EXPERIMENT REPORT

Student Name	Trung Kien Hoang
Project Name	Customer behavior models in buying a new car
Date	28/04/2023
Deliverables	<Experiment 1> < 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>

• **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 the car variables such as 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 helps to reduce running time and prevent bias model. 2. Remove model_19, which is not belong to car models (from 1 to 18). This helps to improve model performance. 3. Drop ID columns: This variable is unnecessary in binary classifications. 4. Remove all observations with missing values: This helps to improve model performance. 5. 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. 6. Data scaling: Scale the data. Constructing or formatting steps are unnecessary because all chosen variables involve car variables. However, constructing or formatting processes such as one hot encoding gender and car segment variables or transforming age groups may be necessary for future experiments.
2.b. Feature Engineering	There are no feature engineering steps performed in this experiment. However, the features such as age categories, gender, vehicle model and type of vehicle are removed from the model because they are non-numeric columns.

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 0.332 and 0.322 respectively. **Logistic Regression Classifier with L1 and L2 Regularization:** F1 score on training and validating sets are 0.332 and 0.322 respectively. **KNN model using Euclidean distance**: F1 score on training and validating sets are 0.800 and 0.722 respectively. **SVC with default hyperparameters:** F1 score on training and validating sets are 0.753 and 0.746 respectively. **Decision tree with default hyperparameters:** F1 score on training and validating sets are 0.999 and 0.788 respectively. The Logistic Regression Classifier and Logistic Regression Classifier with L1 and L2 Regularization indicate poor performance with F1 scores of less than 0.7. On the other hand, the KNN model using Euclidean distance and Decision tree with default hyperparameters performs excellent results for the training set. However, it shows the overfitting of the model to the training data. The reason for underfitting and overfitting data is imbalances in the given dataset. In this experiment, SVC with default hyperparameters is the best model that accurately predicts the target variable in both sets. It indicates that the model generalizes well and does not overfit the training data. The final model (SVC) on the test set: F1 score is 0.746.
3.b. Business Impact	The F1 score of the final model is 0.746, which indicates the good performance of the model. Based on the confusion matrix, this model could use in business to identify customers who would not be likely to buy a new car. However, it could not effectively 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	The oversampling technique is used to prevent the imbalance of the dataset. However, it makes overfitting to models.

• **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 SVC model has performed a satisfactory result. However, it is more important to identify customers who are likely to buy a new car, as demonstrated by the positive cases in this model. As a result, another experiment should be employed to solve this problem.
4.b. Suggestions / Recommendations	Experiment 2: Experiment 1 + feature engineering with gender and car segments. Experiment 3: Experiment 1 + feature engineering with gender, age group and car segments. Steps: • Find the best results of each model based on demonstrated experiments. • Tuning the hyperparameter for those models. Rank 1: Experiments 2 and 3. Rank 2: Two steps