Classification Modelling

This task will require you to develop and deploy a classification model on a product purchase data set.

You are now a Data Scientist working for an international consulting firm. An automotive manufacturer has approached you to help them target existing customers for a re-purchase campaign. The aim of this campaign is to send a communication to customers who are highly likely to purchase a new vehicle. All customers have already purchased at least one vehicle.

The automotive company has supplied a data set of customer demographics, previous car type bought, the age of the vehicle, and servicing details. Note that the servicing details are only for mechanics at official dealerships. You have looked at this data set and already transformed many of the variables to make the modelling easier. There was a lot of noise in the numeric variables so you decided to transform all the numeric variables into deciles (integers 1 to 10, each has a similar number of customers). The deciles can be treated as numeric or factors in R. The data dictionary for this data set is given below:

|  |  |  |
| --- | --- | --- |
| **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, in deciles |
| **sched\_serv\_warr** | Integer | Number of scheduled services (e.g. regular check-ups) used under warranty, in deciles |
| **non\_sched\_serv\_warr** | Integer | Number of non-scheduled services (e.g. something broke out of the service cycle) used under warranty, in deciles |
| **sched\_serv\_paid** | Integer | Amount paid for scheduled services, in deciles |
| **non\_sched\_serv\_paid** | Integer | Amount paid for non scheduled services, in deciles |
| **total\_paid\_services** | Integer | Amount paid in total for services, in deciles |
| **total\_services** | Integer | Total number of services, in deciles |
| **mth\_since\_last\_serv** | Integer | The number of months since the last service, in deciles |
| **annualised\_mileage** | Integer | Annualised vehicle mileage, in deciles |
| **num\_dealers\_visited** | Integer | Number of different dealers visited for servicing, in deciles |
| **num\_serv\_dealer\_purchased** | Integer | Number of services had at the same dealer where the vehicle was purchased, in deciles |

Tasks:

For this task you have been given two csv files, ‘**repurchase\_training.csv’** and **‘repurchase\_validation.csv’**. The tasks below are to be undertaken on ‘**repurchase\_training.csv’** since you have the target variable included.

1. Undertake EDA on this dataset. Justify all decisions you make relating to data processing.

2. Build a linear **classification** model to predict which customers are most likely to repurchase. You can use any technique we used in class with a classification target.

a. Create the confusion matrix & calculate precision, recall, F1 & AUC.

b. Given known best practice for classification tasks and the particular situation at hand, which metric will you use to decide your final model?

3. Build a tree based classification model to predict which customers are most likely to repurchase.

a. How does your tree-based model perform relative to the linear model? (Hint: If you did not pick a good metric in 2.b you will not be able to compare the models!)

b. Discuss the variable importance measures from both types of models (i.e what does ‘variable importance’ mean for these different model types).

c. In this particular task, do they propose different levels of importance for the features? Why do you think this is? Provide intuitive interpretation of these importances, given the context at hand (i.e What do these importances mean in a ‘business sense’)

d. For your tree-based model, construct partial dependency plots for the top 5 most important features. Do you notice anything interesting about any of these plots?

4. With the best model created above, use the variables in **‘repurchase\_validation.csv’** to output both probabilities and class predictions. You will see this file contains a validation data set, but the target variable has been excluded. Your model will be marked against this data set. It is important to follow the following submission guidelines:

1. a. name the file repurchase\_validation\_STUDENTNUMBER.csv So if your number is 1234 the file should be repurchase\_validation\_1234.csv
2. Include only three columns, named as follows, in this order. Note that target\_class must be 1 for purchase and 0 for not. The probability here is for the positive (1) class. No need for two columns with two probabilities
   1. ID
   2. target\_probability
   3. target\_class

5. Reporting (500 – 1000 words)

a. Using all the notes and answers you have above, wrap up all your work into a report for the manager of the automotive firm that follows the CRISP-DM methodology. Remember to consider your audience!