EXPERIMENT REPORT

Student Name	Shalimar Chalhoub
Project Name	Logistic Regression
Date	26/3/2023
Deliverables	<MLAA - Assignment 2 Method 1 – Logistic regression> <Logistic Regression>

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	The hypothesis to be tested is whether Logistic regression model would be ideal for a classification problem where I have to predict whether or not a customer will repurchase a car based on a set of features If my data is linear, Logistic regression will work really well
1.c. Experiment Objective	I think the outcome of the Logistic regression 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	The first thing I did is check on the columns to see if there is any data identifier that should not be there but it was all good. Second, I dropped the ID column as it causes overfitting Third I checked for duplicates and removed them because they are redundant to have Then, I checked to see whether some car models had more than 1 segment and found out that model_17 did, so I fixed them all After that, I mapped my gender and age_band data and replaced them by the dictionary in order to label encode them Next, I did one-hot encoding on the rest of the data Then, after splitting, because my data had a lot of missing values, I imputed them and replaced the missing values by the predicted one and download those datasets for use in the future experiments instead of having to pre process everytime
2.b. Feature Engineering	Feature engineering 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. The only column removed was the ID column as it causes overfitting
2.c. Modelling	The model that was used is LogisticRegression which a model that predicts binary labels for new data based on estimated probabilities. I used the following hyperparameters: max_iter =800 ; because my data was too big it was not going through all of them random_state = 42 C=0.5 ; as my data is imablanced I think for the future, we can experiment with C more

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 f1 scores are listed below: Train set:0.80 Dev set:0.21 Test set: 0.20 These results are very low and I do not recommend deploying this method I think I got these results because the data turned out to be non linear
3.b. Business Impact	Deploying this model would cost the business a lot fo money because even though it might predict correctly most of the true values, it still marks a lot of negative value as true thus majorly wasting the time of employees
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 not been solved 5. Low scores: I could not fix this as the datasets is too imbalanced and non linear

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 logistic regression assumes data linearity and this data is not so my suggestion is to move to another method

4.b. Suggestions / Recommendations	I do not think that this model can be further bettered by using logistic regression however, some methods that can be tried are linearizing the data which could better the results Another thing we can do is to see whether there is some hyperparameters we can tune that might make it better