EXPERIMENT REPORT

Student Name	Shalimar Chalhoub
Project Name	Classification Models
Date	26/3/2023
Deliverables	<MLAA - Assignment 2 Method 2 - KNN> <KNN>

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	Since the data seems to be non-linear, I want to test out if KNN would be a good model that accurately predicts whether or not a customer is more likely to repurchase a second car The reason behind using KNN is because it can handle nonlinear data, and it is easily interpretable meaning that I can easily figure out the reason behind a bad score.
1.c. Experiment Objective	I think the outcome of the KNN 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 to use is data balancing by undersampling using EditedNearestNeighbours. I have tried many other undersampling and oversampling techniques and this one gave me the best results
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 first model used is the EditedNearestNeighbours which is an under-sampling model that removes examples from the training dataset which are classified incorrect by their nearest neighbors thus helping remove noise and irrelevant datapoints. The prediction model used is KNeighborsClassifier which is a classification algorithm that works by finding the K-nearest neighbors and assigning a label based on the majority class of neighbors I chose the following parameters: • N_neighbors = 22 • Metric = ‘euclidean’ I tested many n_neighbors hoping to find the best one and I found that ultimately, 22 gave me the best results. I have also tested Manhattan, euclidean and Hamming metrics and Euclidean worked best

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:0.54 Dev set:0.53 Test set: 0.54 These results are very low and I do not recommend deploying this method
3.b. Business Impact	I think that by using this model, the buisness would not lose money, howver, it would not make as much as it should which is the double, thus, I do not recommend it
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 5. Low scores: I could not fix this as the datasets is too imbalanced and missing a lot of data

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 do think that this experiment is a dead end because knn doesn’t work the best with skewed and imbalanced data however, if we can try to balance it then it can work

4.b. Suggestions / Recommendations	I do not think that this model can be further bettered by using KNN however, some methods that can be tried are balancing or sampling the data which could reduce the error. Another method is to try to differently predict the missing value as this might have had an effect on the data The last thing we can do is check for another metric that might work better on imbalanced data