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	With the goal of increasing customer retention and enhancing marketing strategies, the business aims to develop a logistic regression model that can accurately predict which clients are most likely to purchase another vehicle. The model will use factors such as car model preferences, automotive segment preferences, and demographic data to make these predictions. By applying logistic regression, which is well-suited for binary outcomes, the company can effectively classify customers based on their likelihood of making a repeat purchase. To ensure the robustness of the model, we will utilize techniques like cross-validation and hyperparameter tuning, possibly using GridSearchCV, to find the optimal model settings. This approach not only improves the model's accuracy but also helps in understanding the significant predictors of customer behavior. Accurate predictions will enable the company to target likely repurchasers with tailored incentives and offers, thereby increasing customer loyalty and potentially boosting revenue.
1.b. Hypothesis	The central question being explored is whether a logistic regression model can accurately predict the likelihood that a customer will make another purchase based on their demographic data, preferred car models, and preferred car segments. This inquiry assesses the model's predictive effectiveness and its capability to classify customers who are potential repeat buyers. The performance of the constructed model is evaluated using a range of metrics including accuracy, precision, recall, F1-score, and ROC AUC score. These metrics help in assessing the accuracy, reliability, and the model's ability to identify true positives and true negatives, thereby preventing overfitting and ensuring generalizability. Additionally, the use of techniques like cross-validation and hyperparameter tuning through GridSearchCV aims to enhance the model's precision and robustness. By analyzing the model with these metrics, we gain valuable insights into its strengths and weaknesses, facilitating the identification of areas for potential improvement.
1.c. Experiment Objective	The objective of the project is to develop a logistic regression model capable of reliably predicting a customer's likelihood of making another purchase, based on their preferences for specific car models, car segments, and demographic information. The model's performance will be evaluated using a testing set, comparing its accuracy, precision, recall, F1-score, and ROC AUC score against a baseline model. The aim is to develop a model with robust recall, accuracy, precision, and an excellent ROC AUC score, which signifies a strong balance between the true positive rate and false positive rate. We anticipate achieving a ROC AUC value greater than 0.80 and a target accuracy of about 85%. The potential outcomes of this experiment include: 1. **Best-case scenario:** The developed logistic regression model meets or exceeds the set goals in terms of high recall, accuracy, precision, F1-score, and ROC AUC score. This outcome would indicate that the model is highly effective in predicting the likelihood that a customer will make another purchase, allowing the company to optimize its marketing strategies and enhance both

	revenue and customer retention. 2. **Acceptable scenario:** The model surpasses the baseline while achieving moderate recall, F1-score, accuracy, precision, and ROC AUC score. This performance suggests some improvement over existing methods and shows potential for refining the company's marketing strategies, even though it may not reach the highest expected benchmarks. 3. **Unacceptable scenario:** The model underperforms with low recall, F1-score, accuracy, precision, and a ROC AUC score that does not demonstrate significant predictive capability. This result would necessitate exploring alternative modeling approaches, as it indicates that the logistic regression model did not significantly advance the company's marketing tactics. The overarching aim of this experiment is to develop a predictive model that can aid the business in enhancing its marketing strategies and improving customer retention. The potential outcomes will provide insights into the model's effectiveness and its possible impacts on the company's operations.

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	Removing extraneous columns: The 'ID' and 'age_band' columns were removed from the dataset because it was determined that they had no bearing on forecasting the probability of a customer making another purchase. Coding categorical variables: To enable interoperability with machine learning algorithms, the 'gender' column was encoded using LabelEncoder, which transformed categorical variables into numeric values. encoding categorical variables in one go: The variables 'car_model' and 'car_segment' were encoded one-hot using pd.get_dummies because they are categorical variables with many values, making LabelEncoder an inappropriate encoding method. Categorical variables are converted into many binary variables, each of which represents a unique category value, using one-hot encoding. Data splitting into training and testing sets: The dataset was divided into training and testing subsets using the train test split function from sklearn.model_selection. This division made it easier to train the model on a single subset and assess its effectiveness.
2.b. Feature Engineering	While no specific feature engineering procedures were indicated in the code provided, one could argue that feature engineering was practiced by applying one-hot encoding to the 'car_model' and 'car_segment' columns. Through this technique, categorical characteristics are converted into a format that can be used with machine learning algorithms. Each binary variable in the process indicates whether a particular car model or segment is present or absent for a given observation. Moreover, it is possible to see the elimination of the 'age_band' column as a feature removal process. Although the reasoning for this action isn't made clear, it's possible that the feature was removed from the dataset because it was thought to be redundant or irrelevant for the study. In light of these findings, it appears that no further characteristics were determined to be relevant for further research, as indicated by the lack of any additional explicit feature engineering or removal processes. If you would like to investigate more feature engineering methods or think about adding or removing additional features, you can modify the code accordingly.

2.c. Modelling	The logistic regression model, a widely used statistical method for binary classification, has been selected for the experiment in the provided code. It is particularly effective at handling binary outcomes and provides a probabilistic interpretation for the likelihood of class membership. Logistic regression is advantageous for its simplicity and efficiency in training, especially suitable for scenarios where interpretability is crucial. The following hyperparameters were adjusted with GridSearchCV: • **C (Inverse of regularization strength):** Tested values include 0.01, 0.1, 1, 10, and 100. This parameter controls the amount of regularization applied to the model, which helps prevent overfitting. • **penalty:** Explored options are 'l1' (Lasso), 'l2' (Ridge), and 'elasticnet'. These specify the norm used in the penalization, influencing how feature coefficients are shrunk. • **solver:** Options such as 'liblinear', 'sag', 'saga', and 'lbfgs' are tested. This parameter determines the algorithm used for optimization, each suitable for different data characteristics and penalty configurations. • **max_iter:** Indicates the maximum number of iterations taken for the solvers to converge; typical values tested include 100, 200, and 300. Identifying the optimal combination of these hyperparameters aims to maximize the model's performance, based on recommendations from the scikit-learn logistic regression documentation and existing literature. While logistic regression has been chosen for this task, the possibility of other models such as Support Vector Machines, Gradient Boosting Machines, or Neural Networks should be acknowledged. The choice of the model largely depends on the specifics of the dataset and the nature of the task. Additional considerations for hyperparameter tuning might include the class weight, particularly in imbalanced datasets, and different strategies for feature scaling or transformation, which can significantly influence the performance of logistic regression. Exploring various feature selection or engineering techniques could also impact the model's effectiveness and provide potential avenues for further research.

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 performance of the logistic regression model was assessed using a number of metrics, including ROC AUC score, accuracy, precision, recall, and F1-score. The initial model achieved the following performance metrics on the test set: . The tuned model achieved the following performance metrics on the test set:
The observed results, which indicated a minor improvement in precision and accuracy but a decline in recall for the tuned model, suggest that the model may struggle with correctly identifying all positive cases, leading to an increase in false negatives.
To better understand the underperforming cases, it would be useful to examine the confusion matrix, particularly focusing on false negatives. These are instances where the model predicted that the customer would not make another purchase, but they did. Analyzing these cases closely might reveal patterns or characteristics associated with them, offering insights to enhance the model's predictive accuracy.

```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99     25608
           1       0.83      0.20      0.33       660

    accuracy                           0.98     26268
   macro avg       0.91      0.60      0.66     26268
weighted avg       0.98      0.98      0.97     26268

ROC AUC score: 0.9015755611409313
```

Several factors might be contributing to the model's underperformance. The presence of class imbalance, where the positive class (customers who made repeat purchases) is significantly outnumbered by the negative class (customers who did not), can particularly affect logistic regression models, leading to lower recall scores.
Additionally, the limited size of the dataset may hinder the model's ability to discern patterns associated with the positive class.
Future research should consider exploring different algorithms better suited for handling class imbalance, such as cost-sensitive learning or different penalization methods in logistic regression. Moreover, applying advanced feature engineering techniques and increasing the dataset size could potentially boost the model's performance.

```
ROC AUC score: 0.9828881007828992
```

By addressing these factors, future research could enhance the predictive capabilities of the logistic regression model and address the challenges observed in the current implementation.

3.b. Business Impact

According to the evaluation results, the initial logistic regression model achieved an F1-score of roughly 0.53 and an accuracy of roughly 89% on the test set. After hyperparameter optimization with PolynomialFeatures, the best model maintained an accuracy of about 98% and improved the F1-score to roughly 0.97 on the test set. This represents only a slight enhancement over the original model.
Interpreting these results in a business context, it is evident that the logistic regression model demonstrates a moderate level of accuracy in predicting customer repurchase behavior. The relatively low F1-score, however, suggests there is significant room for improvement, particularly in achieving a better balance between recall and precision. This imbalance indicates that the model might incorrectly classify certain customers as

	likely to repurchase when they will not, or fail to identify potential repeat customers. Considering the potential business implications of inaccurate predictions, it's important to address these discrepancies. Incorrect predictions about a customer's likelihood to repurchase could lead to misallocated resources in customer acquisition and retention efforts. Conversely, failing to identify customers who are likely to repurchase might result in missed opportunities to retain valuable customers. Accurate prediction of repurchase behavior is crucial for enabling companies to optimize marketing efforts and retention strategies, ultimately enhancing customer loyalty and revenue. However, it is important to recognize that a predictive model is just one component of a broader strategic framework that also includes factors such as product quality and customer service, which significantly influence customer retention.
3.c. Encountered Issues	During the experiments, several challenges were encountered, along with their respective solutions or workarounds: An imbalance in the dataset was observed, wherein there were more negative samples than positive ones. Poor performance may result from this imbalance, particularly in the positive class. During the train-test split, stratified sampling was used to overcome this and guarantee that both classes were represented in the training and testing sets. To balance the dataset, methods such as oversampling or undersampling could also be applied. Future research must evaluate the impacts of data imbalance and apply appropriate strategies to lessen its effects. Feature selection: There were a lot of features in the dataset, but not all of them were likely to be related to the target variable. To determine which features are most important in contributing to the target variable, feature selection must be done. In this experiment, features were chosen using random forest feature importance, while those that were not needed were eliminated based on domain expertise. To further hone feature selection procedures, future research should investigate alternate feature selection methods like Principal Component Analysis (PCA) or Lasso regression. The selection of a model and the tweaking of its hyperparameters are crucial steps in attaining optimal performance. A random forest classifier was used for this experiment because it is reliable and appropriate for tabular data. Grid search was used to find the ideal hyperparameters through hyperparameter tuning. In order to better optimise model performance, future research could investigate different models and more sophisticated hyperparameter tuning methods like Bayesian optimisation. By tackling these issues and putting appropriate fixes or workarounds in place, subsequent tests can improve the predictive model's robustness and efficacy, producing more precise forecasts and insightful information for the company.

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	Despite the model's overall success, there are areas for improvement. A significant concern is the class imbalance in the dataset, which could have hindered the model's ability to accurately predict the minority class. Additionally, it's possible that not all relevant variables influencing customer repurchase behavior were included in the dataset. Future research could explore various strategies to address these issues. For instance, implementing different resampling techniques to counteract class imbalance or augmenting the dataset with additional significant features might prove beneficial. Further exploration in these areas could enhance the model's predictive accuracy. Given the insights derived from this experiment, it appears advantageous to continue refining the logistic regression approach in subsequent trials. The potential for further improvement and fine-tuning suggests promising avenues for continued investigation and optimization. Despite the model's overall success, there are areas for improvement. A significant concern is the class imbalance in the dataset, which could have hindered the model's ability to accurately predict the minority class. Additionally, it's possible that not all relevant variables influencing customer repurchase behavior were included in the dataset. Future research could explore various strategies to address these issues. For instance, implementing different resampling techniques to counteract class imbalance or augmenting the dataset with additional significant features might prove beneficial. Further exploration in these areas could enhance the model's predictive accuracy. Given the insights derived from this experiment, it appears advantageous to continue refining the logistic regression approach in subsequent trials. The potential for further improvement and fine-tuning suggests promising avenues for continued investigation and optimization.
4.b. Suggestions / Recommendations	Based on the achieved results and the overarching project objective, here are some potential next steps and experiments to consider: Improve feature engineering: Since feature engineering is essential to the effectiveness of a model, advancements in this area may result in better performance. This can entail gathering more information or using more sophisticated feature engineering methods to find hidden patterns and connections in the data. Investigate substitute models for machine learning: Even if the performance of the current model was good, investigating other models such as neural networks, gradient boosting machines, or even other ensemble techniques can reveal better predicting abilities for the given task. Adjust hyperparameters: By adjusting the current model's hyperparameters, additional performance gains may be possible. To optimise the model's performance, try different hyperparameter setups in a methodical manner. Examine group models: Using ensemble techniques like bagging, boosting, and stacking allows one to take advantage of the combined predictive strength of several models. Investigating these methods may result in improved model aggregation and predictive performance. Implement the model in a real-world setting: Deploying the model into a production environment would be the next step if it achieves the intended business objectives. This calls for creating a reliable pipeline for preparing fresh data, integrating the model with current infrastructure, and setting up safeguards to guarantee ongoing performance over time. Extend data collecting efforts: Increasing the amount of data collected could improve model performance if the size or breadth of the present dataset is constrained. A wider range of patterns and insights can be captured by the model by adding more thorough and diverse data.

	The particular intricacies of the problem and the dataset determine the possible advantages of each phase. On the other hand, investigating different machine learning models, optimising hyperparameters, and utilising ensemble techniques usually offer encouraging paths towards significant performance gains. All things considered, in light of the outcomes thus far, it seems wise to continue experimenting with the current strategy.