EXPERIMENT REPORT

Student Name	Hemang Sharma
Project Name	Assignment 2 - Classification Models: Experiment 6
Date	28th April 2023
Deliverables	<notebook name: bagging.ipynb> <model name: bagging_clf>

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 is to build a predictive model using Bagging Decision Tree (Ensemble Learning I) that can classify whether an existing customer is more likely to buy a new car. The target variable is binary, where 1 indicates that the customer has purchased more than one vehicle and 0 indicates that the customer has only purchased one vehicle. The results of this project can be used by the business to target customers who are more likely to buy a new car, based on their past purchase behavior and other relevant features. This can help the business to optimize their marketing campaigns and improve customer retention. For example, the business can target customers who have only purchased one vehicle but have a high probability of purchasing another vehicle, with specific offers and promotions. The impact of accurate results can be significant for the business. Accurate predictions can help the business to identify potential customers who are more likely to make another purchase and improve the effectiveness of their marketing campaigns. This can lead to increased sales, revenue, and profitability for the business. On the other hand, incorrect results can have negative consequences for the business. Inaccurate predictions can lead to wasted marketing resources and missed opportunities to target potential customers. This can result in lower sales and revenue for the business, as well as a reduction in customer satisfaction and loyalty if customers are targeted with irrelevant offers. Therefore, it is important to ensure the accuracy and reliability of the predictive model through careful data preparation, model selection, and evaluation.

1.b. Hypothesis	The hypothesis that can be tested using this project is whether it is possible to predict whether an existing customer is more likely to buy a new car based on their past purchase behavior and other relevant features. The question I want to answer is whether a predictive model can accurately classify customers who are more likely to make another purchase based on their previous purchase history and other relevant features, such as their age, gender, car model, and other service-related variables. There are several reasons why this hypothesis is worthwhile considering: 1. Improved marketing campaigns: By accurately predicting customers who are more likely to make another purchase, businesses can tailor their marketing campaigns to target these customers with relevant offers and promotions, which can increase sales and revenue. 2. Enhanced customer retention: Identifying customers who are more likely to make another purchase can help businesses to focus on customer retention efforts and improve customer satisfaction and loyalty. 3. Cost savings: Predicting which customers are more likely to make another purchase can help businesses to optimize their marketing resources and reduce marketing costs by avoiding irrelevant marketing campaigns. Overall, this hypothesis is worth considering because it has the potential to provide valuable insights into customer behavior and can help businesses to make data-driven decisions to improve their marketing campaigns, customer retention efforts, and profitability.

1.c. Experiment Objective	The expected outcome of the experiment is to build a Bagging Decision Tree model that accurately predicts whether an existing customer is more likely to buy a new car based on their past purchase behavior and other relevant features. The goal of this experiment is to achieve a high level of accuracy in the model's predictions. While the exact target accuracy can vary depending on the business needs and the nature of the problem, a realistic goal could be to achieve an accuracy of 85% or higher. Possible scenarios resulting from this experiment include: 1. Accurate prediction: If the model accurately predicts customers who are more likely to make another purchase, the business can use these predictions to optimize their marketing campaigns and customer retention efforts, resulting in increased sales and revenue. 2. Overfitting: Overfitting occurs when a model is too complex and fits the training data too well, but fails to generalize well to new data. This can result in poor performance on the test data and inaccurate predictions. To prevent overfitting, it is important to use techniques such as cross-validation and regularization. 3. Underfitting: Underfitting occurs when a model is too simple and fails to capture the underlying patterns in the data. This can also result in poor performance on the test data and inaccurate predictions. To address underfitting, it may be necessary to use more complex models or increase the number of features used in the model. 4. Data quality issues: If the data used to train the model is of poor quality or contains errors, the model's performance may be affected. It is important to carefully clean and preprocess the data before training the model to ensure that it is of high quality. Overall, the experiment's outcome will depend on the quality of the data, the choice of model, and the hyperparameters used to train the model. By carefully selecting the right model and parameters, and using techniques such as cross-validation to assess model performance, it is possible to build a robust and accurate predictive model that can provide valuable insights into customer behavior.

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 following steps were taken for preparing the data: 1. Label encoding was performed on categorical features - age_band, gender, car_model, and car_segment. This step was performed to convert the categorical features into numerical format that can be used in the machine learning model. 2. The 'ID' and 'Target' columns were dropped from the feature set (X) as 'ID' is not useful for the analysis and 'Target' is the variable I are trying to predict. The above steps were sufficient for the purpose of building the Bagging Decision Tree model.
2.b. Feature Engineering	No explicit feature generation was performed.

2.c. Modelling	In this experiment, a Bagging Decision Tree model was trained to predict whether an existing customer is more likely to buy a new car. The Bagging Decision Tree is an ensemble learning technique that combines multiple decision trees, with each tree trained on a random subset of the training data. The final prediction is based on the majority vote of all the trees in the ensemble. The `DecisionTreeClassifier` is used as the base estimator, which is a simple yet powerful model that can capture non-linear relationships between the features and the target variable. Bagging helps to reduce the variance of the model and improve its generalization performance. The following hyperparameters were tuned in the BaggingClassifier: - `n_estimators`: the number of decision trees in the ensemble. Values tested: [5, 10, 15, 20, 25]. - `max_samples`: the maximum number of samples used to train each tree. Values tested: [0.5, 0.6, 0.7, 0.8, 0.9, 1.0]. - `max_features`: the maximum number of features used to train each tree. Values tested: [0.5, 0.6, 0.7, 0.8, 0.9, 1.0]. These hyperparameters were selected for tuning because they can significantly impact the performance of the model and can help to control overfitting. Other models that could be tested include Random Forest, AdaBoost, Gradient Boosting, and XGBoost. These models are also ensemble techniques that can improve the performance of decision trees. However, Bagging Decision Tree was chosen for this experiment as it is simple to implement, interpretable, and computationally efficient. In terms of the base estimator, other models that could be tested include Support Vector Machines (SVM), Logistic Regression, Naive Bayes, and Neural Networks. These models are more complex than decision trees and can capture more complex relationships between the features and the target variable. However, decision trees were chosen as the base estimator for this experiment because they are simple to interpret, computationally efficient, and can capture non-linear relationships between the features and the target variable. In terms of hyperparameters, future experiments could explore different values for the hyperparameters tested in this experiment, as well as other hyperparameters specific to different models. Additionally, other ensemble techniques could be explored, such as Stacking and Blending, which combine multiple models with different types of base estimators.

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 accuracy of the Bagging Decision Tree model on the test set is 0.993, indicating a high level of performance. The confusion matrix shows that the model correctly predicted 25,562 out of 25,608 negative cases (repurchase = 0) and 521 out of 660 positive cases (repurchase = 1). However, there were 46 false positives (predicted as repurchase = 1 but actually repurchase = 0) and 139 false negatives (predicted as repurchase = 0 but actually repurchase = 1). The precision of the positive class (repurchase = 1) is 0.92, which means that when the model predicts a customer will repurchase, it is correct 92% of the time. The recall of the positive class is 0.79, which means that the model correctly identifies 79% of the customers who actually repurchased. The f1-score for the positive class is 0.85, which is a weighted harmonic mean of the precision and recall. The high number of false negatives may indicate that the model is not sensitive enough in detecting customers who are likely to repurchase. This may be due to the imbalanced nature of the dataset, where there are many more negative cases than positive cases. One potential solution to address this is to use techniques such as oversampling or undersampling to balance the dataset. On the other hand, the high number of false positives may indicate that the model is too sensitive in predicting customers who are likely to repurchase, leading to unnecessary marketing efforts and costs. This may require further analysis on the cost-benefit of false positives and false negatives to determine the optimal trade-off. Overall, the model's performance is quite good, but there is room for improvement in detecting customers who are likely to repurchase.
3.b. Business Impact	Based on the results of the experiment, the Bagging Decision Tree model achieved a high accuracy of 99.3%, indicating that it is a strong model for predicting repurchases. The precision and recall for class 1 (repurchased) were also relatively high at 0.92 and 0.79 respectively. This indicates that the model was able to accurately identify a majority of the customers who are likely to repurchase. However, there were still a few false negatives, meaning that some customers who are likely to repurchase were not identified by the model. This could potentially result in missed opportunities for the business to target these customers with retention efforts. On the other hand, there were a few false positives, meaning that some customers who are not likely to repurchase were identified by the model. While this may result in some wasted resources on retention efforts, it is less harmful than missing out on potential repurchases. The impact of incorrect results for the business can vary. If the model predicts that a customer is not likely to buy a new car when they actually are, the business may miss out on potential sales and revenue. On the other hand, if the model predicts that a customer is likely to buy a new car when they actually are not, the business may spend resources on marketing and sales efforts that do not result in a sale. Overall, the model can be used as a tool to help the business identify which customers are more likely to buy a new car and tailor their marketing and sales efforts accordingly. However, it is important to note that the model is not perfect and there may be cases where the model misclassifies a customer. Therefore, it is recommended that the model's predictions be used in conjunction with human judgment and other sources of information to make informed business decisions.

3.c. Encountered Issues	During the experiment, some issues were encountered, including: 1. Data cleaning: One of the main issues was cleaning the data, which included missing values, incorrect data types, and inconsistent data. To solve this issue, various data cleaning techniques were applied, such as imputing missing values and converting data types. 2. Feature selection: There were many features in the dataset, some of which may not be relevant to the business objective. Feature selection techniques were used to identify the most important features. However, there may be some features that were not identified as important in this experiment but could be important in future experiments. 3. Imbalanced data: The dataset was imbalanced, with a higher number of customers who purchased only one vehicle compared to those who purchased more than one vehicle. This can lead to biased model performance. Techniques such as oversampling or undersampling can be used to balance the data. 4. Hyperparameter tuning: The hyperparameters of the models used in the experiment needed to be tuned for optimal performance. Grid search or random search can be used for hyperparameter tuning. However, this can be time-consuming and computationally expensive. 5. Model evaluation: The accuracy score was used to evaluate model performance. However, accuracy alone may not be enough to evaluate the model's performance, especially for imbalanced datasets. Other metrics such as precision, recall, and F1-score should also be considered. 6. Interpretability: Some models, such as ensemble methods, can be difficult to interpret. Model interpretability is important for business decisions and should be considered when selecting a model. Overall, these issues were solved or mitigated in this experiment. However, they should be considered in future experiments to ensure accurate and reliable results.

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	Based on the experiment's results, the Bagging Decision Tree model achieved a high accuracy score of 0.993, suggesting that it can effectively predict whether a customer will repurchase the product. The model's precision score for predicting the positive class is also high, indicating that it is capable of correctly identifying customers who will repurchase the product. However, the model's recall score for predicting the positive class is relatively low, indicating that the model is less successful at identifying all customers who will repurchase the product, potentially leading to missed opportunities to target these customers with tailored marketing campaigns. One potential explanation for the model's lower recall score could be the class imbalance in the data, where there are significantly more customers who did not repurchase the product compared to those who did. This can lead to the model being biased towards predicting the majority class and underperforming on the minority class. Overall, the results suggest that the Bagging Decision Tree model can be a useful tool for predicting repurchase behavior in customers, but there is room for further improvement in terms of identifying all customers who will repurchase. Therefore, it may be worth exploring different modeling techniques or techniques to handle class imbalance, such as oversampling or undersampling the minority class, in future experiments.
4.b. Suggestions / Recommendations	Based on the results achieved and the overall objective of the project, the following potential next steps and experiments can be considered: 1. Feature engineering: Further feature engineering can be done to identify more relevant features that can improve the performance of the model. For example, new features such as customer loyalty, purchase frequency, and customer lifetime value can be added. 2. Model tuning: Additional model tuning can be performed to improve the model's performance further. Hyperparameters such as the number of estimators, max depth of the decision tree, and learning rate can be further optimized. 3. Model selection: Alternative models such as Random Forest, Gradient Boosting, or Neural Networks can be evaluated to improve the model's performance. 4. Data collection: Additional data points such as customer satisfaction scores or product reviews can be collected to add to the model's predictive power. 5. Deployment: If the model's performance is satisfactory, it can be deployed into production to provide real-time predictions for the business. This would require integrating the model into the business's IT infrastructure and implementing a process for updating the model regularly. In terms of ranking, feature engineering and model tuning are likely to provide the highest potential uplift in the model's performance. Model selection and data collection can also provide improvements but may require more effort and resources. Finally, deployment is a crucial step that requires careful planning and consideration of the business's IT infrastructure. Overall, based on the results achieved in this experiment, it is recommended to pursue more experimentation with the current approach to further improve the model's performance.