Product re-purchase prediction Report

Assessment task 1 - part B: classification modelling

Ganesh Arunagiri Rajan **13662388** **30/04/2020**

Business Understanding :

The client who is a manager at an Automotive manufacturing company plans to conduct a vehicle re-purchase campaign with an aim to send communication to customers who are most likely to purchase a new vehicle. He has approached us to help their company in predicting which of their existing customers would be most interested in purchasing a new vehicle.

Data Understanding :

The client has provided 2 datasets(Fig 1). The repurchase_training dataset contains the personal, vehicle and vehicle servicing details of 131337 unique customers. The data dictionary of the 17 different details recorded for each customer can be see in Fig 2.

Dataset name	Purpose
repurchase_training	To train, develop and evaluate all types of suitable models for the problem and choose the best performing model.
repurchase_validation	Real-time deployment of the chosen model on this dataset to predict the/identify the customers likely to purchase a new vehicle.

Fig 1: Datasets and their purpose

Among the 17 details recorded for each customer, 2 details(age_band and gender) had missing values in their columns(Fig 3). Approximately 85%(112375) of the values in age_band and 50%(69308) of the values in gender are missing(Fig 4 and Fig 5). On adding the missing values count from both columns, approximately 8% (181683) of the total data (131337 rows x 17 columns = 2232729) was observed to be missing; which is almost one-tenth of the entire dataset(Fig 3). As seen in fig 6, out of the entire set of total missing data(181683)

- 656 values were only missing in gender
- 43723 values were only missing in age_band
- 68652 values were missing in both age-band and gender

Field	Data Type	Description
ID	Unique ID	Unique ID of the customer
Target	Integer	Model target. 1 if the customer has purchased more than 1 vehicle, 0 if they have only purchased 1.
age_band	Categorical	Age banded into categories
gender	Categorical	Male, Female or Missing
car_model	Categorical	The model of vehicle, 18 models in total
car_segment	Categorical	The type of vehicle
age_of_vehicle_years	Integer	Age of their last vehicle
sched_serv_warr	Integer	Number of scheduled services (e.g. regular check-ups) used under warranty
non_sched_serv_warr	Integer	Number of non-scheduled services (e.g. something broke out of the service cycle) used under warranty
sched_serv_paid	Integer	Amount paid for scheduled services
non_sched_serv_paid	Integer	Amount paid for non scheduled services
total_paid_services	Integer	Amount paid in total for services
total_services	Integer	Total number of services
mth_since_last_serv	Integer	The number of months since the last service
annualised_mileage	Integer	Annualised vehicle mileage
num_dealers_visited	Integer	Number of different dealers visited for servicing
num_serv_dealer_purchased	Integer	Number of services had at the same dealer where the vehicle was purchased

Fig 2 : Data dictionary of details recorded for each customer

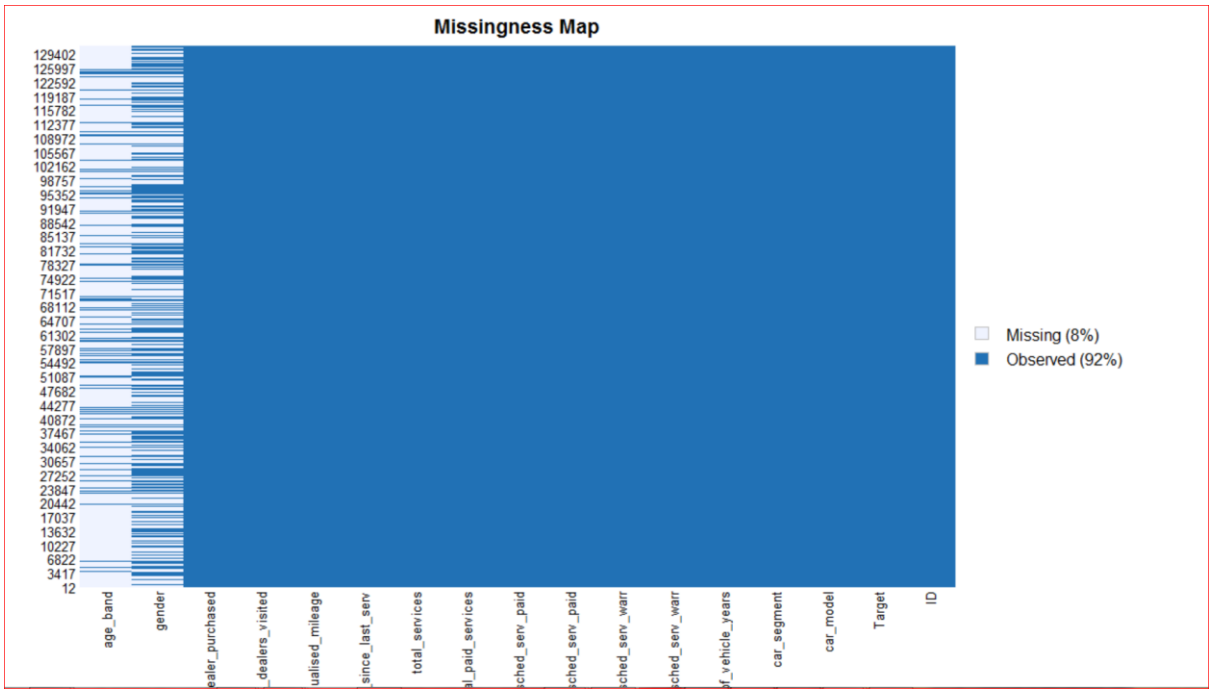


Fig 3 : Missingness map showing which details had missing values in their columns

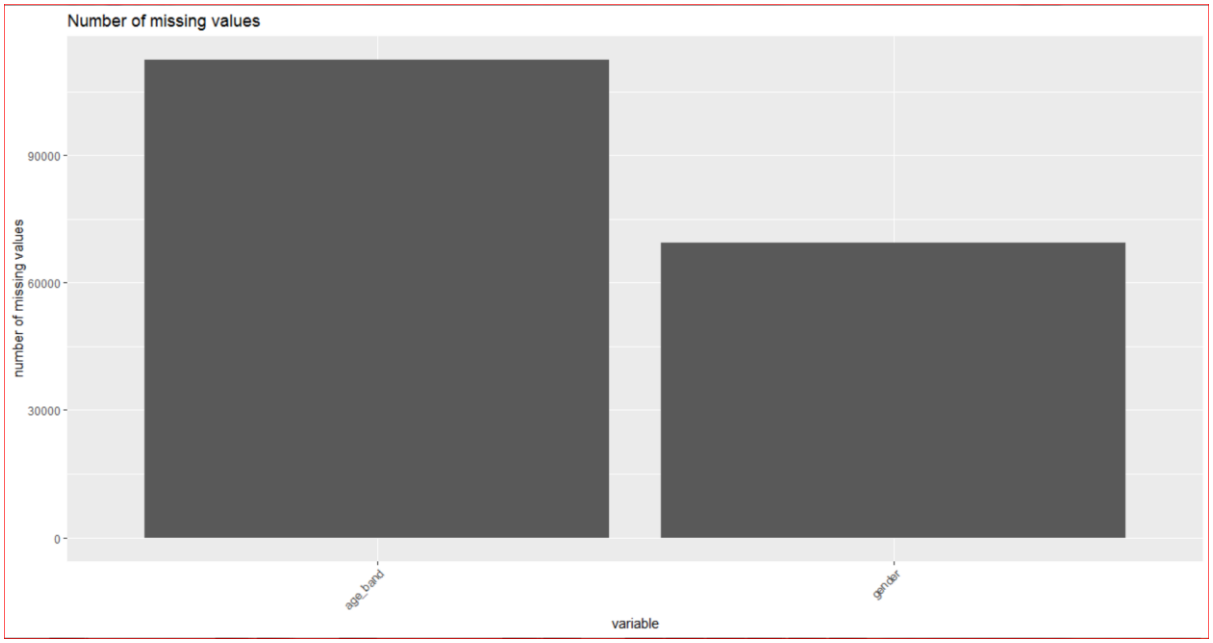


Fig 4 : Number of missing values in gender and age_band

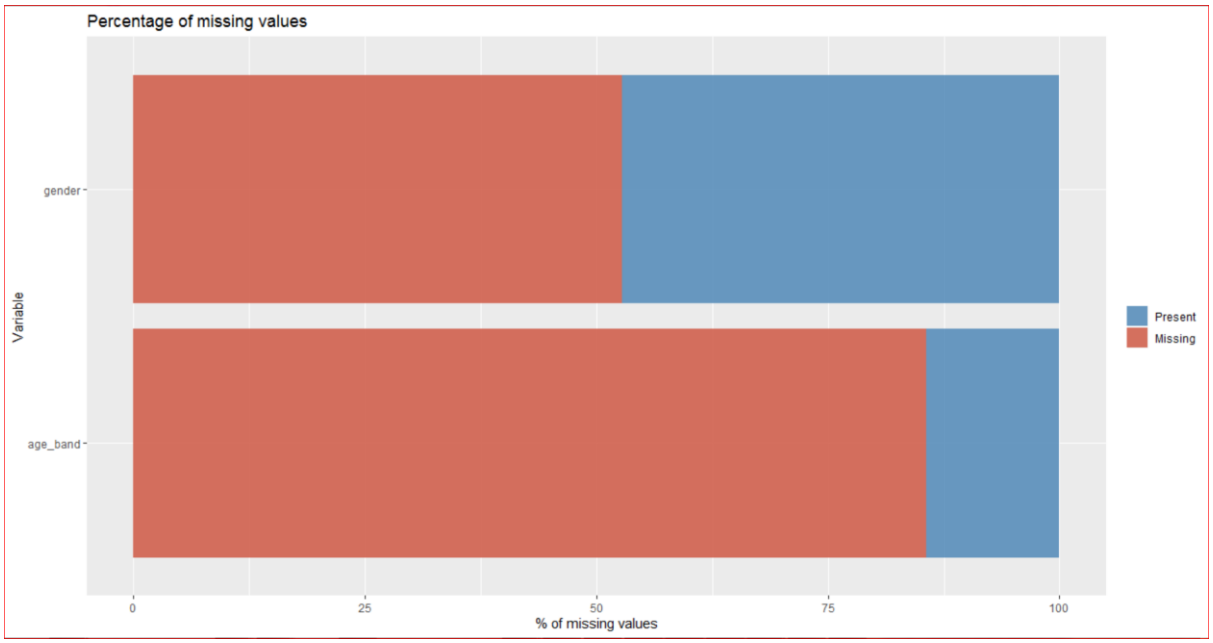


Fig 5 : Percentage of missing values in gender and age_band

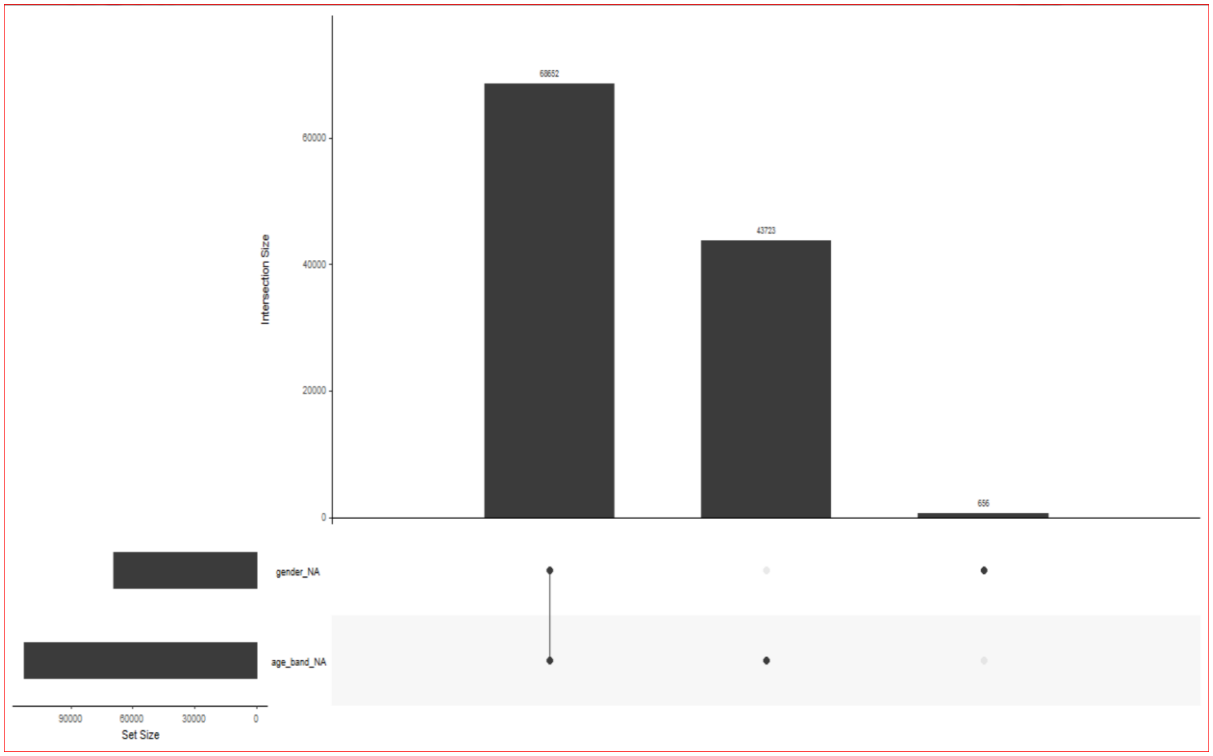


Fig 6 : Intersection set of total missing data

Approximately 97%(127816) of the total observations(131337) belong to Target class 0 and only a mere 3%(3521) belong to Target class 1(Fig 7). This illustrates that there is an imbalance in the Target classes.

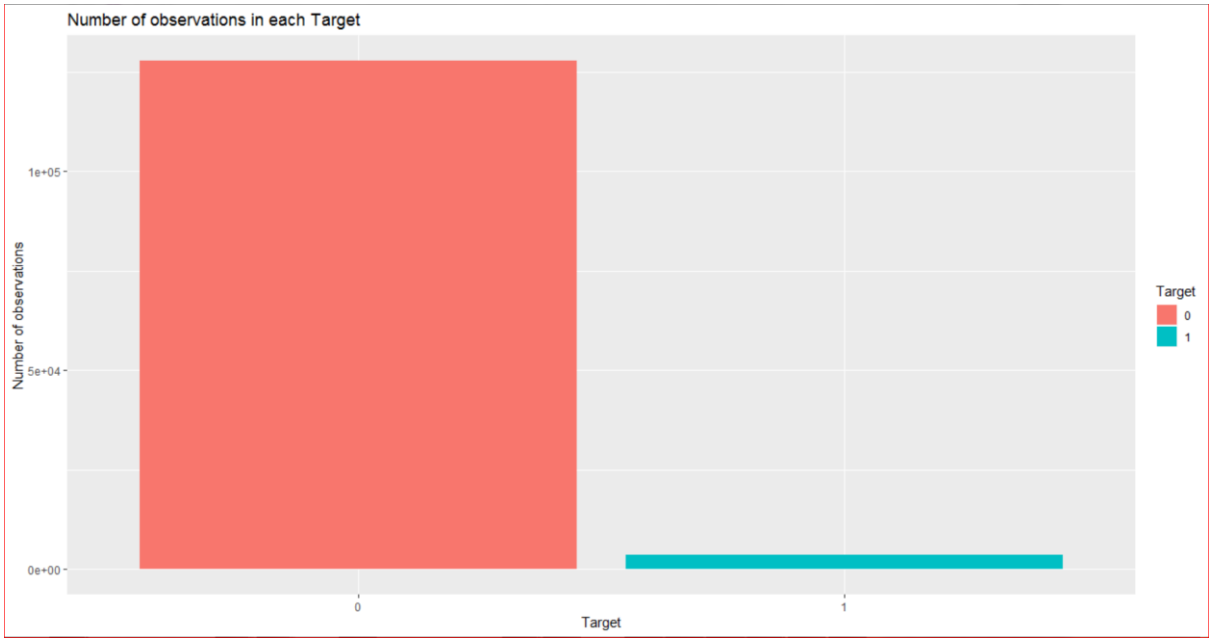


Fig 7 : Number of observations in each Target class

Data Preparation :

There was a lot of noise observed in the columns containing numeric variables(Fig 8). Hence the numeric variables in those columns were converted into decile Integer value range from 1 to 10 with equal number of customers grouped for each Decile value.

Field
age_of_vehicle_years
sched_serv_warr
non_sched_serv_warr
sched_serv_paid
non_sched_serv_paid
total_paid_services
total_services
mth_since_last_serv
annualised_mileage
num_dealers_visited
num_serv_dealer_purchased

Fig 8 : List of columns which were converted from numeric to deciles(Integers), in order to reduce noise

Moreover, the Target field and other fields containing character/categorical values were made as a factor for modelling purposes(Fig 9).

Field	Data Type	Format Conversion
Target	Integer	Made as a Factor
age_band	Categorical	Made as a Factor
gender	Categorical	Made as a Factor
car_model	Categorical	Made as a Factor
car_segment	Categorical	Made as a Factor

Fig 9 : Format conversion of target field and fields containing character values

The ID field was removed from the dataset as it does not add any value to the dataset and the model.

Modelling :

The dataset was further split into train and test set for the purpose of modelling. The dataset was split randomly into train and test sets, with a proportion of 75%(98502 observations) data for training and 25%(32835 observations) data for testing.

Considering that the Business problem requires us to classify each customer into either Target class 0 or 1, Classification based Machine Learning models(Linear classification model & Tree based model) would be ideal to solve the problem as the Target variable although being a numeric is non-continuous, treated as a factor and contains only 2 values(0 and 1). A model each was developed in both the Linear classification and Tree based model types(Fig 10).

Model Class	Chosen Model
Linear Classification	Logistic Regression
Tree Based	Boosting (GBM)

Fig 10 : Model developed for each classification model class type

Class imbalance was also observed in the Training data as approximately 97.3%(95887 observations) were Target class 0 and only 2.7%(2615 observations) were Target class 1(Fig 11).

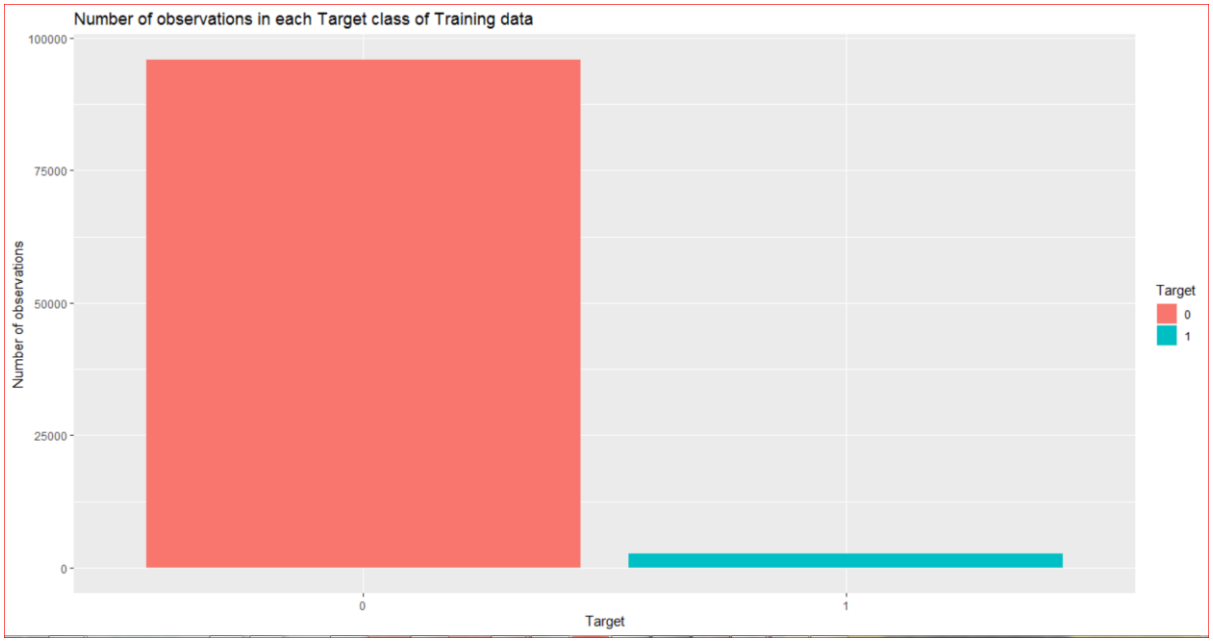


Fig 11 : Number of observations in each Target class of Training data

With Imbalanced Class in the training data, Classification algorithms wouldn't get sufficient information about the minority class(Target 1) to make an accurate prediction thus resulting in biased prediction towards the majority class(Target 0). In order to rectify the class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied on the training data where it oversamples the rare event(Target 1) by utilising k-nearest neighbours and bootstrapping to synthetically create extra observations for that event.

After applying SMOTE, the Target class observations in the training data was balanced(Fig 12 and Fig 13).

Fig 12 : Graph of Number of observations in each Target class of Training data after SMOTE

Target Class	Number of observations after SMOTE	Percentage of observations after SMOTE
0	5230	50%
1	5230	50%

Fig 13: Number and percentage of observations in each Target class of Training data after SMOTE

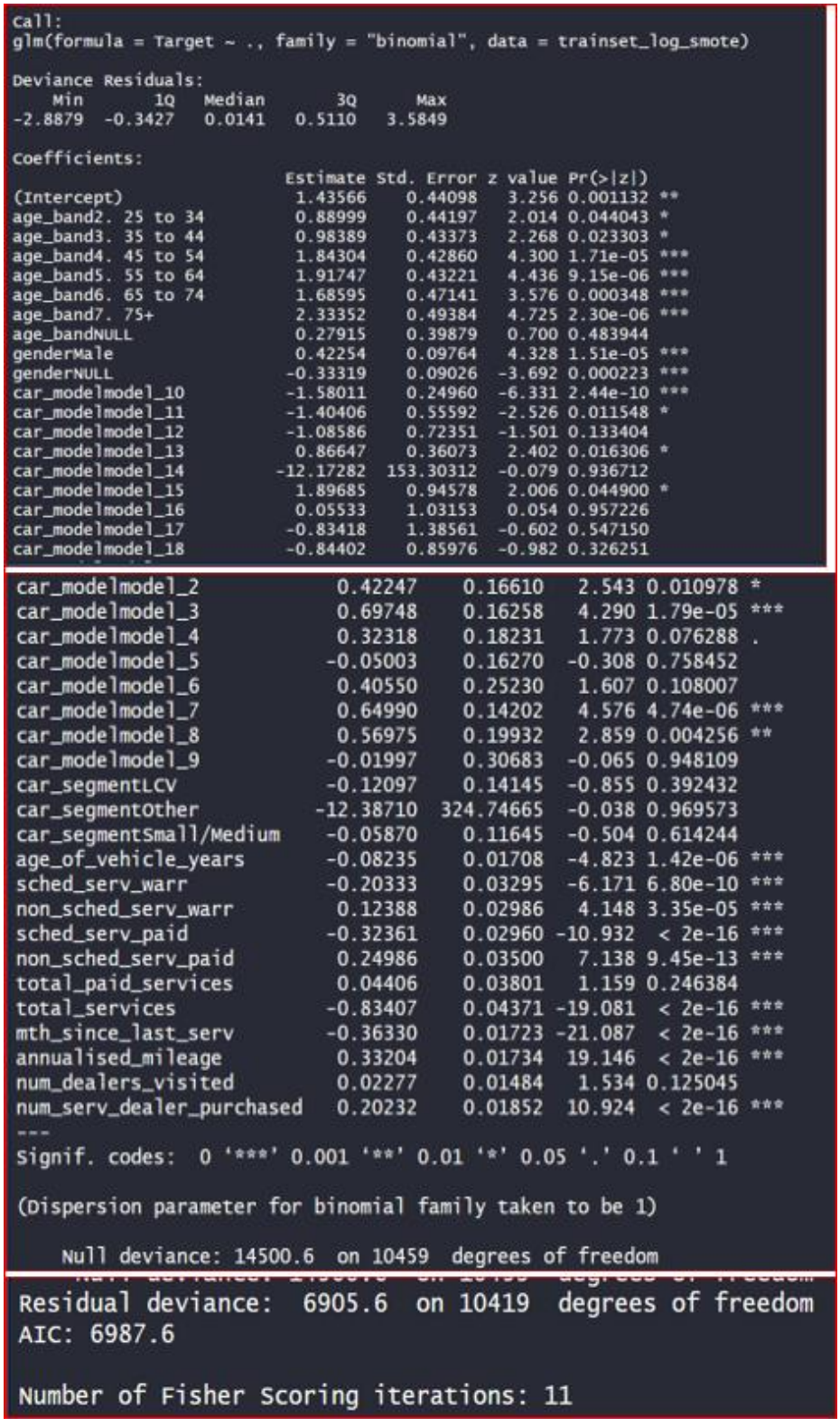
The Target and predictor variables selected for both the models can be found in Fig 14.

Field	Variable Type
Target	Target
age_band	Predictor
gender	Predictor
car_model	Predictor
car_segment	Predictor
age_of_vehicle_years	Predictor
sched_serv_warr	Predictor
non_sched_serv_warr	Predictor
sched_serv_paid	Predictor
non_sched_serv_paid	Predictor
total_paid_services	Predictor
total_services	Predictor
mth_since_last_serv	Predictor
annualised_mileage	Predictor
num_dealers_visited	Predictor
num_serv_dealer_purchased	Predictor

Fig 14 : Target and Predictor variables

Logistic Regression :

The logistic regression model was trained on the SMOTE training data and the AIC was 6987.6(Fig 15).



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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8879  -0.3427   0.0141   0.5110   3.5849

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.43566    0.44098   3.256 0.001132 **
age_band2. 25 to 34       0.88999    0.44197   2.014 0.044043 *
age_band3. 35 to 44       0.98389    0.43373   2.268 0.023303 *
age_band4. 45 to 54       1.84304    0.42860   4.300 1.71e-05 ***
age_band5. 55 to 64       1.91747    0.43221   4.436 9.15e-06 ***
age_band6. 65 to 74       1.68595    0.47141   3.576 0.000348 ***
age_band7. 75+            2.33352    0.49384   4.725 2.30e-06 ***
age_bandNULL              0.27915    0.39879   0.700 0.483944
genderMale                0.42254    0.09764   4.328 1.51e-05 ***
genderNULL               -0.33319    0.09026  -3.692 0.000223 ***
car_modelmodel_10        -1.58011    0.24960  -6.331 2.44e-10 ***
car_modelmodel_11        -1.40406    0.55592  -2.526 0.011548 *
car_modelmodel_12        -1.08586    0.72351  -1.501 0.133404
car_modelmodel_13         0.86647    0.36073   2.402 0.016306 *
car_modelmodel_14       -12.17282  153.30312  -0.079 0.936712
car_modelmodel_15         1.89685    0.94578   2.006 0.044900 *
car_modelmodel_16         0.05533    1.03153   0.054 0.957226
car_modelmodel_17        -0.83418    1.38561  -0.602 0.547150
car_modelmodel_18        -0.84402    0.85976  -0.982 0.326251
car_modelmodel_2          0.42247    0.16610   2.543 0.010978 *
car_modelmodel_3          0.69748    0.16258   4.290 1.79e-05 ***
car_modelmodel_4          0.32318    0.18231   1.773 0.076288 .
car_modelmodel_5         -0.05003    0.16270  -0.308 0.758452
car_modelmodel_6          0.40550    0.25230   1.607 0.108007
car_modelmodel_7          0.64990    0.14202   4.576 4.74e-06 ***
car_modelmodel_8          0.56975    0.19932   2.859 0.004256 **
car_modelmodel_9         -0.01997    0.30683  -0.065 0.948109
car_segmentLCV           -0.12097    0.14145  -0.855 0.392432
car_segmentOther        -12.38710  324.74665  -0.038 0.969573
car_segmentSmall/Medium   -0.05870    0.11645  -0.504 0.614244
age_of_vehicle_years      -0.08235    0.01708  -4.823 1.42e-06 ***
sched_serv_warr           -0.20333    0.03295  -6.171 6.80e-10 ***
non_sched_serv_warr        0.12388    0.02986   4.148 3.35e-05 ***
sched_serv_paid           -0.32361    0.02960 -10.932  < 2e-16 ***
non_sched_serv_paid        0.24986    0.03500   7.138 9.45e-13 ***
total_paid_services        0.04406    0.03801   1.159 0.246384
total_services            -0.83407    0.04371 -19.081  < 2e-16 ***
mth_since_last_serv       -0.36330    0.01723 -21.087  < 2e-16 ***
annualised_mileage         0.33204    0.01734  19.146  < 2e-16 ***
num_dealers_visited        0.02277    0.01484   1.534 0.125045
num_serv_dealer_purchased  0.20232    0.01852  10.924  < 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: 14500.6  on 10459  degrees of freedom
Residual deviance:  6905.6  on 10419  degrees of freedom
AIC: 6987.6

Number of Fisher Scoring iterations: 11
```

Fig 15 : Model statistics of the Logistic Regression model trained on the SMOTE training data.

The trained logistic regression model was tested by making predictions on the testing data and the confusion matrix was created (Fig 16).

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 25602   153
         1  6327   753

               Accuracy : 0.8026
                 95% CI : (0.7983, 0.8069)
    No Information Rate : 0.9724
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1468

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.83113
            Specificity : 0.80184
         Pos Pred Value : 0.10636
         Neg Pred Value : 0.99406
             Prevalence : 0.02759
         Detection Rate : 0.02293
   Detection Prevalence : 0.21562
      Balanced Accuracy : 0.81648

       'Positive' Class : 1

>
```

Fig 16 : Confusion Matrix and statistics of the logistic Regression model

The Recall, Precision,F1 and AUC metrics were calculated from the confusion matrix and ROC curves for Train and test was plotted(Fig 17 & 18).

Metric	Meaning	Formula	Value
Precision	Measure of correctness achieved in Positive prediction	Precision = TP / (TP + FP)	0.1063559
Recall	Measure of actual observations which are predicted correctly	Recall = TP / (TP + FN)	0.8311258
F1	Measure of effectiveness of classification in terms of ratio of weighted importance on either recall or precision	F1 = (2 x Recall x Precision) / (Recall + Precision)	0.18858
Train AUC	Measure of the entire 2-Dimensional area underneath the ROC curve		0.9307059
Test AUC	Measure of the entire 2-Dimensional area underneath the ROC curve		0.9104673

Fig 17: Precision, Recall,F1,Train and Test AUC values of the Logistic Regression model.

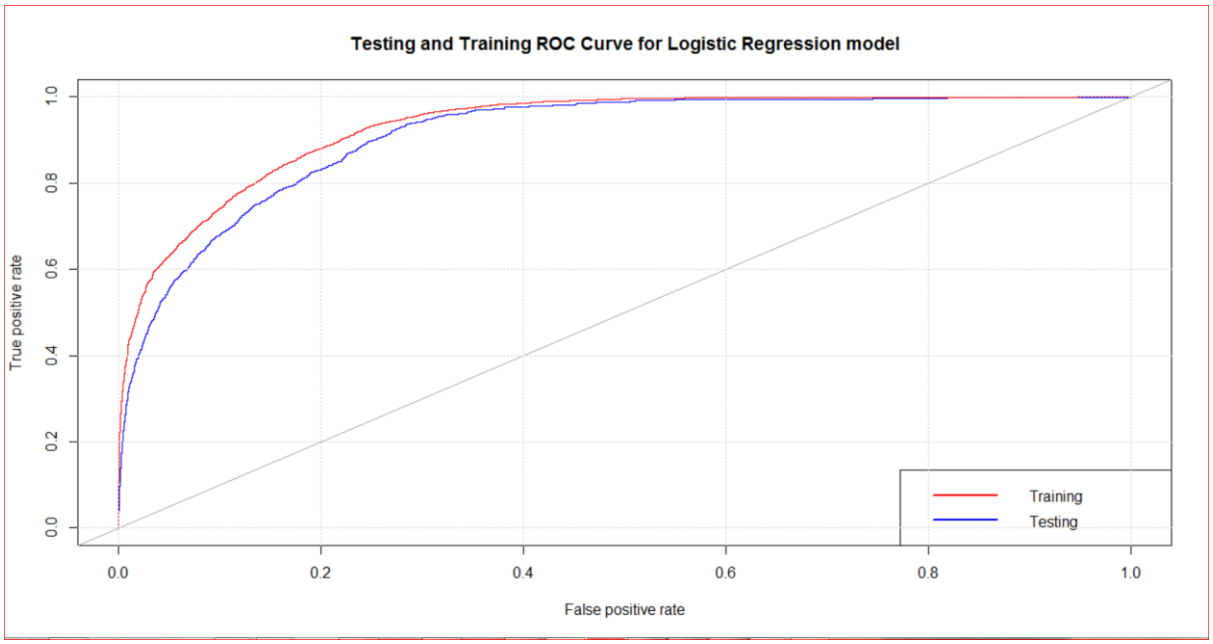


Fig 18 : Testing and Training ROC curves for logistic Regression model.

Fig 19 shows the 10 variables in the logistic regression model with high variable Importance. mth_since_last_serve is the most important variable.

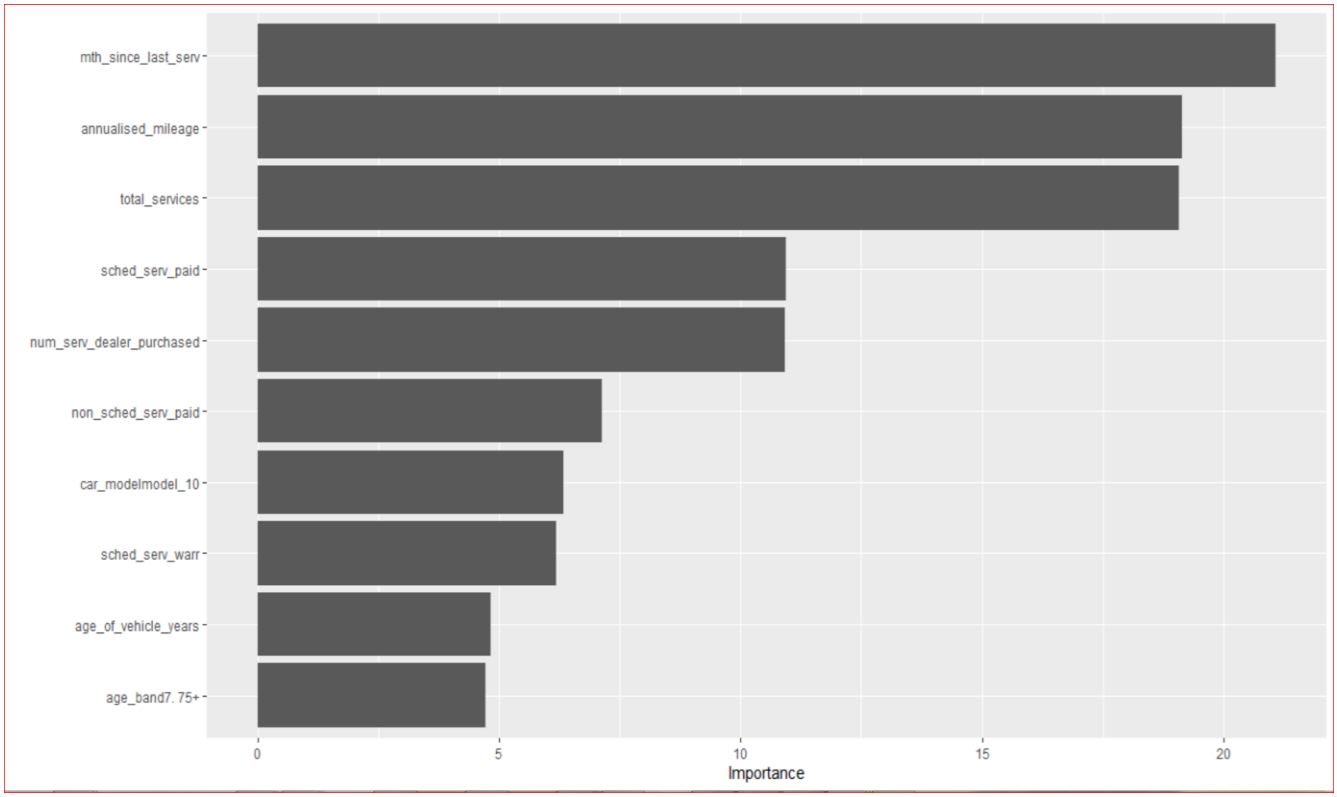


Fig 19 : Variable Importance plot for Logistic regression model

Boosting(GBM) :

Once the hyper-parameter values mentioned in fig 20 were set; the GBM model was trained on the SMOTE training data and variable importance of the predictor variables were calculated(Fig 21).

Hyper-parameter	Meaning	Value
gbm_depth	maximum nodes per tree	20
gbm_n_min	minimum number of observations in the trees terminal	15
gbm_shrinkage	learning rate	0.01
gbm_cv_folds	number of cross-validation folds to perform	70
num_trees	Number of iterations	700

Fig 20 : Hyper-parameter values of GBM model

```
                                                  var     rel.inf
sched_serv_warr                       sched_serv_warr 35.49140773
mth_since_last_serv                mth_since_last_serv 21.16075085
annualised_mileage                  annualised_mileage 11.64720297
sched_serv_paid                        sched_serv_paid  6.40019226
total_services                          total_services  5.66347856
num_serv_dealer_purchased num_serv_dealer_purchased  4.36687955
age_of_vehicle_years            age_of_vehicle_years  4.03474206
non_sched_serv_warr              non_sched_serv_warr  3.14573195
num_dealers_visited              num_dealers_visited  2.30289364
total_paid_services              total_paid_services  2.13984926
gender                                        gender  1.48801876
car_model                                  car_model  1.26844067
non_sched_serv_paid           non_sched_serv_paid  0.68398216
age_band                                    age_band  0.19074403
car_segment                              car_segment  0.01568554
>
```

Fig 21: Variable importance table for GBM

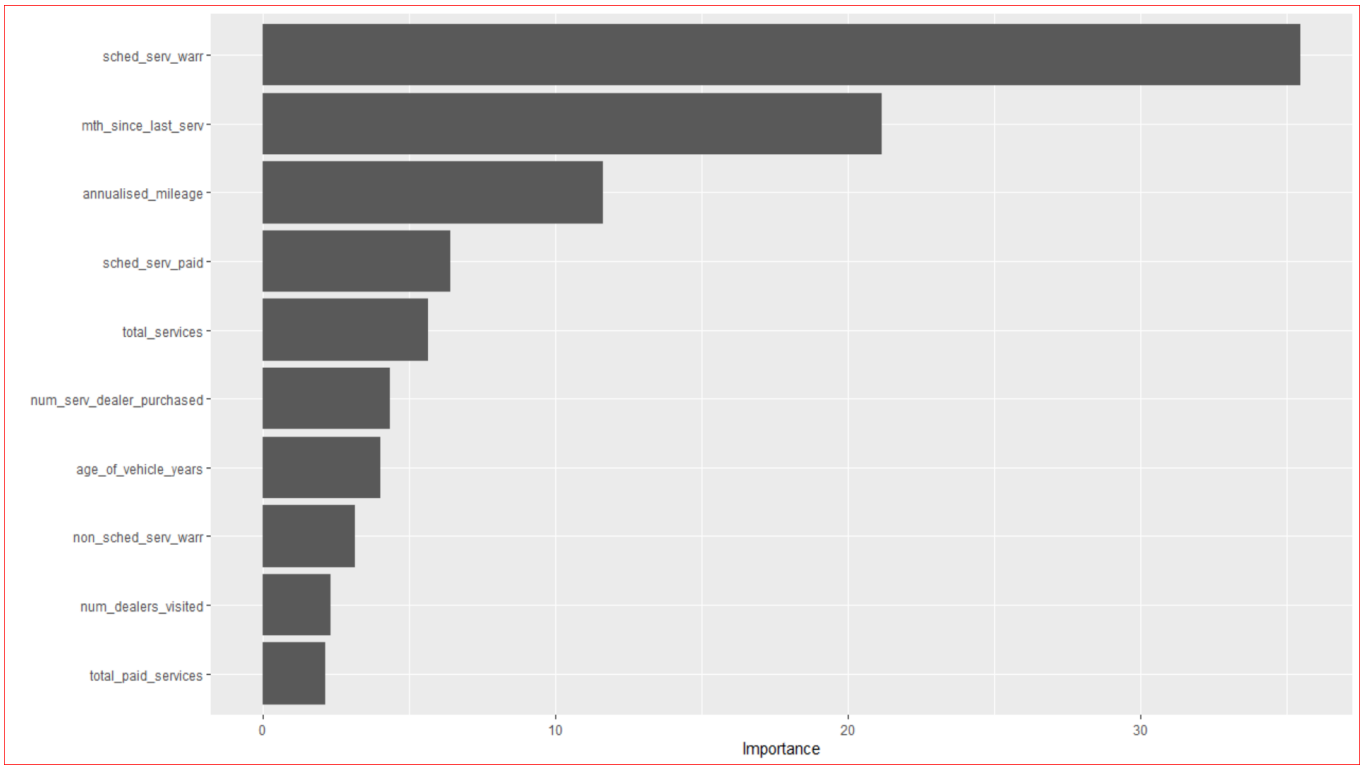


Fig 22 : Variable importance graph for GBM

In logistic regression, the variables with least Pr(>|z|) value less than or equal to 0.05 are more statistically significant than other variables and hence have high variable importance. However in GBM the variables with highest relative influence are regarded as most important. The variables have differing importance in both the models as both the models learn and interpret the variables differently; the logistic regression tries to create a linear relationship with the variable whereas the GBM tries to classify based on the variable value range.

The trained GBM was tested by making predictions on the testing data and the confusion matrix was created (Fig 23).

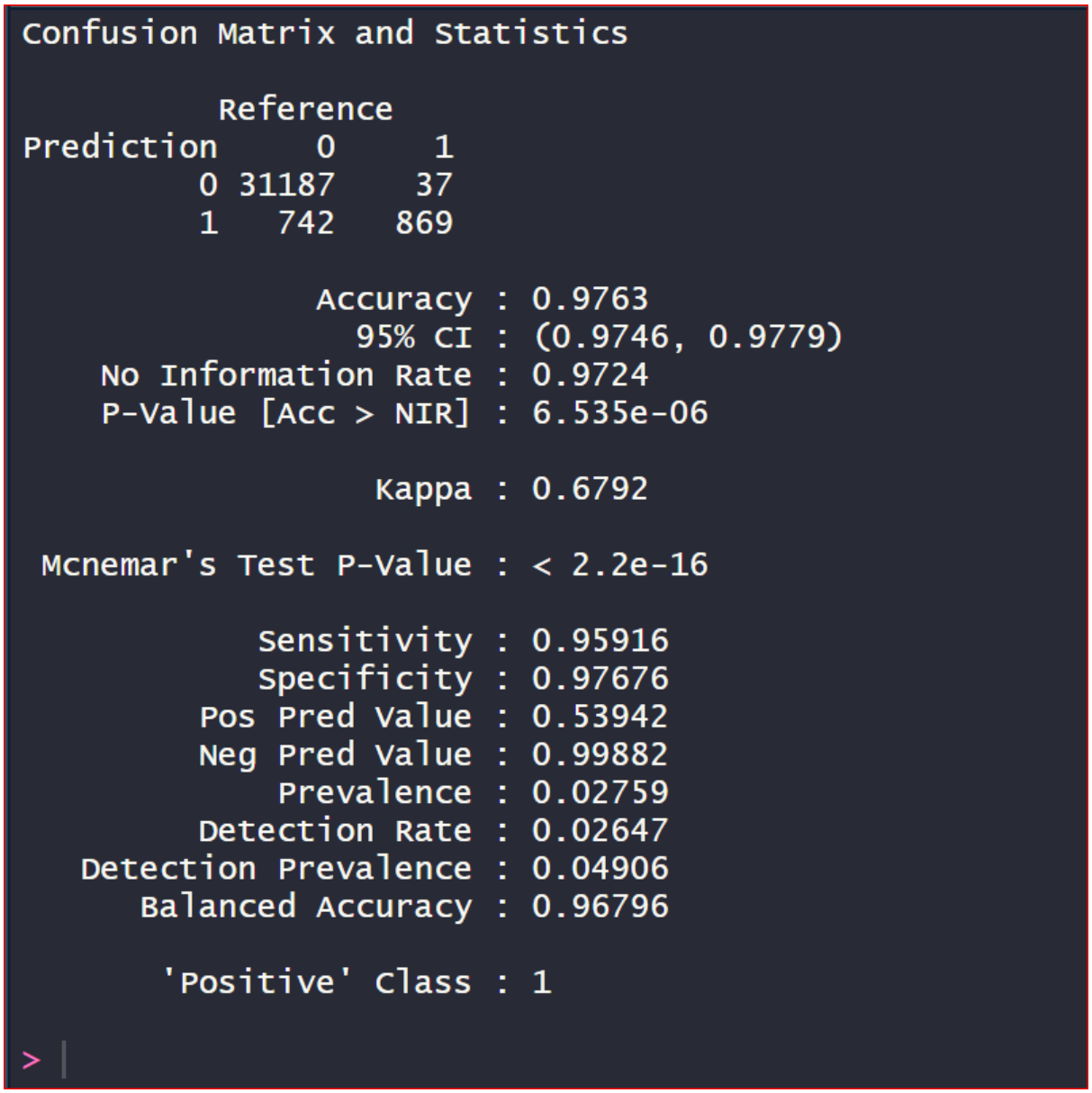


Fig 23 : Confusion Matrix and statistics of GBM

The Recall, Precision,F1 and AUC metrics for GBM were calculated from the confusion matrix and ROC curves for Train and test was plotted(Fig 24 and Fig 25).

Metric	Meaning	Formula	Value
Precision	Measure of correctness achieved in Positive prediction	Precision = TP / (TP + FP)	0.5394165
Recall	Measure of actual observations which are predicted correctly	Recall = TP / (TP + FN)	0.9591611
F1	Measure of effectiveness of classification in terms of ratio of weighted importance on either recall or precision	F1 = (2 × Recall × Precision) / (Recall + Precision)	0.6905046
Train AUC	Measure of the entire 2-Dimensional area underneath the ROC curve		0.9991321
Test AUC	Measure of the entire 2-Dimensional area underneath the ROC curve		0.9959278

Fig 24: Precision, Recall,F1,Train and Test AUC values of GBM

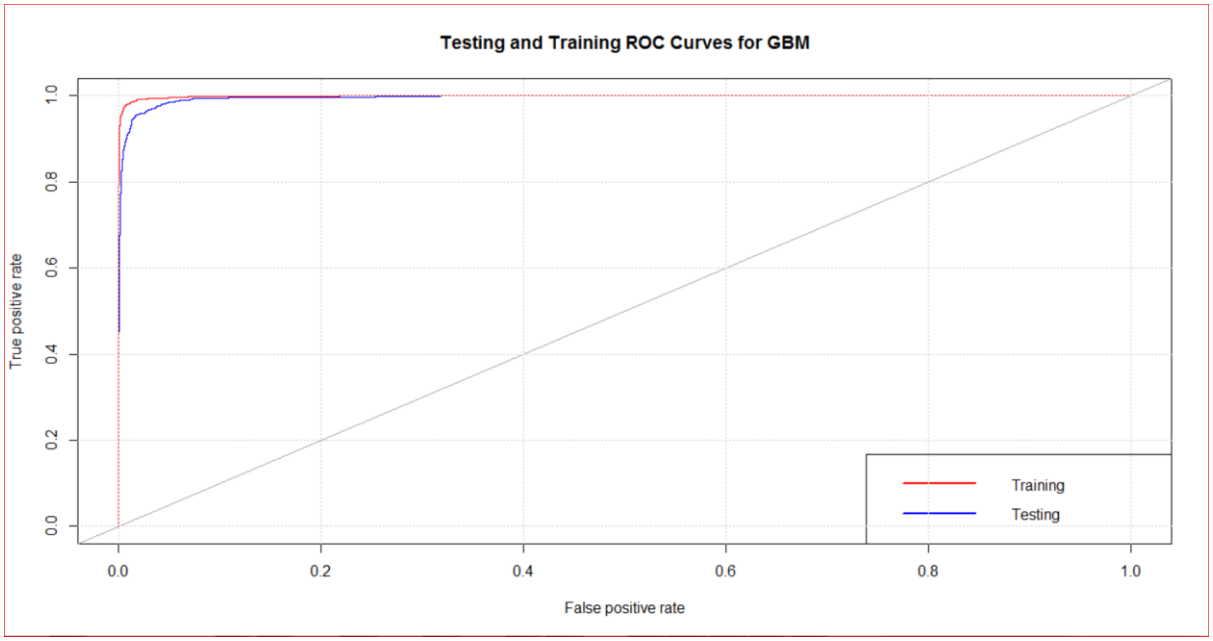


Fig 25 : Testing and Training ROC curves for GBM

Fig 26,27 and 28 shows the partial dependency plots of the top 5 most important variables; where it is interesting to note that for all 5, the probability of predicting Target 1 decreases with an increase in the value of the variable.

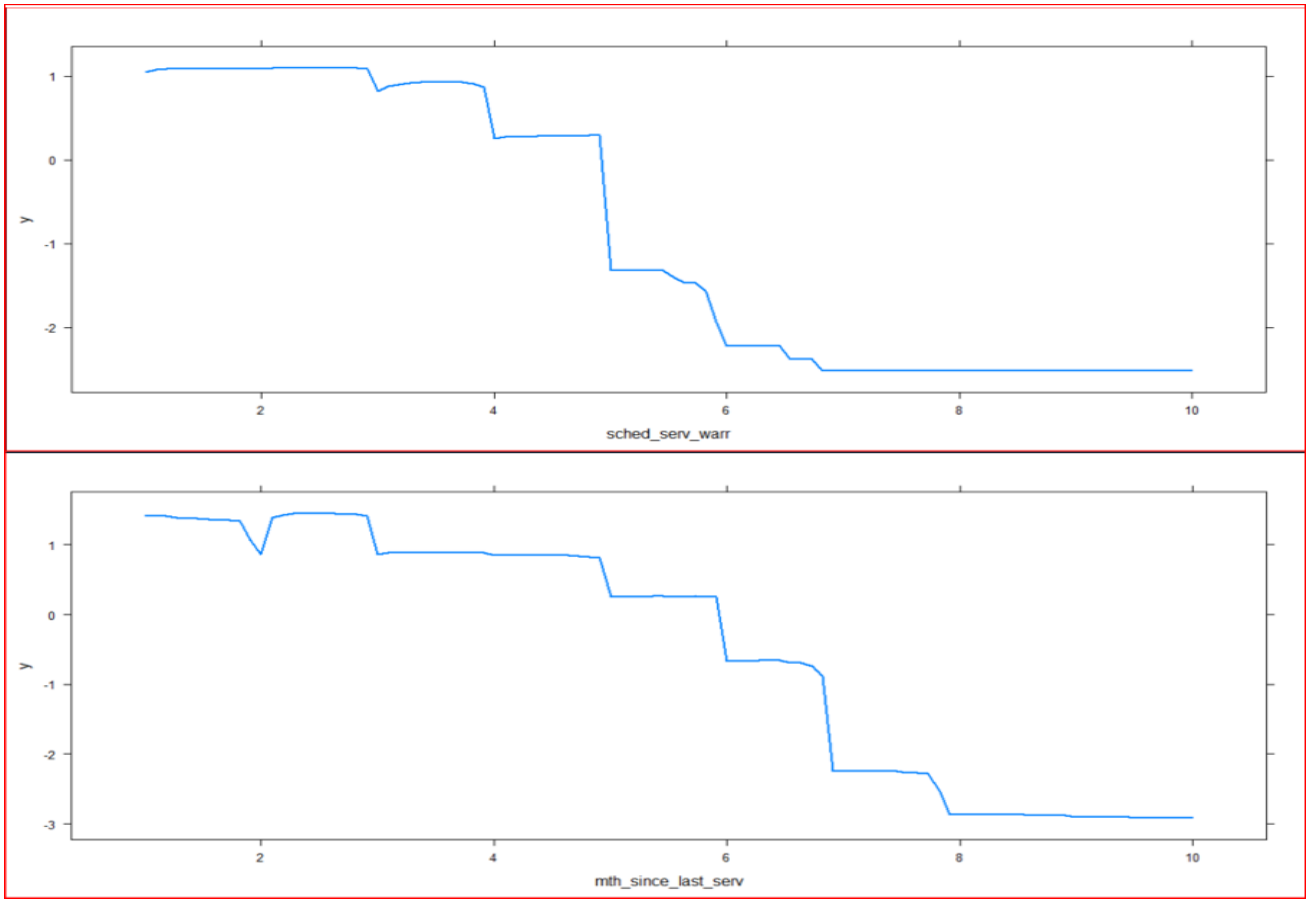


Fig 26 : Partial dependency plots for sched_serv_warr and mth_since_last_serv variables

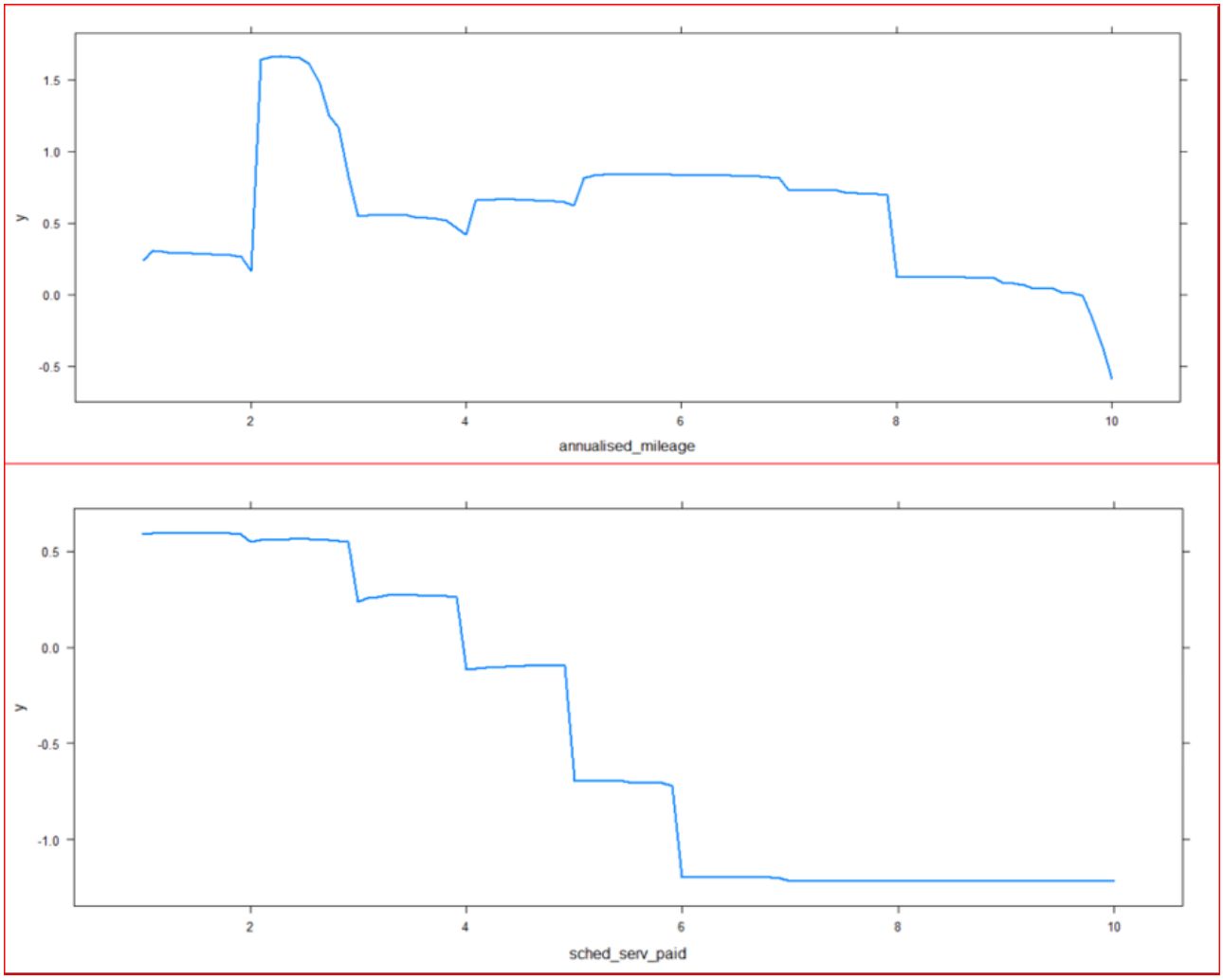


Fig 27 : Partial dependency plots for annualised_mileage and sched_serv_paid variables

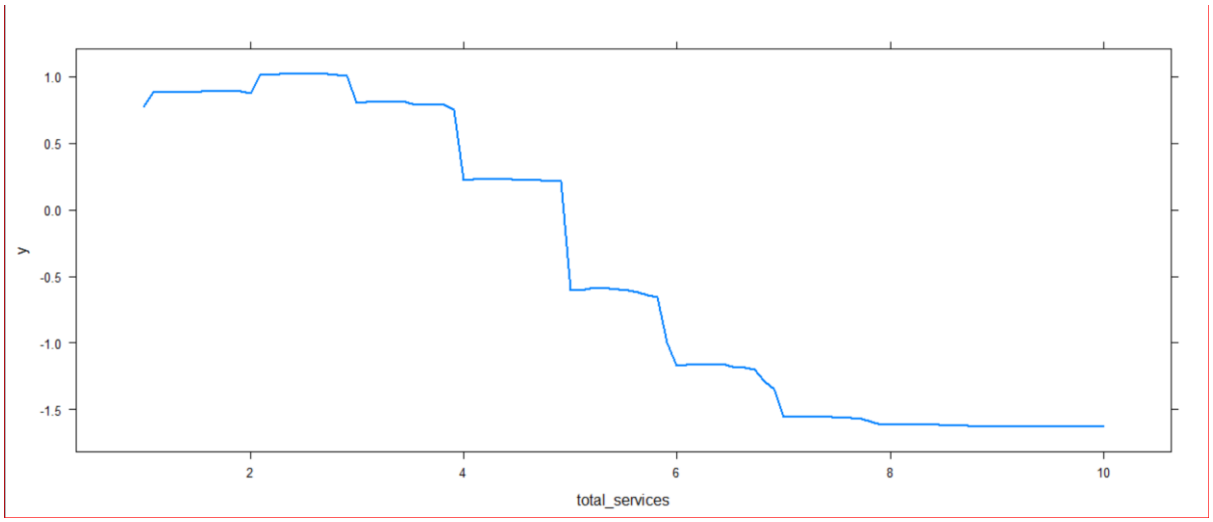


Fig 28 : Partial dependency plots for total_services

Evaluation :

In this Business problem, the weightage of predicting correctly the customers likely to repurchase vehicle is more important as the automotive company has to send communication to all the possible repurchase customers. Hence we are more interested in measuring the number of observations of positive class predicted correctly, which also known as Recall or Sensitivity and the model with the highest value of Recall would be the most appropriate for the problem. GBM has the highest Recall values, hence it is the best model for prediction in this problem.

Model	Recall Value
Logistic Regression	0.8311258
GBM	0.9591611

Fig 29: Recall values of both the models

Deployment :

The GBM model was used to predict the Target class for the repurchase_validation dataset and the percentage of predictions for each Target class was calculated.

Target	Observations	Percentage
0	45904	91.9
1	4096	8.1

Fig 30 : Total Number and percentage of observations predicted for each Target class