J]UJWNR JSY%WJUTWY

Student Name	Shalimar Chalhoub
Project Name	Classification Models
Date	25/3/2023
Deliverables	<MLAA - Assignment 2 Method 4 - Random Forest> <Radom Forest>

---

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 Random forest with manually chosen parameters can accurately help predict whether a customer is more likely to repurchase a car The reason behind me choosing random forest is because it usually has a high accuracy because the decision is being taken by many trees instead of one and it works well with noise and outliers and can handle big non-linear datasets such as the one I have. It also has feature importance, so it measures the importance for every feature of the dataset which helps the interpretability of the model The reason why I decided to manually choose my parameters is to have more room for experimentation to perfect the model
1.c. Experiment Objective	I think the outcome of the random forest model would be positive and I would have a recall score above 75% since this model should adapt to my datal. 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 RandomForestClassifier which is an ensemble method that combines multiple decision trees which will be trained on a subset of the data and features and the best prediction will be chosen by majority voting. The hyperparameters tuned: 1. n_estimators =100 2. Max_depth=15 3. min_samples_leaf=10 4. max_features = None I tested multiple values on them trying to reduce the overfitting and finally found the best results using the values above. For n_estimators I tried a range from [50-200]increasing by 50 and found 100 the best. For Max_depth I tested it on a model without specifying and it gave me 27 do I tried to reduce from there. For min_sample_leaf and max_features I experimented and found that this combination performs the best I decided not to sample for this experiment as I found it gave me biased results, however, for the future we can try other sampling techniques

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:0.78 Dev set:0.73 Test set: 0.73 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 okay well with the business objective as it helps them focus on the clients that are likely to repurchase again. 73% is an acceptable score for this type of targeted marketing however, not the result we wished to achieve
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	I think overall, this experiment did work well and a lot of insights were gained on how hyperparameters affect the overfitting and underfitting of the model. I think this project is not a dead end however, there might be a need to fix the issue of the imbalanced dataset using a method that will not affect the biasing upon prediction
4.b. Suggestions / Recommendations	I think even though this model did not reach the desired above 75% outcome, it can still be deployed and from it, we can extract a lead list with customers that are likely to purchase a second car and we can target a marketing campaign for them either by email or phone call and try to at least get 80% to repurchase since the model predicts 73% and with the marketing campaign's offer it can be turned to 80.