2022 AUTUMN SESSION

36106, MACHINE LEARNING AND ALGORITHMS

ASSESSMENT TASK 1 – PART B

LINEAR REGRESSION AND CLASSIFICATION MODELLING

ANNA LY (12945604)

1 Business Understanding

An automotive manufacturer wants to target existing customers for a re-purchase campaign. International consulting company analytics team analysed the dataset and used Rstudio for analysis and modelling. The aim was to send communication to customers who are highly likely to purchase a new vehicle.

By using CRISP-DM model, we can study the target existing customers for a repurchase campaign.
The following three business questions are addressed in this report:

- What classification algorithms can be applied to training data and future data sets?
- What are the top predictors for customers purchasing a new vehicle?
- What levels of importance are the vehicles features?

2 Data Understanding

The automotive company has supplied two data sets: repurchase_training.csv and repurchase_validation.csv which included customer demographics, previous car type bought, the age of the vehicle, and servicing details that was only for mechanics at official dealerships.

<u>Exploratory Data Analysis (EDA)</u>

- Needed to change ID to character in both datasets, so they are not counted as numbers in analysis.
- Find missing values in data to find out which columns have empty records. As shown in Figure 1 below, was found that:
    - ‘age_band’ had very few values across the training set.
    - gender only has values in about 50% of rows.

Neither of these variables are likely to be usable in our model

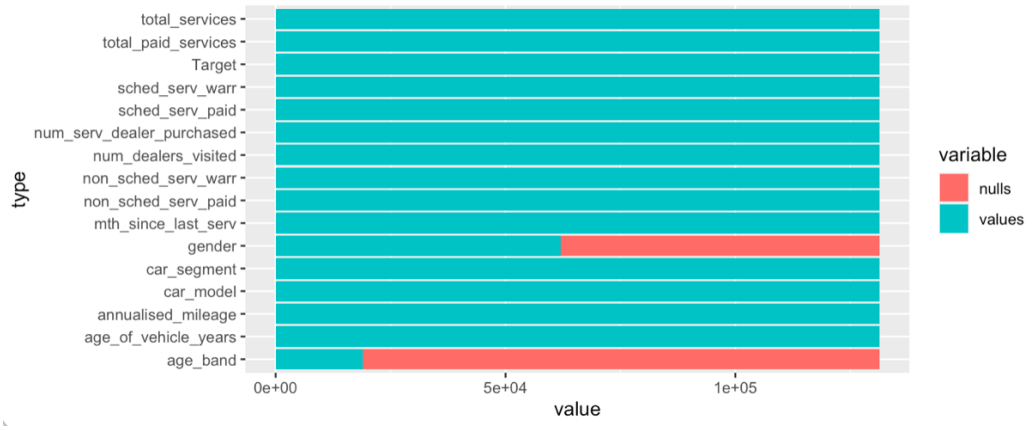


Figure 1 Missing values in data set

- First we write a function that will take a characteristic and return a data frame that gives us the count of entries with target = 1 and target = 0.

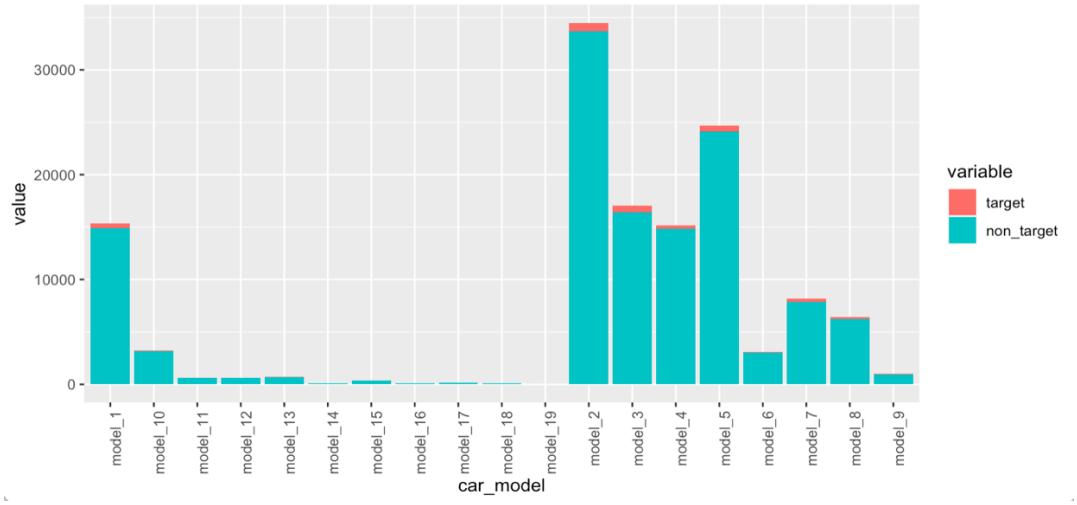


Figure 2 Melt and plot car models

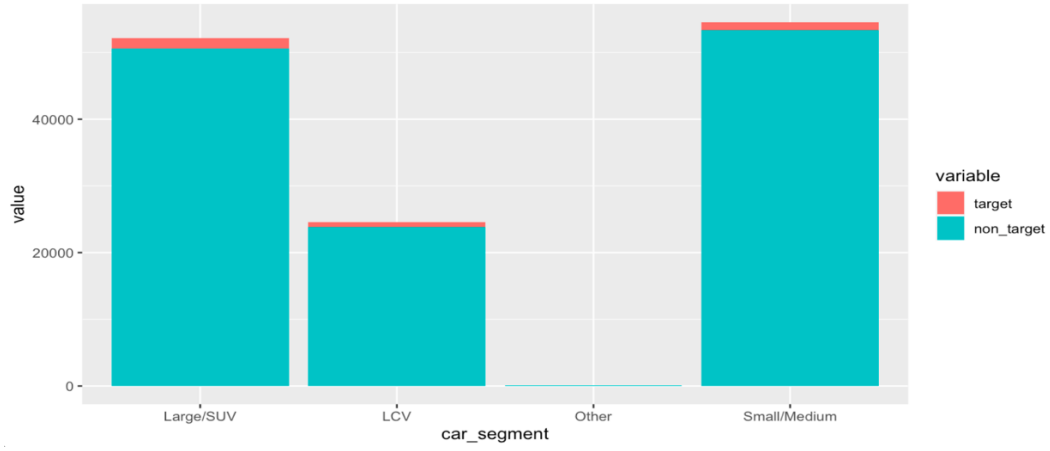


Figure 3 Plot car segments variable

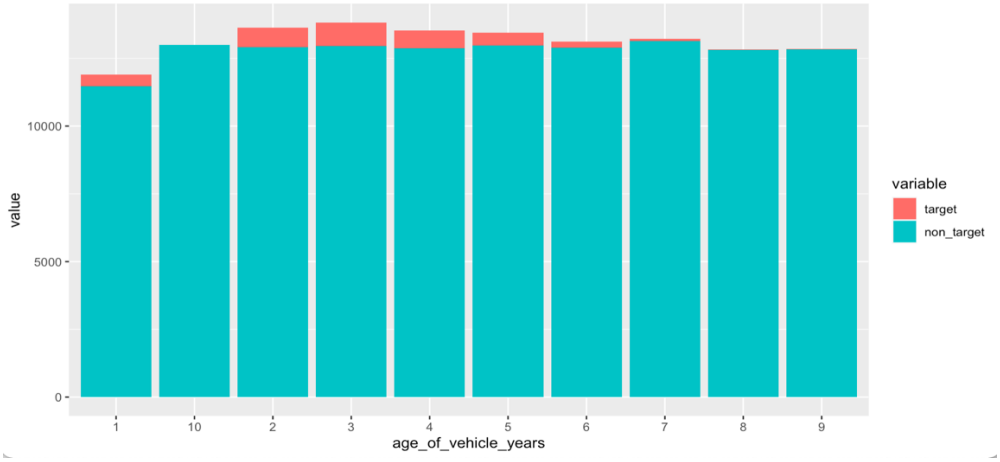


Figure 4 Plot age of vehicle in years variable

- Look at some service history characteristics of the cars to see if we can detect any patterns

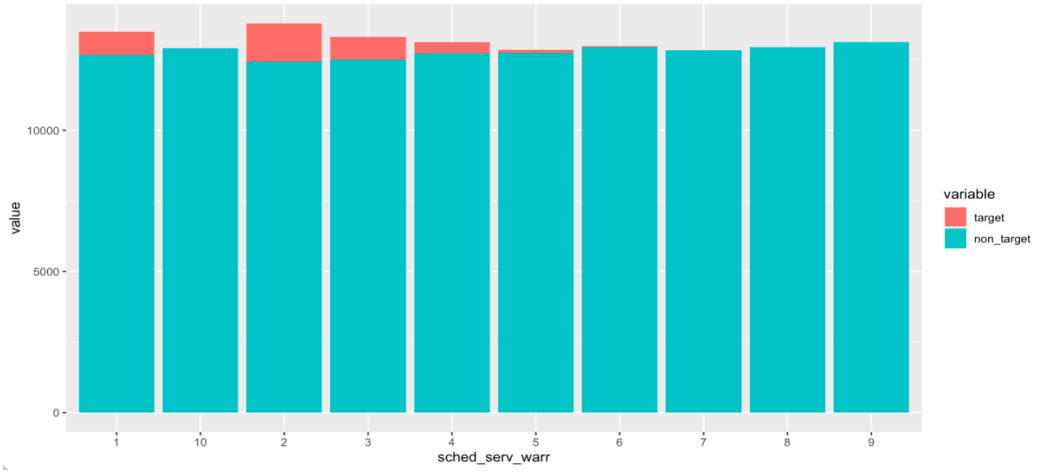


Figure 5 Number of scheduled services used under warranty

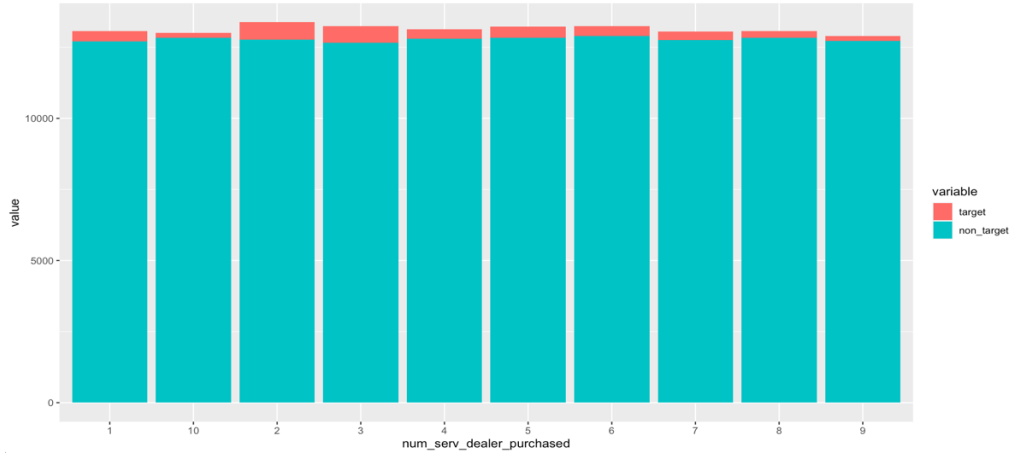


Figure 6 Number of services had at the same dealer where the vehicle was purchased

Linear classification model:

- Run a linear regression model (lrm) on out test set, excluding IDs

```
> glm.model = glm(formula = Target ~ ., family = binomial, data = lrm_data[,-1])
> glm.model

Call:  glm(formula = Target ~ ., family = binomial, data = lrm_data[,
    -1])

Coefficients:
              (Intercept)       age_band2. 25 to 34       age_band3. 35 to 44
                -4.622319                  1.229115                  1.404695
      age_band4. 45 to 54       age_band5. 55 to 64       age_band6. 65 to 74
                 1.656169                  1.774010                  1.489551
           age_band7. 75+             age_bandNULL                genderMale
                 1.985368                  1.782812                  0.426431
               genderNULL        car_modelmodel_10         car_modelmodel_11
                -0.109306                 -0.793655                 -1.368900
        car_modelmodel_12         car_modelmodel_13         car_modelmodel_14
                -1.014117                  1.444475                -11.535142
        car_modelmodel_15         car_modelmodel_16         car_modelmodel_17
                 1.611429                  0.840084                 -1.291977
        car_modelmodel_18         car_modelmodel_19          car_modelmodel_2
                -0.336041                -11.630748                  0.647902
         car_modelmodel_3          car_modelmodel_4          car_modelmodel_5
                 0.821045                  0.656971                  0.309173
         car_modelmodel_6          car_modelmodel_7          car_modelmodel_8
                 0.447349                  0.843444                  0.909639
         car_modelmodel_9           car_segmentLCV          car_segmentOther
                -0.001115                        NA                  0.343629
  car_segmentSmall/Medium      age_of_vehicle_years           sched_serv_warr
                      NA                 -0.037538                 -0.313279
     non_sched_serv_warr          sched_serv_paid       non_sched_serv_paid
                 0.024260                 -0.294926                  0.321299
     total_paid_services          total_services       mth_since_last_serv
                -0.048576                 -0.956269                 -0.354847
      annualised_mileage       num_dealers_visited  num_serv_dealer_purchased
                 0.449782                  0.045751                  0.472887

Degrees of Freedom: 131336 Total (i.e. Null);  131297 Residual
Null Deviance:      32430
Residual Deviance: 20540        AIC: 20620
```

Figure 7 Linear regression model 1

- We want to keep our Target as Factor
- Create training and test sets of our lrm_data
    - 70% of the lrm_data, use floor to round down to nearest integer

- Try linear regression model on our newly cleaned test set

```
> summary(glm.model)

Call:
glm(formula = Target ~ ., family = binomial, data = lrm_testset[,
    -1])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2825  -0.1994  -0.0574  -0.0212   4.6102

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.00383    0.17979 -11.146  < 2e-16 ***
car_model                 -0.01671    0.01218  -1.372   0.1701
car_segment               -0.09131    0.04827  -1.892   0.0586 .
age_of_vehicle_years      -0.02845    0.01830  -1.554   0.1201
sched_serv_warr           -0.31713    0.03947  -8.035 9.38e-16 ***
non_sched_serv_warr        0.03048    0.03511   0.868   0.3853
sched_serv_paid           -0.29485    0.03417  -8.630  < 2e-16 ***
non_sched_serv_paid        0.34254    0.04232   8.094 5.79e-16 ***
total_paid_services       -0.06078    0.04317  -1.408   0.1591
total_services            -0.98927    0.05221 -18.947  < 2e-16 ***
mth_since_last_serv       -0.36283    0.02205 -16.456  < 2e-16 ***
annualised_mileage         0.46998    0.02034  23.103  < 2e-16 ***
num_dealers_visited        0.02567    0.01812   1.417   0.1565
num_serv_dealer_purchased  0.47946    0.02558  18.741  < 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: 9942.1  on 39401  degrees of freedom
Residual deviance: 6299.0  on 39388  degrees of freedom
AIC: 6327

Number of Fisher Scoring iterations: 9
```

Figure 8 Linear regression model 2

- Now predict probabilities on lrm test set
- Create a confusion matrix

```
> glm_confusion_matrix
     true
pred     0     1
   0 38254   838
   1    62   248
```

- We will want to look at evaluation measures regularly, so create function to calculate and return them

```
> lrm_evaluation_measures
                name  accuracy precision   recall        F1
1 Linear Regression 0.9771585       0.8 0.228361 0.3553009
```

- Now to get AUC, write a function and add it to our evaluation measures data frame
  - lrm_evaluation_measures as the first row

```
> evaluation_measures
                name  accuracy precision   recall        F1
1 Linear Regression 0.9771585       0.8 0.228361 0.3553009
```

LASSO

- Now we want to try LASSO to perform grid search to find optimal value of lambda IE regularise the model

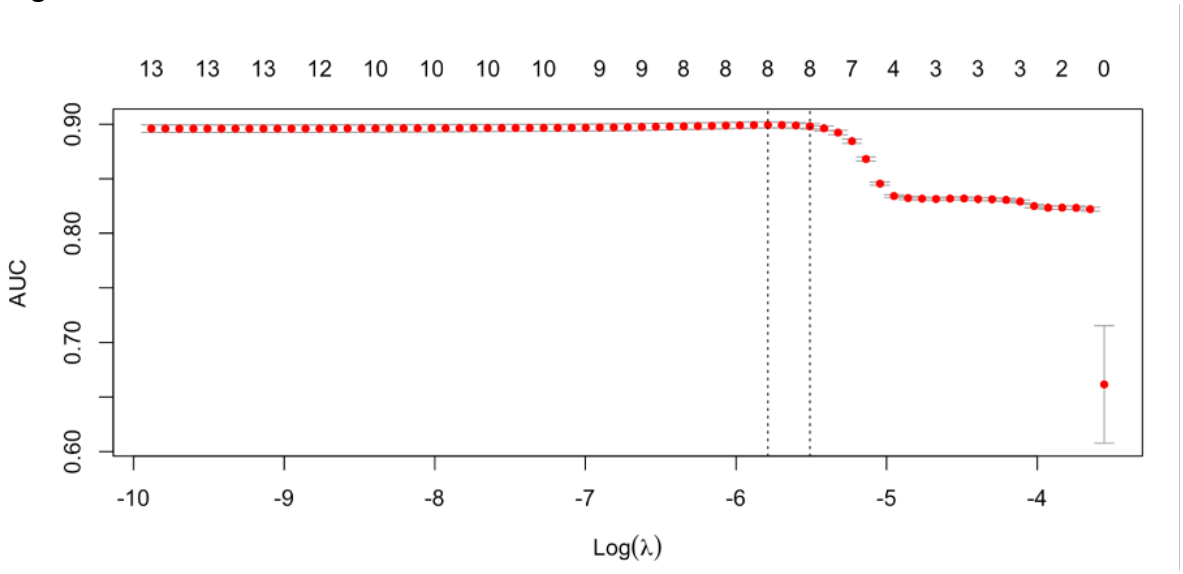


Figure 9 LASSO Grid Search

- Using lambda of 1se rather than the minimum lambda, see what predictors are discarded
  - Min value of lambda: 0.003063237
  - Best value of lambda: 0.004049419

- Regression co-efficient:

```
> coef(lrm_cv.out,s=lrm_lambda_1se)
15 x 1 sparse Matrix of class "dgCMatrix"
                                    s1
(Intercept)                -2.02885943
(Intercept)                 .
car_model                   .
car_segment                 .
age_of_vehicle_years        .
sched_serv_warr            -0.21925165
non_sched_serv_warr         .
sched_serv_paid            -0.22020908
non_sched_serv_paid         0.09489035
total_paid_services         .
total_services             -0.25693734
mth_since_last_serv        -0.16585332
annualised_mileage          0.15973725
num_dealers_visited         0.01266522
num_serv_dealer_purchased   0.15410491
```

- Convert test data to a model matrix
- Get prediction probabilities
- Lasso confusion matrix

```
> lasso_confusion_matrix
     true
pred     0     1
   0 38316  1084
   1     0     2
```

- Get the AUC and add it to our evaluation measures data frame

```
> lasso_evaluation_measures
   name  accuracy precision      recall          F1       AUC
1 Lasso 0.9724887         1 0.001841621 0.003676471 0.9059771
```

- Visually compare two models evaluation metrics

```
> evaluation_measures
               name  accuracy precision      recall          F1       AUC
1 Linear Regression 0.9771585       0.8 0.228360958 0.355300860 0.9099919
2             Lasso 0.9724887       1.0 0.001841621 0.003676471 0.9059771
```

3 Data Preparation

In the data preparation phase, Rstudio was used for to clean data. We were able to transform many variables to make modelling and analysis easier.

4 Modelling

The data is modelled using we used a tree-based classification model to predict which customers are most likely to repurchase as structured below.

Tree based classification

- Remove columns with high NAs – excluding age and gender and ID

- Create training and test sets
  - 70% of the sample size, use floor to round down to nearest integer
  - Assign observations to training and test sets
  - This is a classification problem so set method = “class” and exclude ID
- Create a confusion matrix

```
> rpart_confusion_matrix
      true
pred     0     1
   0 38204   492
   1   112   594
```

  - Cannot get AUC

```
> evaluation_measures
               name  accuracy precision      recall          F1       AUC
1 Linear Regression 0.9771585 0.8000000 0.228360958 0.355300860 0.9099919
2             Lasso 0.9724887 1.0000000 0.001841621 0.003676471 0.9059771
3    RPart unpruned 0.9846708 0.8413598 0.546961326 0.662946429        NA
```

- Value of the complexity parameter (alpha) for that gives a tree of that size
  - ‘Prune’ the tree using that value of the complexity parameter

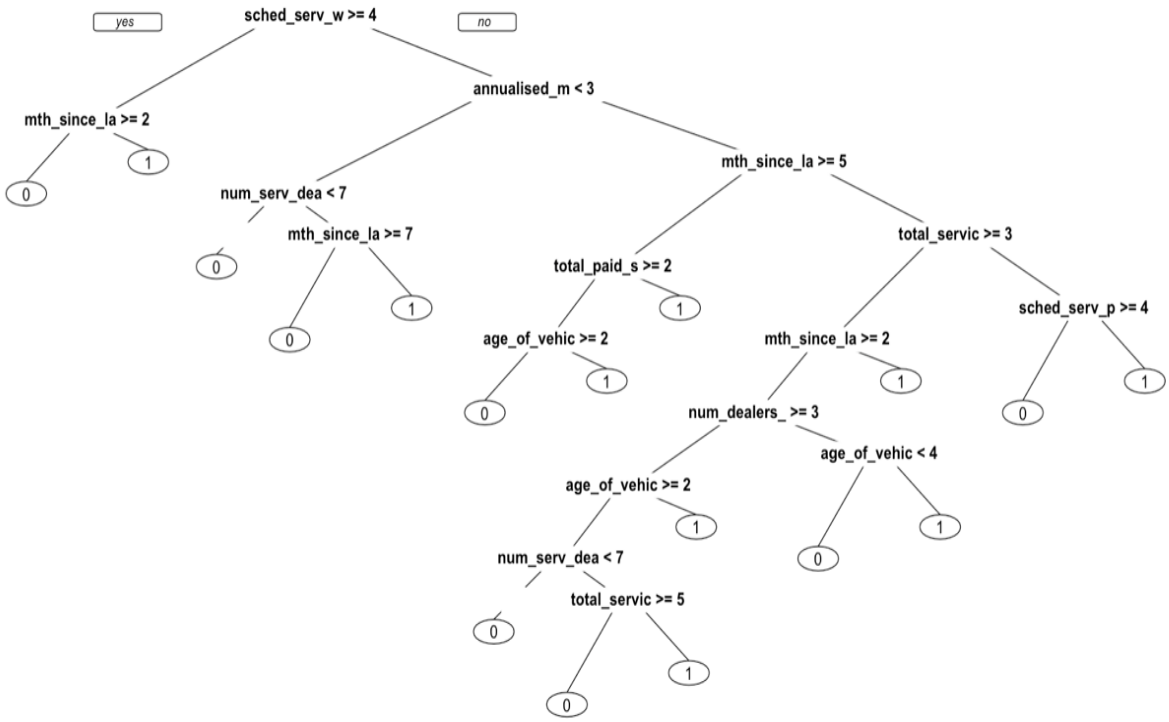


Figure 10 Plot pruned tree

- Confusion matrix (Pruned Model):

```
> rpart_pruned_confusion_matrix
    true
pred     0     1
   0 38204   492
   1   112   594
```

- Convert confusion matrix to data frame
- Cannot get AUC

```
> evaluation_measures
              name  accuracy precision      recall          F1       AUC
1 Linear Regression 0.9771585 0.8000000 0.228360958 0.355300860 0.9099919
2            Lasso 0.9724887 1.0000000 0.001841621 0.003676471 0.9059771
3    RPart unpruned 0.9846708 0.8413598 0.546961326 0.662946429        NA
4      RPart pruned 0.9846708 0.8413598 0.546961326 0.662946429        NA
```

Partial Dependency Plots (PDP)

- Construct top 5 predictors Partial Dependency Plots

- Build Random Forest Model and confusion matrix

```
> rf_confusion_matrix
    true
pred     0     1
   0 38255   270
   1    61   816
```

- Get AUC and add it to our evaluation measures data frame

```
> evaluation_measures
              name  accuracy precision      recall          F1       AUC
1 Linear Regression 0.9771585 0.8000000 0.228360958 0.355300860 0.9099919
2            Lasso 0.9724887 1.0000000 0.001841621 0.003676471 0.9059771
3    RPart unpruned 0.9846708 0.8413598 0.546961326 0.662946429        NA
4      RPart pruned 0.9846708 0.8413598 0.546961326 0.662946429        NA
5     Random Forest 0.9915994 0.9304447 0.751381215 0.831380540 0.9726184
```

- Quantitative measure of variable importance

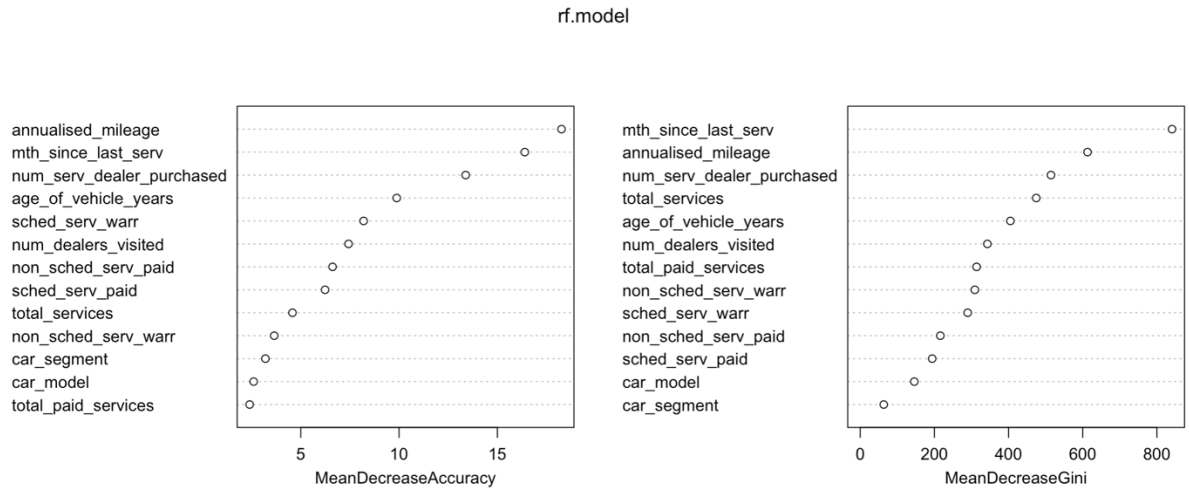
```
> importance(rf.model)
                                0          1 MeanDecreaseAccuracy MeanDecreaseGini
car_model                2.392773  1.3505779             2.605272        145.78399
car_segment              3.730051 -0.7295766             3.204021         63.48697
age_of_vehicle_years    10.033931  1.0472400             9.879635        404.93732
sched_serv_warr          8.012983  8.0134959             8.196708        289.77039
non_sched_serv_warr      3.555333  5.7902630             3.651554        309.09483
sched_serv_paid          6.160170  4.2716116             6.236445        194.07129
non_sched_serv_paid      6.621974  1.2565487             6.617860        216.11870
total_paid_services      2.359423  3.5130409             2.400776        313.83332
total_services           4.466082  6.2869701             4.577858        474.71505
mth_since_last_serv     16.189961 10.3869204            16.393161        840.83222
annualised_mileage      18.134159  1.1102167            18.260267        613.01008
num_dealers_visited      7.148902  5.9586025             7.430406        342.91994
num_serv_dealer_purchased 13.236560  0.5015671            13.387480        514.39543
```

Variables with high importance are drivers of the outcome and their values have a significant impact on the outcome values.

The random forest model fits multiple trees, and each tree is built by randomly selecting different features in the dataset. The overall prediction of forest is made of averages of individual trees.

When model is fit on training data on linear regression model it is different, modelling nonlinearities and interactions. Shrinkage methods are used (LASSO) in case linear regression breaks down.

- Look at variable importance using built in varImpPlot from randomForest



- Use vip package instead

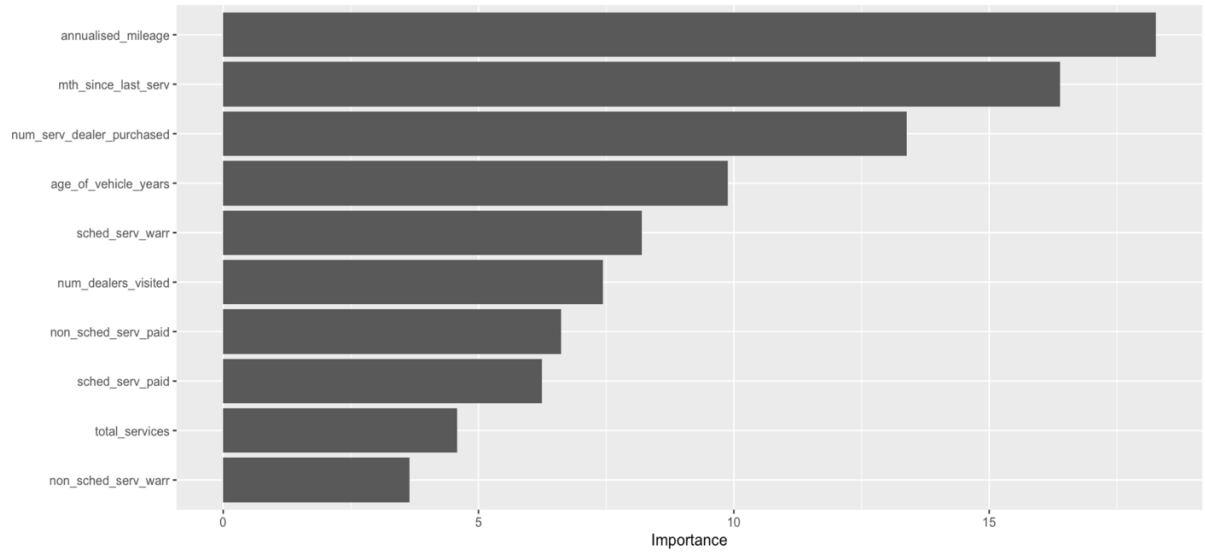
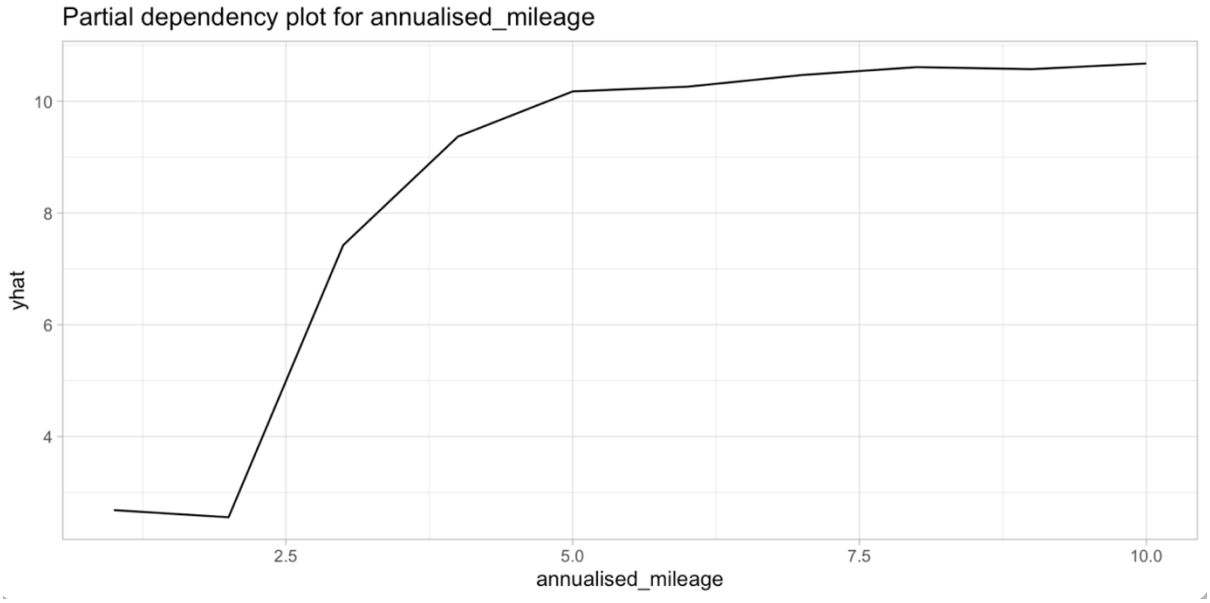


Figure 11 Plot variable importance with vip package

- Create PDP of top 5 predictors

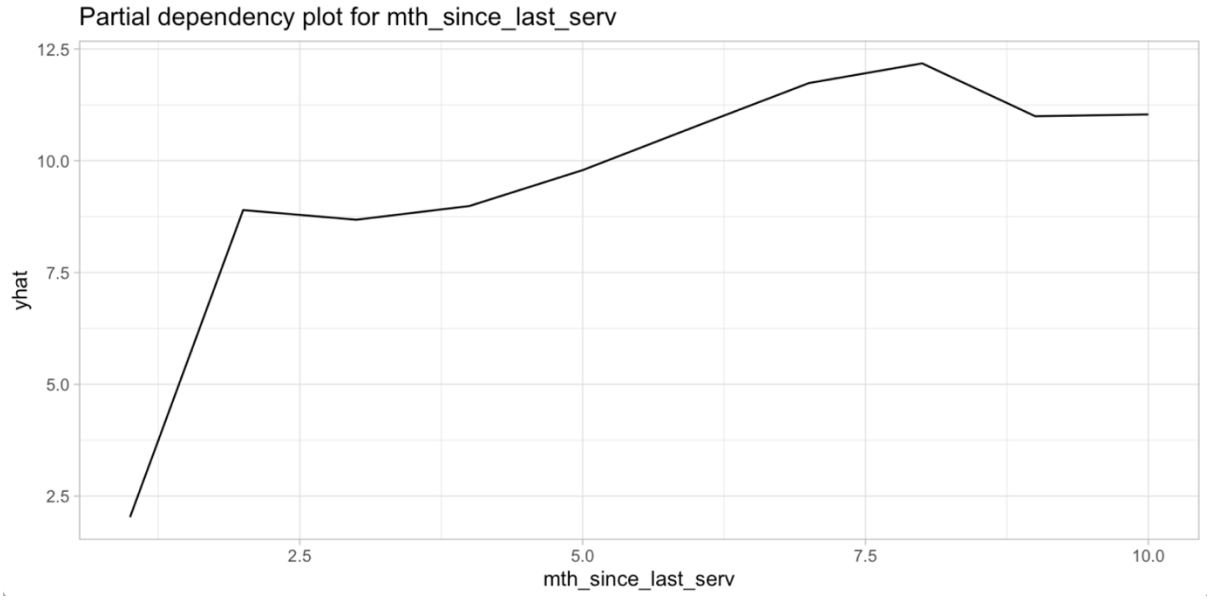
In this case, random forest model was used to predict the number validation set and use Partial Dependence Plots (PDP) to visualise the relationships the model has learned. PDP for predicting purchasing car and the top 5 predictors seen below:

1. Annualised mileage



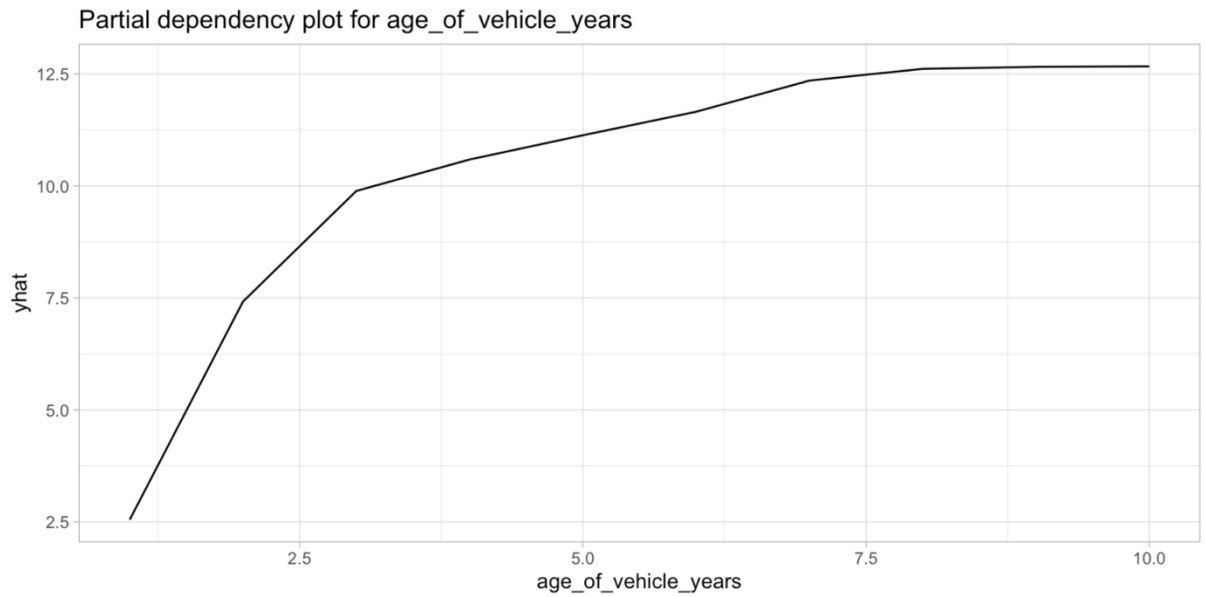
The trend line between the feature and target is positive. People who had more mileage on their car had better chances of purchasing a new vehicle.

2. Months since last service



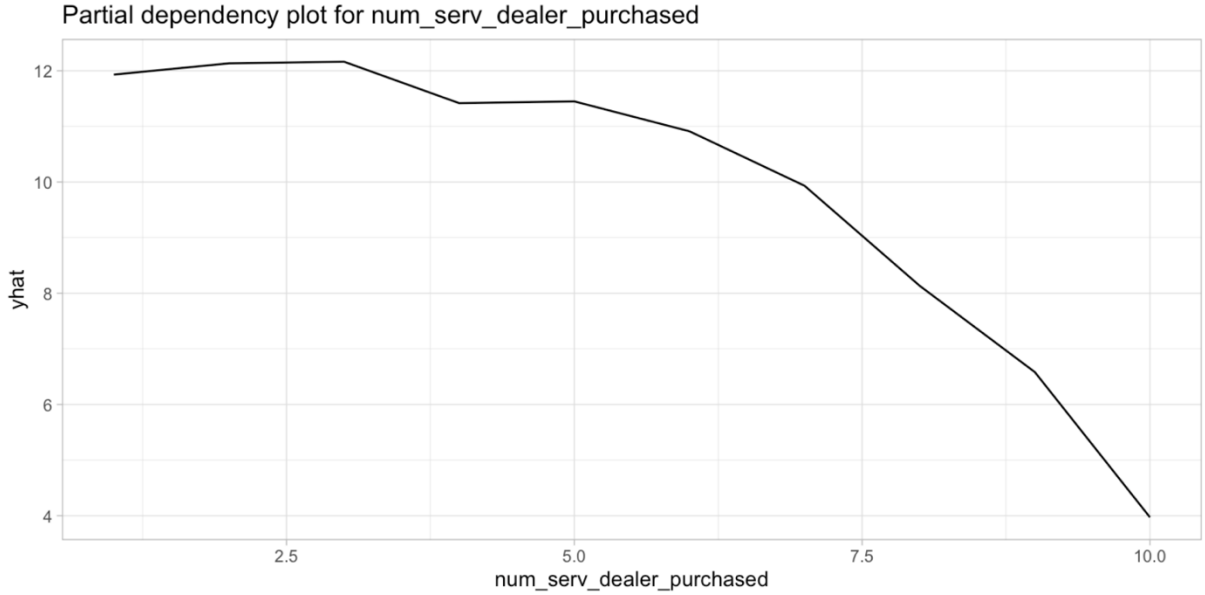
The longer since last service, the better chance the customer is more likely to repurchase a new vehicle.

3. Age of Vehicle in years



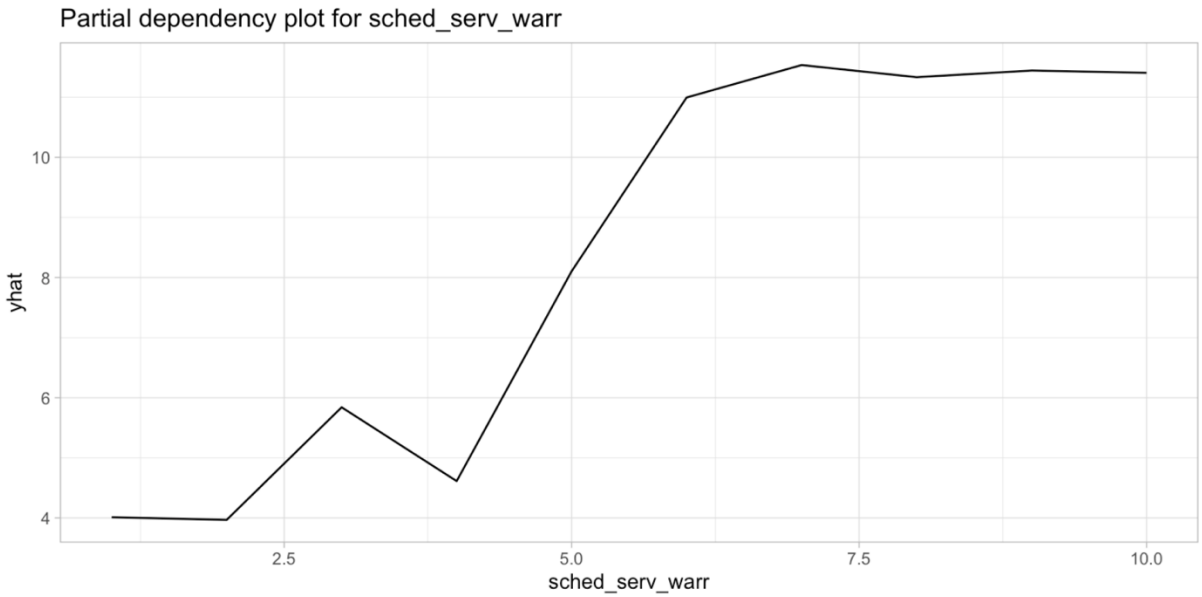
The older the vehicle, the better customer is more likely to repurchase a new vehicle. This may be because of vehicle wear and tear.

4. Number of services had at the same dealer where the vehicle was purchased



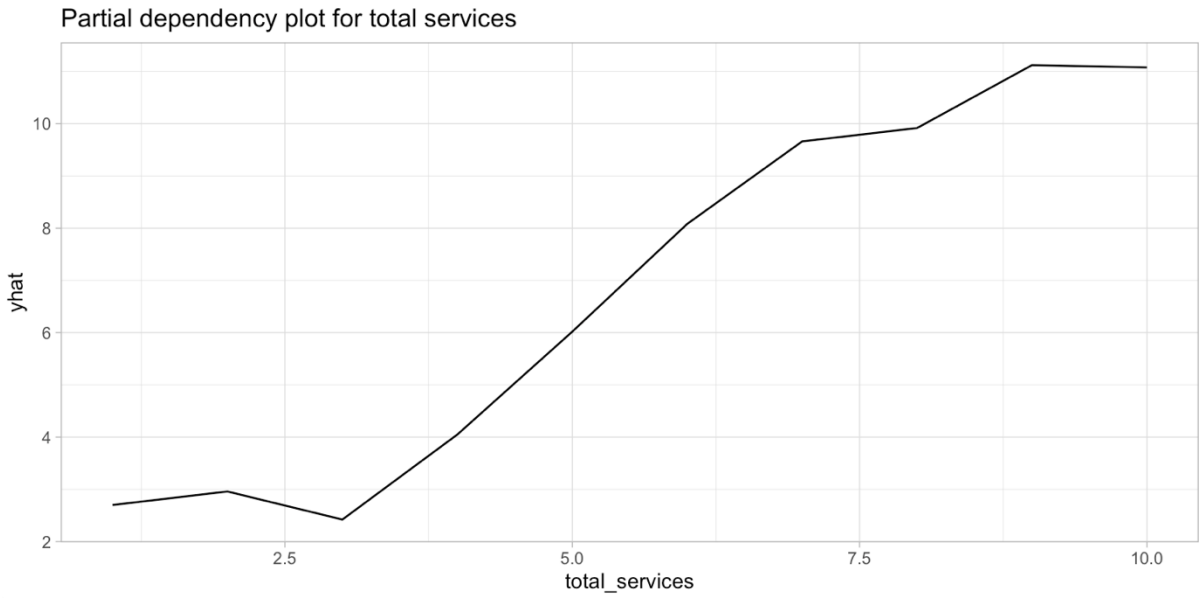
The trend line between the feature and target is negative. As number of services had at the same dealer where the vehicle was purchased increases, the customer is less likely to repurchase a new vehicle.

5. Number of schedules services used under warranty



As scheduled services used under warranty increase, the higher the chances of repurchasing a new vehicle. However in between 2.6-4 services, there seems to be a less likelihood of repurchasing. This reason can be further investigated below.

6. Lets do a 6th for total services



As total services increase, the higher the chances of repurchasing a new vehicle. This may be due to low ongoing maintenance costs.

<u>Random Forest model to predict validation set</u>

- Train set is the full repurchase training set after cleaning for the tree models
- Car model 19 was filtered from the training set because it doesn't exist in the validation data set
- Run the model (don't include ID, age or gender columns in the test set)
- Final confusion matrix

```
> final_confusion_matrix

        0      1
0  127653    738
1     161   2783
```

With the best model created the variables in 'repurchase_validation.csv' was used to output both probabilities and class predictions. This can be found in appendix.

5 Evaluation

Discussion of ethics, privacy issues and mitigation

The use of machine learning in marketing to influence decision making can have huge potential to profit a business, but also raise ethical. In this section, we discuss organisations such as automotive manufacturers and strategies and ethical and privacy issues of machine learning in marketing and research.

Informed consent to use:
Automotive manufacturers collect customer data by providing services. They can collect information such as their demographics, age of vehicle and servicing details. There is a need to examine under what circumstances the principles of informed consent should be deployed in the Artificial Intelligence space. Recent legislation concerning data protection has focused on protecting consumer data from unethical use. As AI becomes more prevalent, so with additional legislation to consumer privacy, data collection and algorithmic bias.

Algorithmic fairness and biases:
AI has the capability to improve the customer experience but just with any machine learning system, they will only be as trustworthy and fair as the data that it is trained with. Since algorithms behind AI are created by humans, they can be programmed with biases and can reinforce social and racial bias. It should be considered the risk for biases when deciding what ML technologies are used to train algorithms and datasets used for programming. Bias may occur with socio-economic status or age. This can be resolved through increased data availability and form minority populations and specify which populations the algorithm is or is not appropriately used.

6 Deployment

To conclude, the tree-based classification model can now be deployed for the sales team to view and understanding the parameters to help predict which customers are most likely to repurchase. They can increase their business potential by targeting customers with their vehicles using data above from top 5 predictors.

7 Appendix

The model generated above ‘repurchase_validation_12945604.csv’ contains three columns:

- ID
- target_probability
- target_class

repurchase_validation_12945604.csv