Machine Learning Algorithms and Applications

Assessment 2 | Part D

Nathan Collins

12062131

Assessment 2:	Classification Models
Type:	Individual Assessment
Deliverables:	Jupyter Notebook (x5) Experiment Report (x5) Final Report ~1500 words
Weight:	100 pts
Due:	Friday, 28 April, 23:59

Assessment Criteria:

- *The soundness of justification for the selected technique*
- *Quality of code and visualisations*
- *Accuracy of results and evidence supporting claims*
- *The breadth of evidence of collaborative work (e.g. meeting minutes, details of contributions etc.)*
- *Criticality and specificity in evaluating assumptions and potential ethical issues*
- *Appropriateness of communication style to the audience*

Section 1: Business Understanding

[1.1] Objective, Situation, Data Mining Goals

The aim of this project was to utilise analytical and statistical methods to determine the likelihood that an existing customer of an automotive manufacturer will purchase a new vehicle. The success of the model will empower business stakeholders to implement a cost-effective re-purchase campaign targeting existing customers who are prospective in purchasing a new or secondary vehicle.

As the business seeks a novel and actionable solution, the data mining objective is to identify a set of features that correspond to a class of individuals within a provided historical dataset of existing customers. Classification of the customers who have purchased a second vehicle can ultimately be applied to prospective customers.

[1.2] Project Plan

1. Data Collection: Provided by the automotive dealer.

2. Data Cleaning and Pre-processing: Data was cleaned for suitability prior to analysis. This included retaining two data frames, one including the age bands and genders of customers, the other without. For each data frame, missing values were dropped, categorical variables were re-engineered to integers, NaN values were removed, and the ID column was removed.

3. Feature Selection: All features were incorporated within all models.

4. Model Selection and Instantiation: Modelling choices: Univariate Logistic Regression (ULR), Multivariate Logistic Regression (MLR), Support Vector Classification (SVC), Random Forest Classification (RFC), and XGBoosting Classification (XGB).

5. Model Evaluation: A confusion matrix and precision-recall curve served as the key performance metrics throughout each model.

[1.3] Ethical Considerations

Ethical and privacy implications can occur from misused data. While the dataset provided retains limited personal information about individuals (age and gender), it is essential to remain mindful of how the results may be applied by third parties.

Privacy: Despite limited descriptions of customers, information about purchase behaviour is present. It is important to ensure this information is not misused, as results could be reversed-engineered to impact vehicle insurance premiums based on servicing histories.

Fairness: If the dataset was compiled by the automotive dealer, it may represent a bias towards certain demographics, leading to a biased model.

Section 2: Data Understanding and Preparation

[2.1] Understanding the Data

The provided dataset consisted of 131,337 observations across 17 variables.

> Each observation represented a previous customer and features associated with that customer. Some features included the target repurchase outcome ("Target", where values of 1 indicate the customer has purchased a second vehicle), the car model, the car segment and the age of the vehicle.
>
> See the appendix for a complete list of the features.

The first goal was to understand the variables in relation to the business objective.

By inspecting each variable's contents, key themes were identified:

> **Customer-centric data:**
> such as their corresponding age band and gender.
>
> **Current vehicle-centric data:**
> such as the car model, segment, age and mileage.
>
> **Servicing-related data:**
> such as the amount of scheduled and unscheduled services, the amount paid and the months since their last service.
>
> **Dealership-centric data:**
> such as the number of dealers and whether a second vehicle was purchased at the same dealership.

[2.2] Data Preparation

Prior to modelling, data must be organised by applying a range of standard cleaning and manipulation techniques.

The first comprises the removal of duplicated entries, so as not to weigh a specific outcome higher than others. By removing the ID column, 2726 duplicate entries were revealed and subsequently removed.

The next technique involved amending missing values. Visualising these values with the missingno library revealed 109,668 age band and 67,455 gender values were absent (Figure 1). As the number of missing values was substantial while the weight of the classification feature was considered strong, these features were retained in a separate data frame (Table 1). Oversampling was not applied, as the number of missing values in these fields was deemed too substantial.

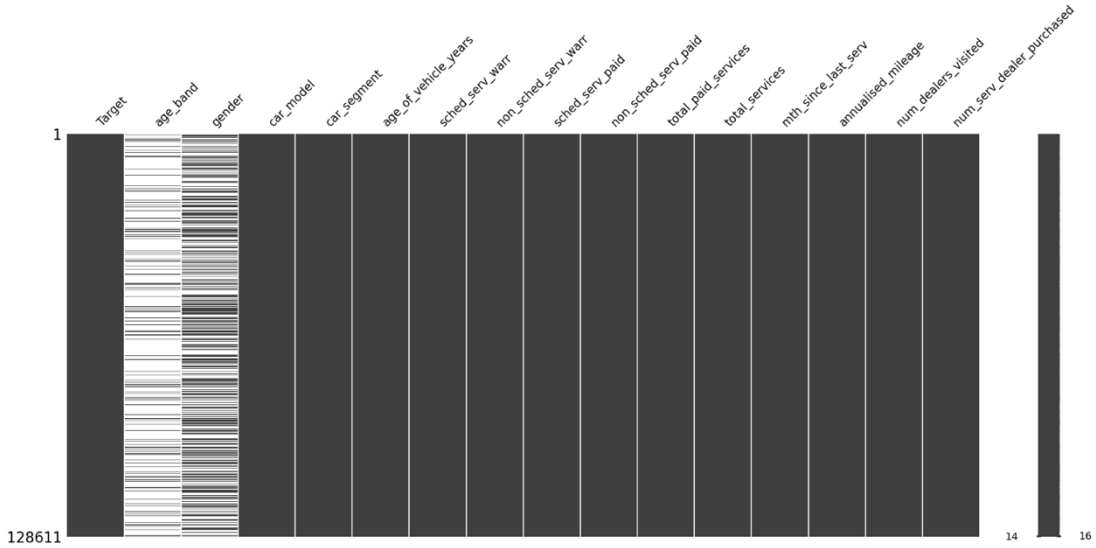

Figure 1 NaN values visualised with Missingno.

Next, statistical characteristics were screened by reviewing the count, mean, max, min and std. By displaying the ranges in an array, it helped identify possible outliers or data collection errors. None were detected.

Lastly, four features were re-engineered from categorical to numerical variables with one-hot-encoding. These variables were age_band, gender, car_model and car_segment.

This resulted in 42 features present within the dataset containing all variables and 34 features in the one without (Table 1).

Table 1 Data frames applied for analysis and modelling.

Data Frame	Feature
cars_ALL	"*all features*" 18289 entries, 42 features
cars_NAG	"*no age-gender*" 128611 entries, 34 features

[2.3] Exploratory Data Analysis

Prior to modelling, the data was explored through graphical representations. The first examined the Target variable, revealing that only 2.7% of the cars_NAG dataset was the group of interest, making this dataset unbalanced (Figure 2).

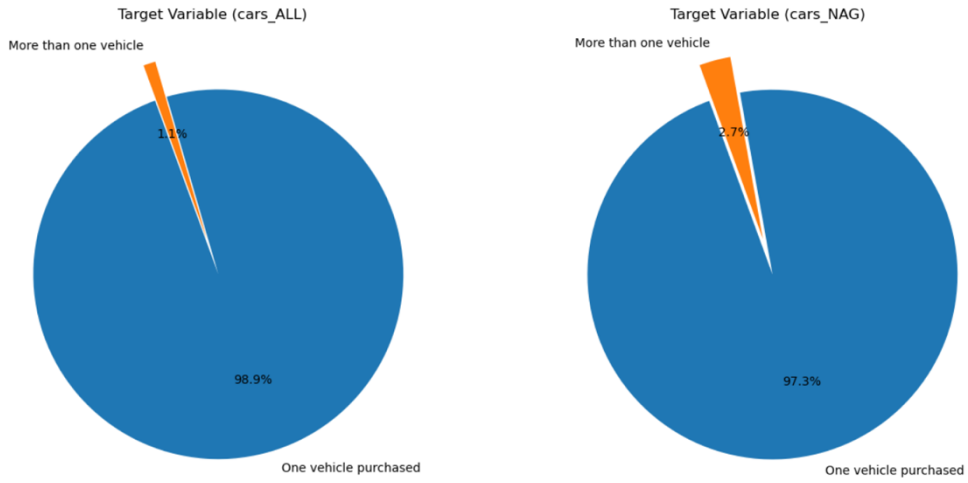

Figure 2 Pie graph visualisations of the Target variable and the class of interest.

Further visuals were constructed to compare the categorical features of the customers that resided within this Target cohort. Despite this dataset representing a fraction of the total observations, it revealed males were more likely to purchase a second vehicle, with the highest age band as 45-54 (Figure 3).

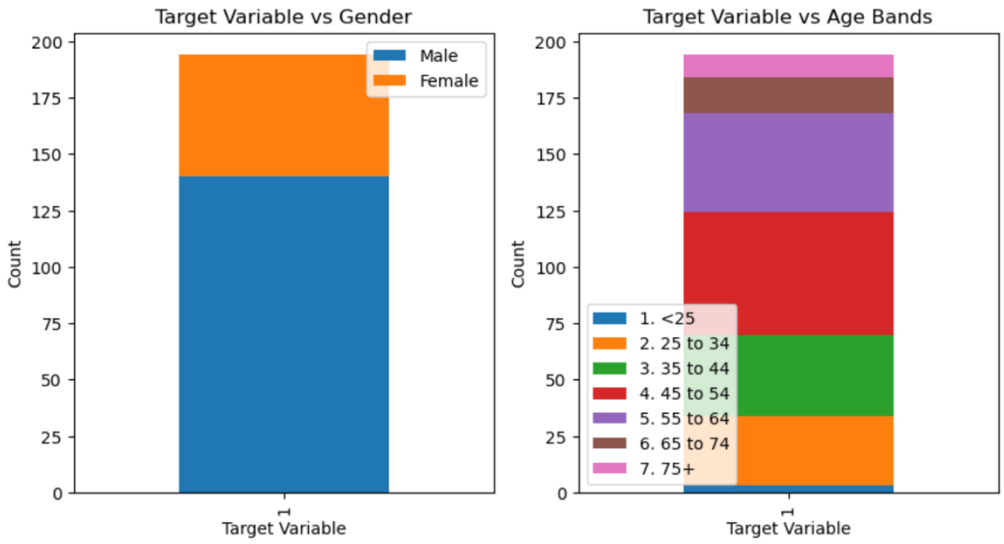

Figure 3 A segmented bar graph illustrating the male and female cohorts, beside the age band cohorts, with Target values of 1.

A graph comparing the preference of car segments and models between the categorical distinctions of the cohort of interest was also constructed to illustrate vehicle popularity (Figure 4). While both sexes preferred Large/SUV models, it was found to be proportionately more popular with males.

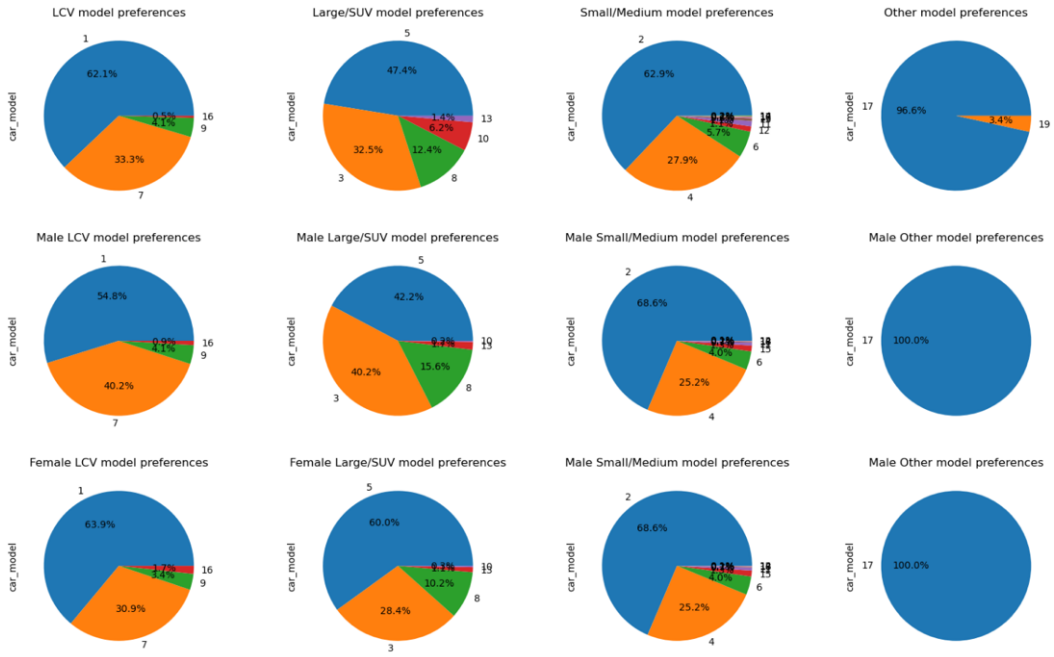

Figure 4 A series of pie charts illustrating the popularity of car models against the variety of car segments. The first row covers the overall majority, while the second and third rows cover the male and female preferences.

The final visual explored notable features of the target cohort's vehicle and its corresponding servicing. These customers typically have 2-3 (deciles) total servicing fulfilments prior to purchasing a new vehicle (Figure 5).

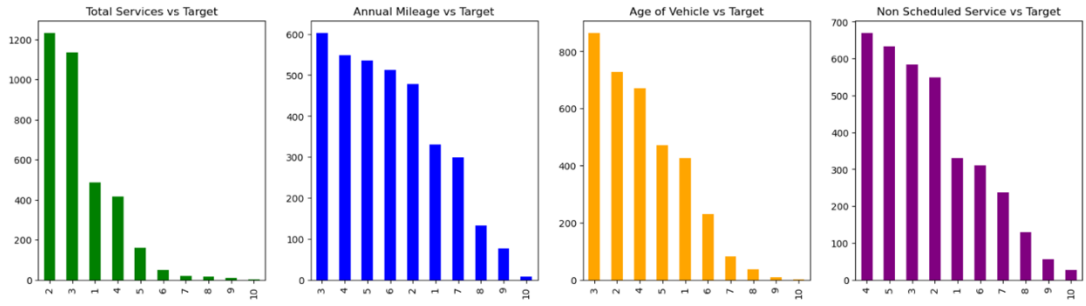

Figure 5 Four bar graphs illustrating total_services, annualised_mileage, age_of_vehicle_years, and non_sched_serv_war in deciles, against the target cohort with values of 1.

Section 3: Modelling

[4.1] Applying Techniques

Six separate modelling techniques were applied in this project, each carried out through an iterative process, applying the learnings from one technique to the one which followed. The performance metric to gauge the performance of each model was a Precision-Recall curve.

Univariate Logistic Regression (ULR) & KNeighbour Classification (KNC)

The first approach applied a simple regression classification of all features against the target feature by applying a ULR and KNC model.

No significant classification insights were acquired (Table 2 & 3).

Table 2 Precision and Recall scores of the univariate logistic regression function and the KNeighbours classification function on the cars_NAG data frame.

Feature	Precision_Train	Recall_Train	Recall_Test	Precision_Test
age_of_vehicle_years	0.0	0.0	0.0	0.0
sched_serv_warr	0.0	0.0	0.0	0.0
non_sched_serv_warr	0.0	0.0	0.0	0.0
total_paid_services	0.0	0.0	0.0	0.0
total_services	0.0	0.0	0.0	0.0
mth_since_last_serv	0.0	0.0	0.0	0.0
annualised_mileage	0.0	0.0	0.0	0.0
num_dealers_visited	0.0	0.0	0.0	0.0
num_serv_dealer_purchased	0.0	0.0	0.0	0.0
annualised_mileage	0.0	0.0	0.0	0.0
car_model_1	0.0	0.0	0.0	0.0
car_model_2	0.0	0.0	0.0	0.0
car_model_3	0.0	0.0	0.0	0.0
car_model_4	0.0	0.0	0.0	0.0
car_model_5	0.0	0.0	0.0	0.0
car_model_6	0.0	0.0	0.0	0.0
car_model_7	0.0	0.0	0.0	0.0
car_model_8	0.0	0.0	0.0	0.0
car_model_9	0.0	0.0	0.0	0.0
car_model_10	0.0	0.0	0.0	0.0
car_model_11	0.0	0.0	0.0	0.0
car_model_12	0.0	0.0	0.0	0.0
car_model_13	0.0	0.0	0.0	0.0
car_model_14	0.0	0.0	0.0	0.0
car_model_15	0.0	0.0	0.0	0.0
car_model_16	0.0	0.0	0.0	0.0
car_model_17	0.0	0.0	0.0	0.0
car_model_18	0.0	0.0	0.0	0.0
LCV	0.0	0.0	0.0	0.0
Large/SUV	0.0	0.0	0.0	0.0
Small/Medium	0.0	0.0	0.0	0.0
Other	0.0	0.0	0.0	0.0

Table 3 & 4 Confusion matrix results of the univariate logistic regression function and the KNeighbours classification function on the cars_NAG data frame.

CM_train	
100071	0
2817	0

CM_test	
25019	0
704	0

Multivariate Logistic Regression (MLR)

Following the ULR and KNC modelling phase, an MLR function was developed and applied to all features for both datasets. The resulting Precision-Recall curve for cars_ALL was not uniform and dropped prior to hyperparameter tuning. The most successful combination of hyperparameter tuning applied with this model and the cars_NAG dataset is displayed in the figures below. No significant features were discernible with this model.

Dataset: cars_NAG
Hyperparameters tuned: 'penalty' (none)
'C' (0.1, 0.5, 1.0)
'solver' (none)

Figure 6 The Precision-Recall curve generated from MLR modelling, against the testing set.

Table 8, 9, 10 The confusion matrix results, precision & recall results, Accuracy score and F1 score, of both MLR Training and Testing sets.

Confusion Matrix Training Set	
99951	120
2193	624
Confusion Matrix Testing Set	
24991	28
563	141

Training Set	
Precision	0.839
Recall	0.222
Testing Set	
Precision	0.834
Recall	0.200

Accuracy Training Set	0.98
Accuracy Testing Set	0.98
F1 Score Training Set	0.97
F1 Score Testing Set	0.97

Support Vector Classification (SVC)

Following the MLR modelling phase, an SVC function was developed and applied to all features for both datasets. Once more, the resulting Precision-Recall curve for cars_ALL was not uniform and dropped prior to hyperparameter tuning. The most successful combination of hyperparameter tuning applied with this model and the cars_NAG dataset is displayed in the figures below. Evaluation for feature significance is not supported with this model.

Dataset: cars_NAG
Hyperparameters tuned: 'C' (200)
'kernal' (rbf)
'gamma' (auto)

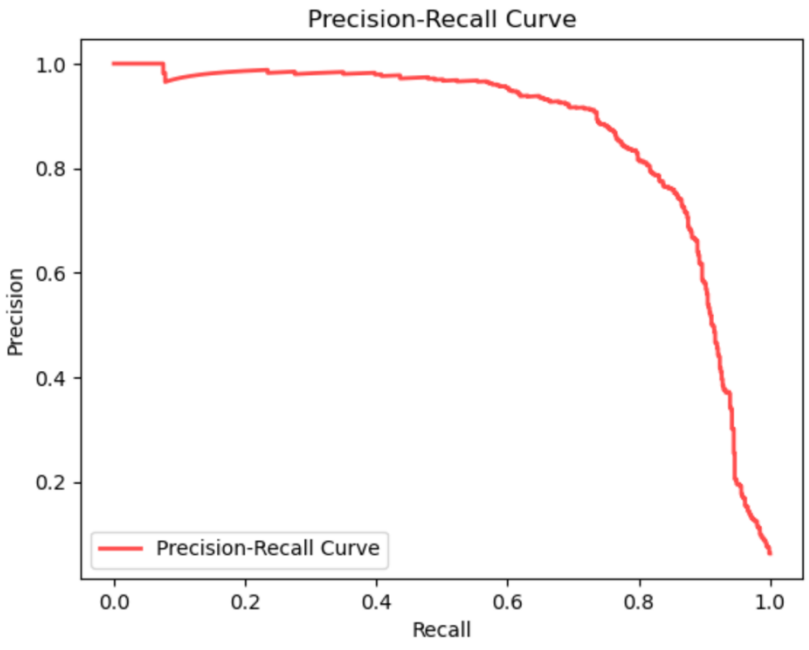

Figure 7 The Precision-Recall curve generated from SVC modelling, against the testing set.

Table 11, 12, 13 The confusion matrix results, precision & recall results, Accuracy score and F1 score, of both SVC Training and Testing sets.

Confusion Matrix Training Set	
99951	120
619	2198
Confusion Matrix Testing Set	
24967	52
188	516

Training Set	
Precision	0.948
Recall	0.780
Testing Set	
Precision	0.908
Recall	0.733

Accuracy Training Set	0.99
Accuracy Testing Set	0.99
F1 Score Training Set	0.99
F1 Score Testing Set	0.99

Random Forest Classification (RFC)

Following the SVC modelling phase, an RFC function was next developed and applied to all features for both datasets. The resulting Precision-Recall curve for cars_ALL remained ununiform and was dropped prior to hyperparameter tuning. The most successful combination of hyperparameter tuning applied with this model and the cars_NAG dataset is displayed in the figures below. Significant features of the cars_NAG dataset are displayed in Figure 9.

Dataset: cars_NAG
Hyperparameters tuned: 'max_depth' (40)
'n_estimators' (200)
'criterion' (entropy)

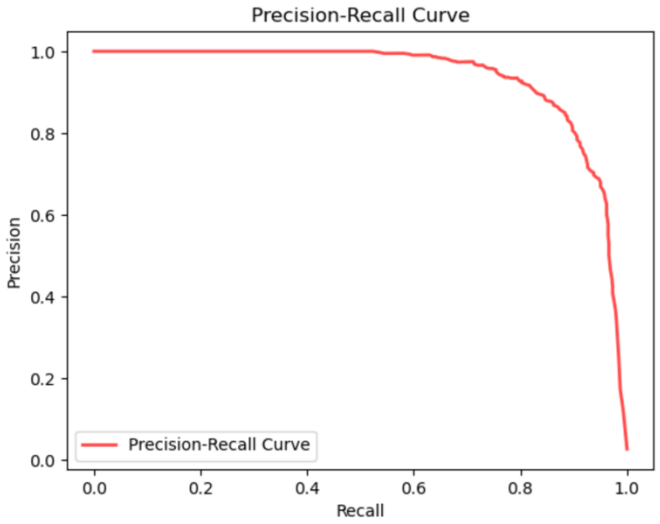

Figure 8 The Precision-Recall curve generated from RFC modelling, against the testing set.

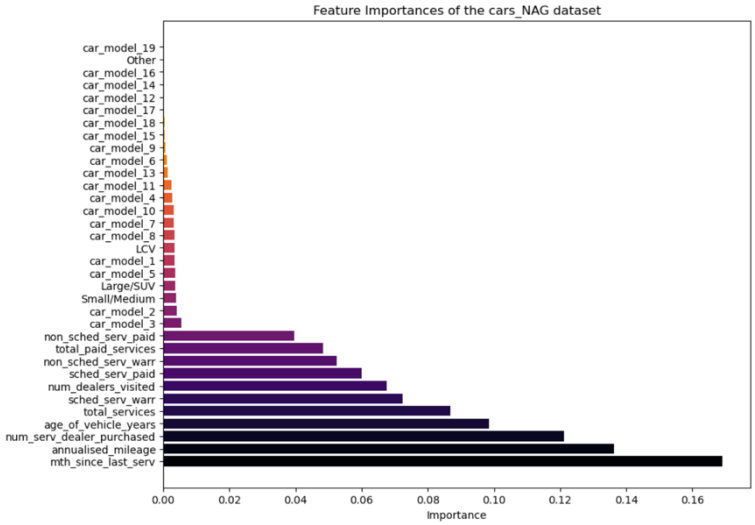

Figure 9 The significance of each feature in the RFC model.

Table 14, 15, 16 The confusion matrix results, precision & recall results, Accuracy score and F1 score, of both RFC Training and Testing sets.

Confusion Matrix Training Set	
100071	0
1	2816
Confusion Matrix Testing Set	
25001	18
190	514

Training Set	
Precision	1.0
Recall	1.0
Testing Set	
Precision	0.966
Recall	0.73

Accuracy Training Set	1.0
Accuracy Testing Set	0.99
F1 Score Training Set	1.0
F1 Score Testing Set	0.99

XGBoosting Classification (XGB).

Following the RFC modelling phase, a final function was developed for XGB modelling and applied to all features for both datasets. The resulting Precision-Recall curve for cars_ALL continued to remain ununiform and was dropped prior to hyperparameter tuning. The most successful combination of hyperparameter tuning applied with this model and the cars_NAG dataset is displayed in the figures below. This model received three additional features, highlighted in Figure 11, alongside the significant features of the dataset.

Dataset: cars_NAG
Hyperparameters tuned: 'max_depth' (5)
'min_child_weight' (1)
'subsample' (0.8)
'colsample_bytree (0.8)
'learning_rate' (0.2)

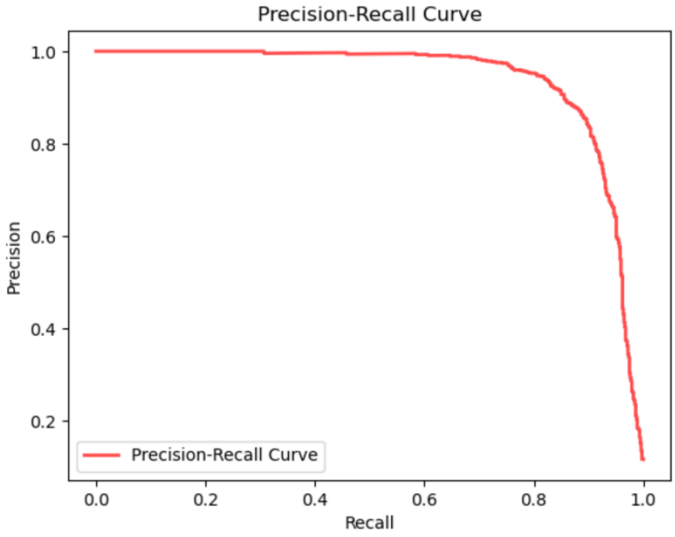

Figure 10 The Precision-Recall curve generated from RFC modelling, against the testing set.

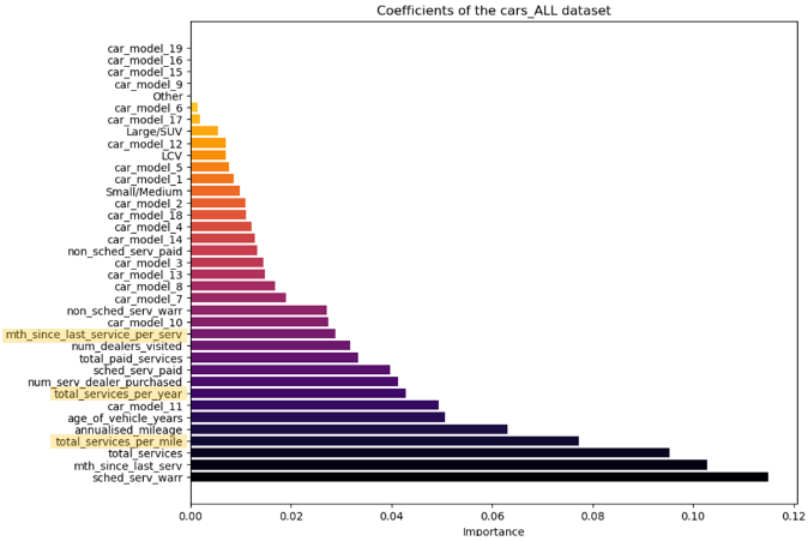

Figure 11 The significance of each feature in the RFC model, including engineered features.

Table 17, 18, 19 The confusion matrix results, precision & recall results, Accuracy score and F1 score, of both RFC Training and Testing sets.

Confusion Matrix Training Set	
99962	109
422	2395
Confusion Matrix Testing Set	
24986	33
131	573

Training Set	
Precision	0.961
Recall	0.851
Testing Set	
Precision	0.948
Recall	0.808

Accuracy Training Set	0.99
Accuracy Testing Set	0.99
F1 Score Training Set	0.99
F1 Score Testing Set	0.99

Section 4: Evaluation

[4.1] Results

To meet the requirements of the business, the nominated model needed to offer a clear and accurate classification of the features associated with the cohort of interest. Based on the project's exploration, the outcome provides a choice between two models: RFC and XGB, where sacrificing a little over a percentage in precision can provide an almost 10-fold increase in recall.

> **RFC** provides a precision of **96.6**%, with a recall of **73**%.
>
> **XGB** offers a precision of **94.8**%, with a recall of **80.8**%.

If XGB is considered, it successfully identifies the needs of the business by providing four key metrics to assess when searching for leads for prospective customers:

> i. The number of scheduled services under warranty.
> ii. The number of months since the last service.
> iii. The total number of services.
> iv. The annualised vehicle mileage.

[4.2] Recommendation

Based on the analysis conducted and sequential modelling phases, it is recommended that the dealership considers proceeding with its marketing campaign. Should prospective customers share the features above and possess similar ranges to existing customers who already have purchased a second vehicle, it is likely that these prospective customers should be the target for the campaign.

No further modelling is recommended, as the discrepancy in false positives represents a minor percentage overlooked unless this percentage is of particular interest to the dealership.

[4.3] Deployment

This report summarises all modelling phases undertaken for this project. As we are utilising a classification model to categorise a cohort, no technical deployment is required unless automating the screening process of potential customers is required. Deployment is otherwise utilising the recommendations held within this report.

Appendix

ID: *Unique ID of the customer*

target: *Model target. 1 if the customer has purchased more than 1 vehicle, 0 if they have only purchased 1.*

age_band: *Age banded into categories*

gender: *Male, Female or Missing*

car_model: *The model of vehicle, 18 models in total*

car_segment: *The type of vehicle*

age_of_vehicle_years: *Age of their last vehicle, in deciles*

sched_serv_warr: *Number of scheduled services (e.g. regular check-ups) used under warranty, in deciles*

non_sched_serv_warr: *Number of non-scheduled services (e.g. something broke out of the service cycle) used under warranty, in deciles*

sched_serv_paid: *Amount paid for scheduled services, in deciles*

non_sched_serv_paid: *Amount paid for non scheduled services, in deciles*

total_paid_services: *Amount paid in total for services, in deciles*

total_services: *Total number of services, in deciles*

mth_since_last_serv: *The number of months since the last service, in deciles*

annualised_mileage: *Annualised vehicle mileage, in deciles*

num_dealers_visited: *Number of different dealers visited for servicing, in deciles*

num_serv_dealer_purchased: *Number of services had at the same dealer where the vehicle was purchased, in deciles*

Engineered Features:

'total_services_per_mile'*: The total services divided by the annualised.*

'total_services_per_year'*: The total services divided by the age of the vehicle.*

'mth_since_last_service_per_serv'*: The total services divided by the age of the vehicle.*