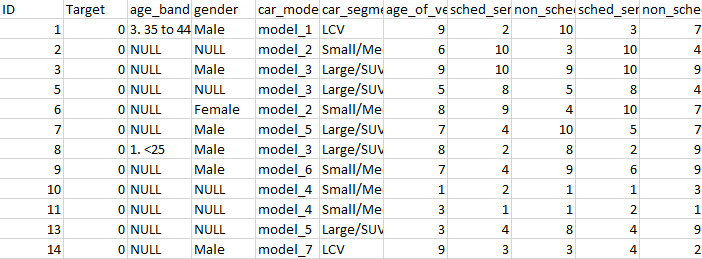
**Repurchase Campaign to Target Existing Customers**

**Business Understanding**We need to target existing customers of this automotive manufacturer for a re-purchase campaign. As it is significantly more expensive to acquire new clients than it is to keep existing ones. A considerable amount of the revenue comes from repeat consumers. According to studies, repeat customers can account for more than 40% of a company's income, and increasing customer retention rates by 5% can boost earnings by 25% to 95%. This type of recurrent purchase can help increase return on investment (ROI) and assure long-term growth. Purchases made by a returning customer are known as repeat purchases. These customers are well-versed in your brand and are frequently motivated by a need for convenience. Why change something that is working?

The goal of this campaign is to send a message to clients who are very likely to buy a new car. Personalized engagement after purchasing a new vehicle can leave a positive and lasting impression on the customers and help them remember your brand when they need to repurchase.

**Data Understanding & Data Preparation**We have been given a customer purchase dataset which has customer demographics, past car type purchases, vehicle age, and servicing information. As there was a lot of noise in the numeric variables so I have to transformed all the numeric variables into deciles. The numeric variables that I transformed to deciles are mentioned below in the table.

Here is a small sample of the data  


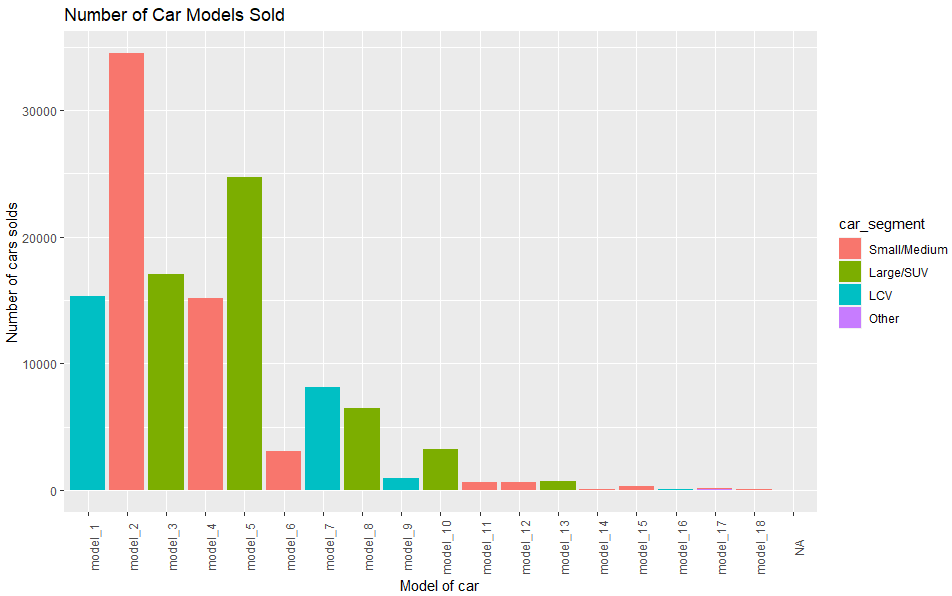
We have been given repurchase\_training dataset which has 131338 rows for training the model and we also have been given repurchase\_validation dataset with 50000 rows to validate the model.

Here is a Data Dictionary to describe the different variables in the dataset

|  |  |
| --- | --- |
| **Variables** | **Variable description** |
| ID | 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 bands of customers |
| gender | Gender of the customer |
| car\_model | The model of vehicle |
| car\_segment | Vehicle Type |
| age\_of\_vehicle\_years | Previous vehicle Age, value is in deciles |
| sched\_serv\_warr | Number of scheduled services used under warranty, value is in deciles |
| non\_sched\_serv\_warr | Number of non-scheduled services used under warranty, in deciles |
| sched\_serv\_paid | Amount paid for scheduled services, value is in deciles |
| non\_sched\_serv\_paid | Amount paid for non-scheduled services, value is in deciles |
| total\_paid\_services | Amount paid in total for services, value is in deciles |
| total\_services | Total number of services, value is in deciles |

As seen above many of the numeric variables are in declies. We Converted age\_band, gender, car\_model and car\_segment variables to categorical variables as well. Upon closer examination, we find the variable gender has 69,308 Null values, variable age\_band has 112,375 Null values and car\_model variable has two missing values. We then delete two rows from the dataset i.e. Row number 26319 & 85668 as these two rows contain the missing values of the car\_model variable. We then divided the dataset into 6 to generate testing and training pairs.

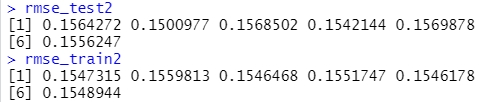
Upon doing some Exploratory data analysis on the dataset we come to know that 97.3% of the customers chose to not repurchase from the manufacturer and only 2.7% of the existing customers chose to purchase again from the manufacturer. ID variable also can be removed as it serves us no purpose, it won’t be of any help for us to predict.



Evidently, Model\_2 and Model\_5 are the most models sold by the manufacturer. Approximately 25,000 model\_2 was purchased by customers and approximately 25000 model5 was purchased. Model\_2 is a small/medium car, whereas model\_5 is a large SUV.

**Modelling**We run the generalized linear model on the created 6 folds. Upon running the glm model on the 6 folds. Fold No. 2 performed the best, with the highest AIC score(Akaike Information criterion). It had an AIC score of – 104363. We also find that the following variables: car\_segment, age\_of\_vehicle\_years & non\_sched\_serv\_paid are statistically insignificant and we remove them from the dataset. After removing the above 3 variables from the dataset. Fold no. 2 gives us the best AIC score of -96037. The AIC value of -96,037 is good as we have a data of 131,337 observations. Upon running the model after removing all age and genders as a lot of the values in that variable was NULL, the AIC score improved only marginally.

Upon testing the Root mean square error of all the 6 folds. Fold No. 2 had the lowest RMSE of 0.1500977. Indicating that it is the better fit. Upon comparing the RMSE of training and testing set of the 6 folds. It looked like the training set had lower values, meaning the model fit better.



The GLM got an AUC (Area under the curve) score: 0.901

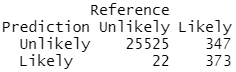
The confusion matrix of the model also yielded satisfactory results. As seen below it had high number of true positives and true negatives.



We got a precision of 87%, Recall of 99.06% and F1 of 92.6%. for this model fold 2 worked best

We then build a gradient boosting model for our tree-based classifier model as our target is unbalanced. GBM model builds trees, where new tree corrects errors made by previous tree

It is excellent at detecting anomaly, especially when data is unbalanced. We split the data into 80-20 ratio of training and testing respectively. Below is the confusion matrix of the GBM model. As seen below it has high number of true positives and true negatives.



IN GBM we got an AUC of 98%, precision of 98.65%, Recall of 99.99% and F1 of 99.28%.

**Evaluation & Deployment**

Variable importance: As discussed earlier for GLM, we had found that the variables: car\_segment, age\_of\_vehicle\_years & non\_sched\_serv\_paid were statistically insignificant. Whereas for the GBM model we find out that mth\_since\_last\_serv & num\_serv\_dealer\_purchased are statistically very significant variables. On the contrary GBM considers age\_of\_vehicle\_years to be significant and GLM though otherwise.

Upon evaluating it is clearly evident that the Gradient boosting model performs much better than the gradient boosting model as the AUC, precision, Recall and F1 of GBM is higher than GLM model.

Upon running our model on the validation dataset we estimate that 49,180 customers are highly unlikely to repurchase a car from the manufacturer, whereas 820 customers from the validation dataset are likely to repurchase. The ID of the above customers is mentioned in the csv file to target them for the company’s repurchase campaign.