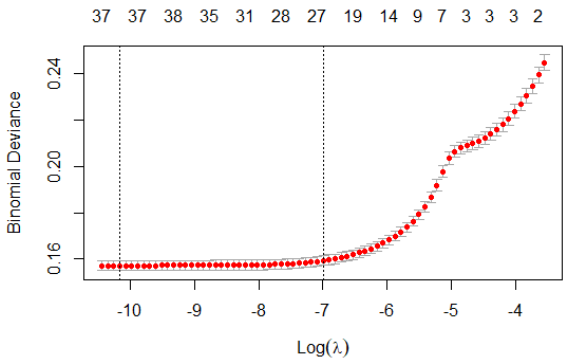
d. Lasso

Model

```
Call:  cv.glmnet(x = x, y = y, family = "binomial", alpha = 1)

Measure: Binomial Deviance

      Lambda Index Measure       SE Nonzero
min 3.87e-05    72  0.1572 0.002235      37
1se 9.15e-04    38  0.1593 0.002090      25
```

```
43 x 1 sparse Matrix of class "dgCMatrix"
                                      1
(Intercept)                -2.92045848
(Intercept)                 .
age_band2. 25 to 34         .
age_band3. 35 to 44         0.08343575
age_band4. 45 to 54         0.72035013
age_band5. 55 to 64         0.78958554
age_band6. 65 to 74         0.56196744
age_band7. 75+              1.02567212
age_bandNULL                0.71101400
genderMale                  0.35216902
genderNULL                 -0.13725234
car_model2                  0.09337695
car_model3                  0.24376688
car_model4                  .
car_model5                 -0.27446664
car_model6                 -0.07596111
car_model7                  0.81236372
car_model8                  0.30479060
car_model9                  0.10527491
car_model10                -1.47978375
car_model11                -1.78588672
car_model12                -1.50219191
car_model13                 0.99138652
car_model14                -3.75997471
car_model15                 1.20956360
car_model16                 0.45049139
car_model17                -1.33622337
car_model18                -0.97174757
car_model19                 .
car_segmentLCV             -0.55931504
car_segmentOther           -2.61472776
car_segmentSmall/Medium     .
age_of_vehicle_years       -0.03069905
sched_serv_warr            -0.32180101
non_sched_serv_warr         0.01631863
sched_serv_paid            -0.29228995
non_sched_serv_paid         0.29829428
total_paid_services        -0.03400547
total_services             -0.92289412
mth_since_last_serv        -0.34675045
annualised_mileage          0.43688735
num_dealers_visited         0.04787768
num_serv_dealer_purchased   0.46624520
```

Plot the model



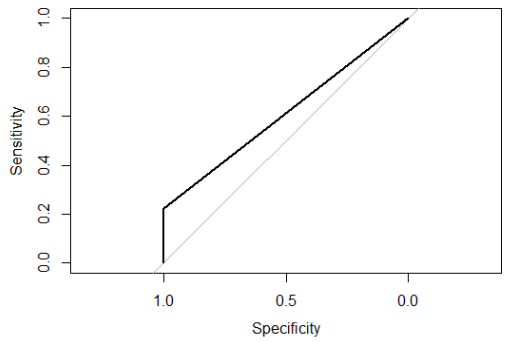
Confusion Matrix

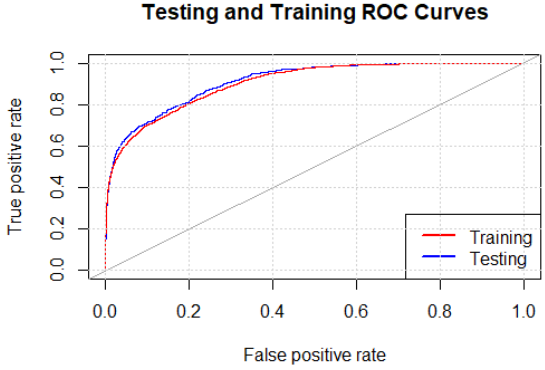
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 31888   705
         1    41   201

               Accuracy : 0.9773
                 95% CI : (0.9756, 0.9789)
    No Information Rate : 0.9724
    P-Value [Acc > NIR] : 1.565e-08

                  Kappa : 0.3425

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.221854
            Specificity : 0.998716
         Pos Pred Value : 0.830579
         Neg Pred Value : 0.978370
             Prevalence : 0.027593
         Detection Rate : 0.006122
   Detection Prevalence : 0.007370
      Balanced Accuracy : 0.610285

       'Positive' Class : 1
```

Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
0.221854305	0.998715901	0.830578512	0.978369589
Precision	**Recall**	**F1**	**Prevalence**
0.830578512	0.221854305	0.350174216	0.027592508
Detection Rate	**Detection Prevalence**	**Balanced Accuracy**	
0.006121517	0.007370184	0.610285103	

AUC

```
Call:
roc.default(response = testset_las$Target, predictor = testset_las$prediction)

Data: testset_las$prediction in 31929 controls (testset_las$Target 0) < 906 cases
 (testset_las$Target 1).
Area under the curve: 0.6103
```





e. Random Forest

Model

```
Call:
 randomForest(formula = Target ~ . - ID, data = trainset_rf, importance = TRUE,
   xtest = testset_rf[, c(-1, -2)], keep.forest = TRUE, ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0.64%
Confusion matrix:
      0    1  class.error
0 95799   88 0.0009177469
1   546 2069 0.2087954111
```

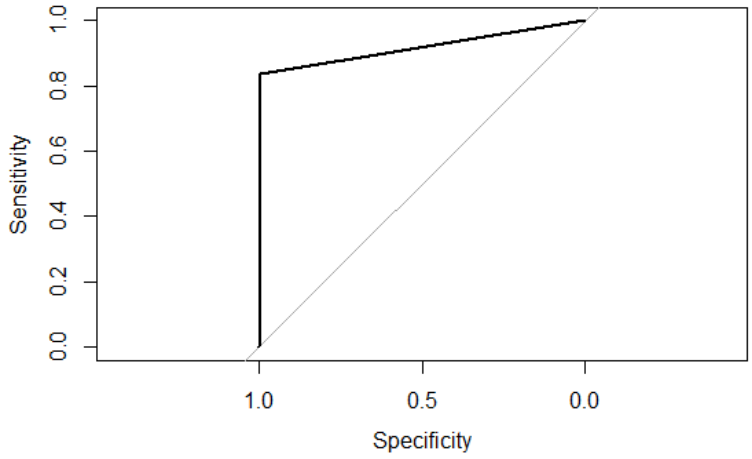
Confusion Matrix

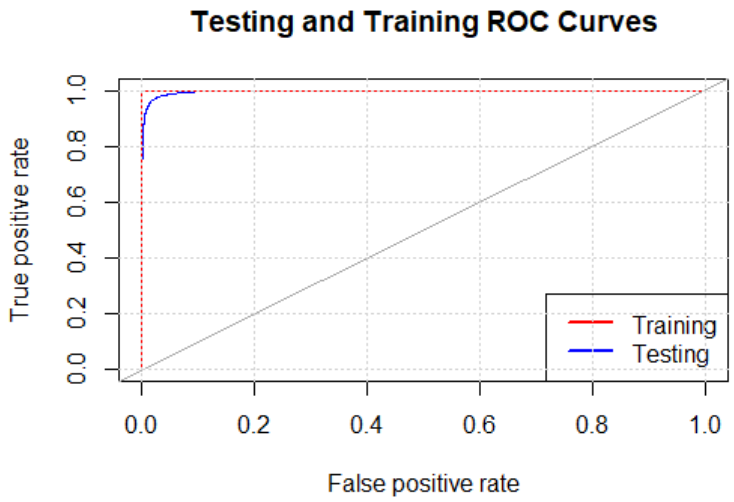
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 31870   149
         1    59   757

               Accuracy : 0.9937
                 95% CI : (0.9927, 0.9945)
    No Information Rate : 0.9724
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.876

 Mcnemar's Test P-Value : 6.784e-10

            Sensitivity : 0.83554
            Specificity : 0.99815
         Pos Pred Value : 0.92770
         Neg Pred Value : 0.99535
             Prevalence : 0.02759
         Detection Rate : 0.02305
   Detection Prevalence : 0.02485
      Balanced Accuracy : 0.91685

       'Positive' Class : 1
```

Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
0.83554084	0.99815215	0.92769608	0.99534651
Precision	**Recall**	**F1**	**Prevalence**
0.92769608	0.83554084	0.87921022	0.02759251
Detection Rate	**Detection Prevalence**	**Balanced Accuracy**	
0.02305467	0.02485153	0.91684649	

AUC

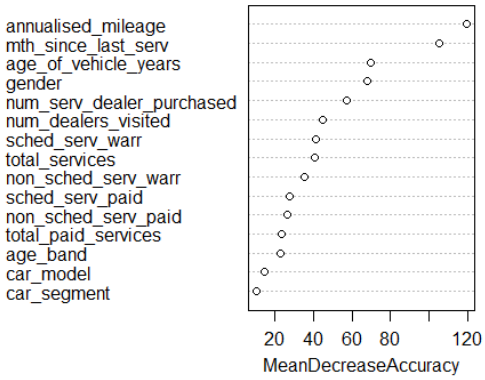
```
Call:
roc.default(response = testset_rf$Target, predictor = testset_rf$prediction)

Data: testset_rf$prediction in 31929 controls (testset_rf$Target 0) < 906 cases
 (testset_rf$Target 1).
Area under the curve: 0.9168
```

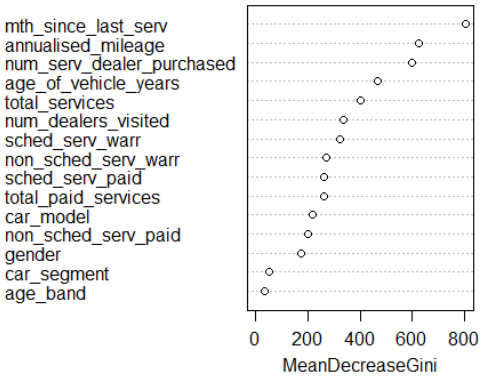




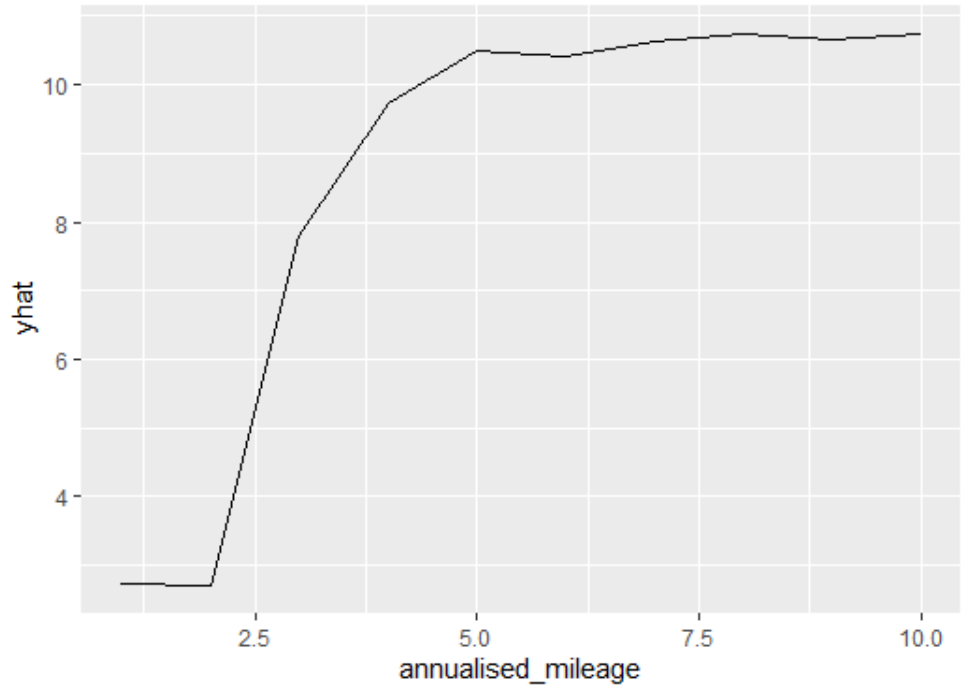
Appendix 3 - Importance of Variable

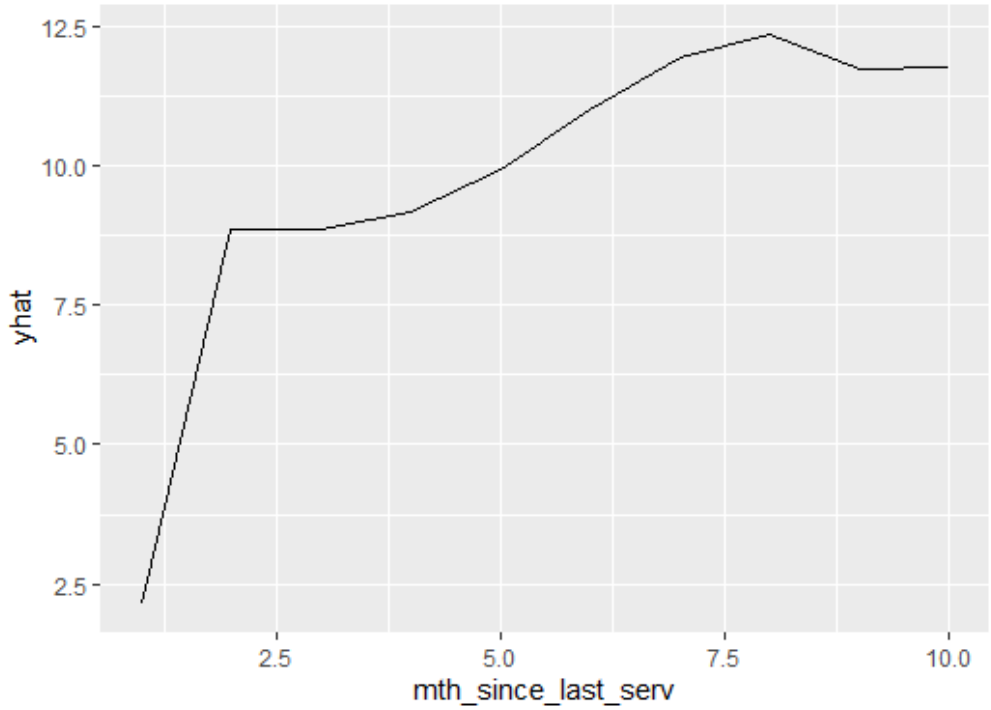
	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
age_band	22.92785	3.854558	22.54283	33.82671
gender	67.41494	22.889241	68.05048	173.82855
car_model	14.53633	1.671146	14.68839	216.72677
car_segment	10.79494	-5.860180	10.34835	51.88446
age_of_vehicle_years	69.09117	13.369631	69.57151	467.75144
sched_serv_warr	39.83900	46.414041	41.06362	321.43766
non_sched_serv_warr	34.43030	31.724244	35.04365	271.60479
sched_serv_paid	26.78388	31.855751	27.44374	263.70272
non_sched_serv_paid	26.33788	19.261041	26.51394	202.96139
total_paid_services	23.34150	17.284429	23.51263	260.71503
total_services	39.85762	48.419248	40.62588	402.01562
mth_since_last_serv	102.46595	79.787299	105.07619	803.95867
annualised_mileage	119.25679	7.506576	119.46656	625.11181
num_dealers_visited	43.48387	50.953036	44.69867	336.25176
num_serv_dealer_purchased	56.94795	15.241720	57.30340	600.75401

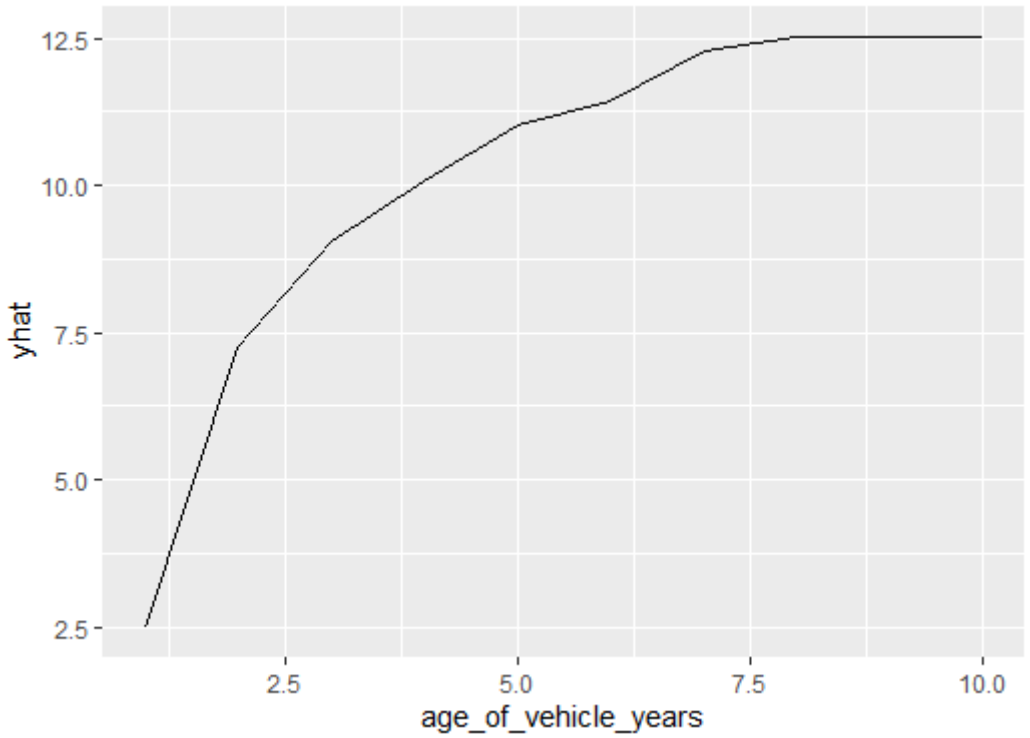


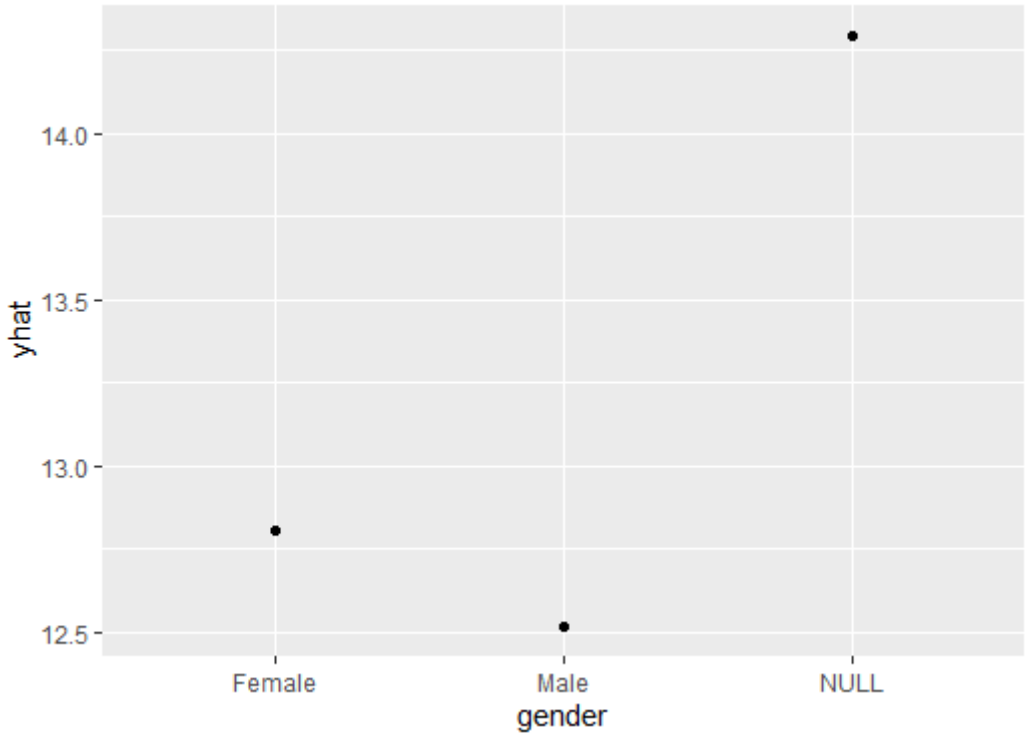


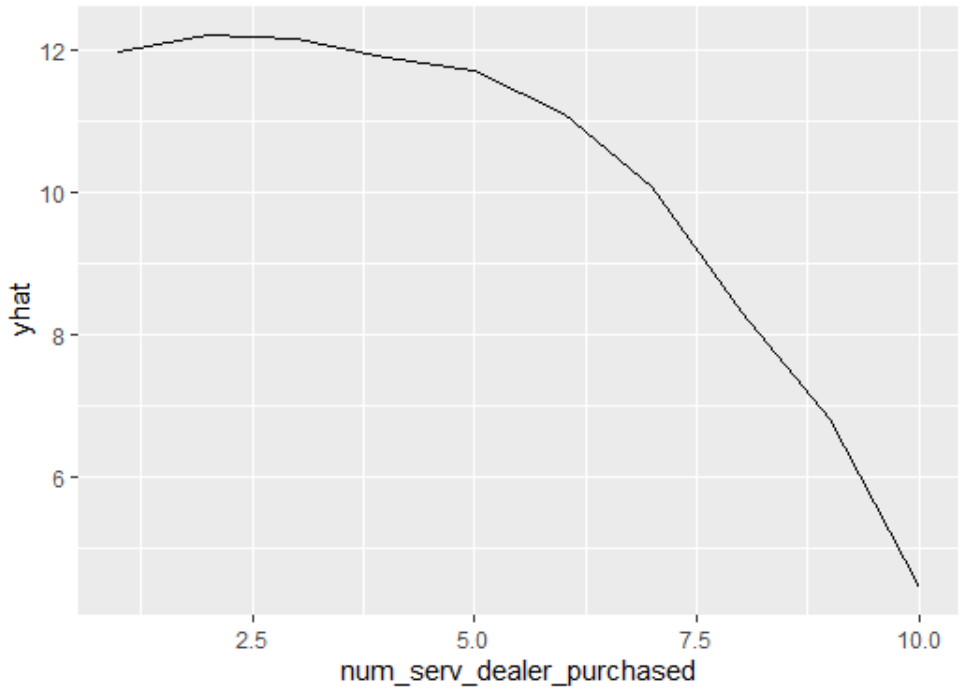
Appendix 4 - Partial Dependency Plots






12.5
10.0
7.5
5.0
2.5
yhat
2.5
5.0
7.5
10.0
age_of_vehicle_years


14.0
13.5
13.0
12.5
yhat
Female
Male
NULL
gender


12
10
8
6
yhat
2.5
5.0
7.5
10.0
num_serv_dealer_purchased