EXPERIMENT REPORT

Student Name	Trung Kien Hoang
Project Name	Customer behavior models in buying a new car
Date	28/04/2023
Deliverables	<Experiment 5> < KNN model with the number of neighbors using Euclidean distance>

- **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 gender 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. This experiment aims to reduce the overfitting in KNN model using Euclidean distance. 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**: Train the KNN model using Euclidean distance with multiple number of neighbors. As a result, store the best model in this experiment to the evaluation section and generalize the best model for this project.

- **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” to “Small/Medium” in all cars model 17. 4. Remove all observation with missing values in gender column: This help to improve model performance. 5. Drop ID columns: This variable is not necessary in binary classifications. 6. 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.
2.b. Feature Engineering	One-hot encoding the gender and car segment variables: This help to transform categories variables to numeric variables.
2.c. Modelling	In this experiment, the KNN model using Euclidean distance is trained with a function that print F1 score for training and validating sets in KNN models with number of neighbors from 1 to 35.

<table>
<tr><th colspan="2">• EXPERIMENT RESULTS</th></tr>
<tr><td colspan="2">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.</td></tr>
<tr><td><b>3.a. Technical Performance</b></td><td>The best number of neighbors for the model is 25 where F1 score on training and validating sets are 0.820 and 0.791 respectively. This indicates that the model generalizes well and does very slightly overfit the training data.
The final model on the test set: F1 score is 0.792.</td></tr>
<tr><td><b>3.b. Business Impact</b></td><td>The F1 score of the final model is 0.792, 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.</td></tr>
<tr><td><b>3.c. Encountered Issues</b></td><td>Changing a number of neighbors one by one takes a very long time and effort. It is solved by creating a function that takes input as a number of neighbors and performs the KNN model.</td></tr>
</table>

• 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 with 25 neighbors using Euclidean distance has performed a great result. This is the final experiment conducted for this project.
4.b. Suggestions / Recommendations	Choose a model that performed exceptionally well for this project in the evaluation section, and then deploy it to the business.