EXPERIMENT REPORT

Student Name	Trung Kien Hoang
Project Name	Customer behavior models in buying a new car
Date	28/04/2023
Deliverables	<Experiment 3> < Logistic Regression Classifier, Logistic Regression Classifier with L1 and L2 Regularization, KNN model using Euclidean distance, SVC with default hyperparameters, Decision tree with default hyperparameters> <Feature engineering with gender, car segments and age group >

• **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	This initiative aims to develop a model to predict whether an existing customer will purchase a new car. This model's output will be used to target potential customers in a marketing campaign. More specifically, the marketing team might employ the model to find clients who are more likely to purchase a new car and target those individuals. It can improve the marketing campaign's effectiveness by preventing the team from wasting efforts on customers who are unlikely to make a purchase. Suppose the model accurately forecasts which customers will likely buy a new car. In that case, the marketing team can use the model to find targeted customers and advertise a new car model, which leads to increased sales and income for the company. Nonetheless, if the model is inaccurate, the marketing team might waste efforts on ineffective campaigns, which could result in a decline in sales and revenue. In addition, the company's relationship with those who dislike marketing communications will have a negative impact.
1.b. Hypothesis	**Hypothesis**: The greater probability of purchasing a new car corresponds to the presence of customers characteristics such as gender and age; and the car variables such as the type of vehicle, the age of the last vehicle, the number of scheduled services, the number of non-scheduled services, the amount paid for scheduled services, the amount paid for non-scheduled services, the amount paid in total for services, the total number of services, the number of months since the last service, the annualised vehicle mileage, the number of different dealers visited for servicing and the number of services had at the same dealer where the vehicle was purchased.

	It is worth considering because gaining an awareness of the specific car factors that influence the purchase of new cars may shed light on ways in which a company can enhance the quality of the products and services it offers to satisfy the requirements of its clientele better.
1.c. Experiment Objective	The anticipated outcome of the endeavour is a model that accurately predicts whether an existing consumer will purchase a new car. Because the given dataset is imbalanced, the F1 score is a metric incorporating recall and precision of data. It is a measurement that balances the model's performance in both classes and considers both false positives and false negatives. Ideally, the model's F1 score would be at least 0.7, where the maximum score is 1.0 and higher values indicate superior performance. **Scenario 1**: Train the selection of algorithms; for all data sets, if the best model's performance is less than 0.7 of the F1 score, take one or more actions: Change data preparation, feature engineering, hyperparameter optimization, or algorithm selection. **Scenario 2:** Train the selection of algorithms; for all data sets, if the performance of a model is greater than 0.70 and less than 0.90 of the F1 score, store the model in the evaluation section and proceed to a new experiment. **Scenario 3:** Train the selection of algorithms; for all data sets, if the performance of a model is greater than 0.9, finish the project and deploy the model to the business.

• **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	1. Removing duplicated records: This help to reduce running time and prevent bias model. 2. Remove model_19, which is not belong to car models (from 1 to 18). This help to improve model performance. 3. Replace the type of car “Other” with “Small/Medium” in all cars model 17. 4. Remove all observation with missing values in the age group column: This help to improve model performance. 5. Remove all observation with missing values in the gender column: This help to improve model performance. 6. Drop ID columns: This variable is unnecessary in binary classifications. 7. Data splitting: The database is divided into three distinct collections for training, validating, and testing. Whereas the training set is used to fit the model, the validating set is used to modify hyperparameters, and the testing set is used to evaluate the model with unseen data. 8. Data scaling: Scale the data.
2.b. Feature Engineering	1. Extracting age group: This help to transform categories variables to numeric variables. 2. One-hot encoding the gender and car segment variables: This help to transform categories variables to numeric variables.
2.c. Modelling	In this experiment, there are five models: Logistic Regression Classifier, Logistic Regression Classifier with L1 and L2 Regularization, KNN model using Euclidean distance, SVC with default hyperparameters and Decision tree with default hyperparameters are trained in training, validating, and testing dataset. In specifically, Logistic Regression and KNN are standard models in binary classification; SVM is used to find the best boundary to separate both classes; and a decision tree is a simple algorithm for modelling non-linear relationships. No hyperparameter is tuned or the values tested because all models are set up to default hyperparameters. However, the number of neighbors in KNN models or “min_samples_split” and “max_depth” hyperparameters in the decision tree may be significant for future experiments.

• **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	**Logistic Regression Classifier**: F1 score on training and validating sets are 0.398 and 0.444 respectively. **Logistic Regression Classifier with L1 and L2 Regularization:** F1 score on training and validating sets are 0.389 and 0.444 respectively. **KNN model using Euclidean distance**: F1 score on training and validating sets are 0.849 and 0.717 respectively. **SVC with default hyperparameters:** F1 score on training and validating sets are 0.847 and 0.667 respectively. **Decision tree with default hyperparameters:** F1 score on training and validating sets are 0.663 and 0.429 respectively. The Logistic Regression Classifier, Logistic Regression Classifier with L1 and L2 Regularization, Decision tree and SVC models indicate poor performance with F1 scores of less than 0.7. In this experiment, KNN model using Euclidean distance is the best model that accurately predicts the target variable in both sets. It indicates that the model generalizes well but does overfit the training data. The final model (KNN) on the test set: F1 score is 0.833.
3.b. Business Impact	The F1 score of the final model is 0.833, which indicates an excellent performance of the model. Based on the confusion matrix, this model could use in business to identify customers likely to buy a new car. The impacts of the incorrect results are that the marketing team might waste efforts on ineffective campaigns, which could result in a decline in sales and revenue. In addition, the company's relationship with those who dislike marketing communications will have a negative impact.
3.c. Encountered Issues	Removing an excessive number of missing values from the gender and age group, which might result in the loss of essential data information.

• **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	For this experiment, the KNN model using Euclidean distance has performed a great result. However, it is more important to reduce the overfitting for the training data, so another experiment should be employed to solve this problem. In addition, there is an increase in the performance of the logistic regression model, but the result is not satisfactory enough to be considered a good performance outcome.
4.b. Suggestions / Recommendations	Experiment 4: Tuning the hyperparameter for decision tree model. Experiment 5: Tuning number of neighbors in the KNN model using Euclidean distance. Experiment 4 and 5 are the same rank.