J]UJWNR JSY%WJUTWY

Student Name	Shalimar Chalhoub
Project Name	Classification Models
Date	25/3/2023
Deliverables	<MLAA - Assignment 2 Method 3 - Decision Tree> <Decision Tree>

---

1. EXPERIMENT BACKGROUND	
Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach.	
1.a. Business Objective	The goal of this project to the business is to accurately identify which customers are going to be repurchasing a second car from the company depending on several factors such as their gender, age and the model and segment of their car as well as some other details about the car. These results can be used to target those types of customers and spend more resources on getting them to buy a second car. Inaccurate results may cause the company to lose money by targeting the wrong customers as well as lose sales.
1.b. Hypothesis	I want to test whether decision trees can help accurately identify if a customer is more likely to repurchase or not. The reason behind using a decision tree is because it can handle nonlinear data, categorical and numerical data as well as large datasets and can focus on the important features.
1.c. Experiment Objective	I think the outcome of the decision tree model would be positive and I would have a recall score above 75% since this model should adapt to my model. Possible outcomes: 1. The model can either have a large recall and thus be efficient to be deployed by the company 2. Or, the model with produce a low recall score and would be dismissed

2. EXPERIMENT DETAILS	
Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them.	
2.a. Data Preparation	For this model, I have used datasets that have already been preprocessed and split during the first experiment and thus no data preparation took place. The rationale has been explained in the first experiment report The steps I decided not to use is data balancing, as I have tried it and it made my model's predictions uneven and skewed.
2.b. Feature Engineering	Feature engineering was performed before splitting the datasets in experiment 1. It included imputation of gender and age_band variables as well as checking whether a model had 2 different car_segments and fixing them It was found that model_17 had the segments small/medium as well as others and thus others was transformed to small/medium for that model
2.c. Modelling	The model used for this experiment is DecisionTreeClassifier which is and algorithm that builds a tree model to classify data based on a set of rules. I chose this model because it can handle imbalanced data as well as large datasets and non-linearity The hyperparameters tuned: 1. Min_samples_split=30 2. Max_depth=15 I tested multiple values on them trying to reduce the overfitting and finally found the best results using the values above. For Min_sample_split, I had to increase it given the complexity of my dataset, as for Max_depth, also that was increased till 15 which reduced overfitting and bettered my results For the future, we can experiment more hyperparameters such as min_samples_leaf and max_leaf_nodes, but for my model, both of these were left untouched

3. EXPERIMENT RESULTS	
Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified.	
3.a. Technical Performance	The recall scores are listed below: Train set:0.80 Dev set:0.73 Test set: 0.75 Whilst I find these results acceptable, they do seem a little low and that is because the dataset is very skewed and imbalanced as well as missing a lot of values which we had to predict
3.b. Business Impact	I think these results work very well with the business objective as it helps them focus on the clients that are likely to repurchase again. 75% is a very good score for this type of targeted marketing
3.c. Encountered Issues	These are the issues encountered with this experiment: 1. Missing data: this was fixed at the beginning of experiment 1 using imputation 2. Imbalanced data: this has not been fixed as the model performs well without it 3. Overfitting: this has been solved by tuning the parameters 4. Underfitting: This has also been solved by parameter tuning

4. FUTURE EXPERIMENT	
Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective.	
4.a. Key Learning	Some key learnings from this experiment is the importance of hyperparameters in the outcome of the model and how important it is to take time to finetune them as it majorly affects the outcome of the model

4.b. Suggestions / Recommendations	I would say this method produces the results I am looking for but nevertheless, they could be bettered by trying to sample or balance the datasets. Trying different or additional parameters can help as well. This method is definitely worth pursuing. For the deployment, it could be a program which takes in a dataset of features and produces a list of potential customers who are worth pursuing and targeting with marketing in order to get them to repurchase