EXPERIMENT REPORT

Student Name	Trung Kien Hoang
Project Name	Customer behavior models in buying a new car
Date	28/04/2023
Deliverables	<Experiment 4> < Decision tree model > <"min_samples_split" and "max_depth" 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. This experiment aims to reduce the overfitting in decision tree model. 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 decision tree with different hyperparameter to find the optimal values of “min_samples_split” and “max_depth”. As a result, store the best model in the evaluation section and proceed to a new experiment.

- **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. Drop ID columns: This variable is unnecessary in binary classifications. 4. Remove all observation with missing values: This help 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.
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, Decision tree is trained with “min_samples_split” and “max_depth” hyperparameters where “min_samples_split” values are 15, 45, 75 and 85; and “max_depth” values are 3, 8, 13 and 18.

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	**min_samples_split = 15:** F1 score on training and validating sets are 0.886 and 0.817 respectively. **min_samples_split = 45:** F1 score on training and validating sets are 0.820 and 0.789 respectively. **min_samples_split = 75:** F1 score on training and validating sets are 0.808 and 0.796 respectively. **min_samples_split = 85:** F1 score on training and validating sets are 0.801 and 0.774 respectively. **min_samples_split = 75 and max_depth = 3:** F1 score on training and validating sets are 0.435 and 0.430 respectively. **min_samples_split = 75 and max_depth = 8:** F1 score on training and validating sets are 0.710 and 0.705 respectively. **min_samples_split = 75 and max_depth = 13:** F1 score on training and validating sets are 0.808 and 0.796 respectively. **min_samples_split = 75 and max_depth = 18:** F1 score on training and validating sets are 0.808 and 0.796 respectively. In this experiment, Decision tree model with “min_samples_split” = 75 and “max_depth” = 13 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 on the test set: F1 score is 0.8.
3.b. Business Impact	The F1 score of the final model is 0.8, 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	The results of F1 score changed every time the model is executed. To solve this problem, set “random_state” = 42 to all trained models.

<table>
<tr><th colspan="2">• FUTURE EXPERIMENT</th></tr>
<tr><td colspan="2">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.</td></tr>
<tr><td><b>4.a. Key Learning</b></td><td>The Decision tree model with hyperparameters has attained the business objective for this endeavor. However, it would be preferable to conduct another experiment to reduce overfitting in the KNN model in order to compare the performance of the model with this experiment.</td></tr>
<tr><td><b>4.b. Suggestions / Recommendations</b></td><td>Experiment 5: Tuning number of neighbors in the KNN model using Euclidean distance.</td></tr>
</table>