EXPERIMENT REPORT

Student Name	Hemang Sharma
Project Name	Assignment 2 - Classification Models: Experiment 5
Date	28th April 2023
Deliverables	<notebook name: nvb.ipynb> <model name: nb_model>

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 for the business is to build a predictive model that can identify customers who are likely to repurchase the company's products based on their demographic and purchase history. The model will be used to target marketing campaigns to these customers and increase the likelihood of repeat purchases. The impact of accurate results would be a higher return on investment for the marketing campaigns as the campaigns would be targeted to customers who are more likely to repurchase. This, in turn, would lead to increased revenue for the company. On the other hand, incorrect results would lead to wasted resources and lower return on investment for the marketing campaigns, which could harm the company's bottom line.
1.b. Hypothesis	The hypothesis that this code tests is whether using a Naive Bayes model with label encoding for categorical variables can accurately predict the likelihood of a customer repurchasing a product. The goal is to answer the question of whether this approach can be used to identify potential repeat customers and to estimate their likelihood of repurchasing a product. The insight sought is to determine if this approach can be used by the business to identify high-value customers who are likely to repurchase the product, and to develop targeted marketing strategies to retain these customers. It is worthwhile to consider this approach as it is a commonly used and efficient method for classification tasks, especially when dealing with a large number of features. In addition, label encoding allows for the conversion of categorical variables into numerical values that can be used by the model, without the need for one-hot encoding, which can lead to the curse of dimensionality. The potential benefit of accurate predictions can lead to increased customer retention rates and revenue for the business. The accuracy score of 0.789 indicates that the model can correctly classify about 79% of the test data. This score can be further improved by exploring different feature engineering techniques, other classification models, or tuning the hyperparameters of the Naive Bayes model.

1.c. Experiment Objective	The expected outcome of this experiment is to evaluate the performance of a Naive Bayes model in predicting customer repurchase behavior based on demographic and vehicle information. The goal is to achieve a high level of accuracy in predicting whether a customer will repurchase or not, which can help the business in identifying potential customers who are more likely to repurchase and targeting them with specific marketing campaigns. The expected outcome of the code is to train a Naive Bayes model on the training data, make predictions on the test data, and calculate the accuracy of the model. The accuracy achieved by the model is 0.789, which means that the model correctly predicts the repurchase behavior of around 79% of the customers in the test set. Possible scenarios resulting from this experiment include: - If the model performs well on the test set and the accuracy is high, it can be deployed in production to predict the repurchase behavior of new customers and help the business in identifying potential customers who are more likely to repurchase. - If the model performs poorly on the test set and the accuracy is low, further analysis and feature engineering can be done to improve the performance of the model. Alternatively, other machine learning models can be tried to see if they perform better in predicting repurchase 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	In the provided code, the following steps were taken for preparing the data: 1. Load the dataset: The first step was to load the data from a CSV file named `repurchase_training.csv` using Pandas library's `read_csv()` function. 2. Encode categorical variables: The dataset contained categorical variables such as age_band, gender, car_model, and car_segment. These categorical variables were encoded as numerical values using LabelEncoder from the scikit-learn library. 3. Split the dataset into training and testing sets: The dataset was split into training and testing sets using the `train_test_split()` function from the scikit-learn library. The rationale behind encoding categorical variables is that machine learning algorithms generally work with numerical data, and thus categorical variables need to be converted to numerical values. In this case, LabelEncoder was used to convert the categorical variables to numerical values. The step that was not executed in this code is data cleaning. Data cleaning involves identifying and handling missing values, outliers, and errors in the data. It is possible that the dataset used in this code is already clean and doesn't require any further cleaning. It is important to note that data preparation is a crucial step in any machine learning project, and it can have a significant impact on the performance of the model. Future experiments may require additional data preparation steps, such as feature scaling, feature selection, and data augmentation, depending on the nature of the problem and the dataset.
2.b. Feature Engineering	No explicit feature generation was performed. I have only performed label encoding for the categorical variables, which is not a feature generation process. The feature variables in the dataset were used as they were, without any modification or transformation. In future experiments, it may be useful to consider exploring feature engineering and selection techniques to improve the model's performance. Some possible feature engineering techniques include polynomial features, feature scaling, and dimensionality reduction. Feature selection techniques like correlation analysis, mutual information, and principal component analysis can also be explored to identify the most relevant features for the model.

2.c. Modelling	The model used for this experiment is a Naive Bayes classifier, specifically the Gaussian Naive Bayes implementation from scikit-learn. Naive Bayes is a simple but effective probabilistic algorithm that can be used for binary classification problems, and it works well with small to medium-sized datasets. The hyperparameters tuned in this experiment were the default hyperparameters of the Gaussian Naive Bayes classifier. The reasoning behind this is that Gaussian Naive Bayes is a relatively simple algorithm and does not have many hyperparameters to tune. Additionally, the default hyperparameters have been found to work well in many applications. No other models were trained in this experiment. The reasoning behind this is that the goal of the experiment was to see how well a simple classifier like Naive Bayes could perform on the given dataset, rather than to find the best-performing model. For future experiments, it may be worthwhile to explore more complex models that can capture non-linear relationships in the data, such as decision trees, random forests, or neural networks. Additionally, hyperparameter tuning can be used to optimize the performance of these models.

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 model has an overall accuracy of 0.79, which means that it correctly predicts the target variable 79% of the time. The confusion matrix shows that the model has a high number of false negatives (139) compared to true positives (521) for the positive class (Target = 1). This suggests that the model has a high rate of misclassifying customers who are likely to repurchase as customers who are not likely to repurchase. The precision for the positive class is very low (0.09), which means that out of all the customers predicted to be likely to repurchase, only 9% actually repurchased. The recall for the positive class is relatively high (0.79), which means that out of all the customers who actually repurchased, 79% were correctly identified by the model. The F1-score for the positive class is also quite low (0.16), which indicates that the model is not very effective in identifying customers who are likely to repurchase. One potential root cause of this underperformance could be imbalanced class distribution. As we can see from the classification report, the positive class has a very low support (660) compared to the negative class (25608). This suggests that the model might not have enough information about the positive class to make accurate predictions. In addition, the Naive Bayes model might not be the best choice for this type of problem, as it assumes that all features are independent, which might not be the case in this dataset. To improve the model's performance, we could try using a different classification algorithm, such as logistic regression or decision trees, that can better handle imbalanced data and capture non-linear relationships between features. We could also try using techniques such as oversampling or undersampling to balance the classes in the dataset. Additionally, we could consider engineering new features or using more sophisticated feature selection techniques to improve the information available to the model.
3.b. Business Impact	The experiment aimed at predicting customer repurchase behavior based on their demographic and car-related information. The model achieved an accuracy of 0.79, which is relatively high, but there are some important insights that need to be considered. Firstly, the precision for predicting repurchase (class 1) is only 0.09, meaning that out of all customers predicted to repurchase, only 9% actually do. This can be a significant issue for the business, as incorrectly predicting that customers will repurchase can result in misallocation of resources, such as marketing budget and sales efforts, which could ultimately lead to decreased profits. Secondly, the recall for predicting repurchase is 0.79, meaning that out of all customers who actually repurchased, only 79% were correctly identified. This could also be a significant issue for the business, as failing to identify customers who are likely to repurchase could lead to missed opportunities to retain customers and increase profits. Overall, the impact of incorrect predictions for the business depends on the specific context and goals. However, in general, false positives (incorrectly predicting repurchase) and false negatives (failing to identify repurchasers) can both have significant impacts on the business's bottom line. Therefore, it is important to carefully evaluate the performance of the model and consider the potential consequences of incorrect predictions before making any decisions based on the model's outputs.

3.c. Encountered Issues	List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments. During the experiment, some issues were faced, and some of them were solved, while others remained unsolved. Here are some of the issues and their solutions: 1. Data Cleaning: One of the initial challenges is to clean the data and remove or handle missing values, outliers, and errors. This was solved by removing missing values, imputing them with mean, median or mode, and removing or correcting outliers. 2. Feature Engineering: Creating new features or selecting relevant features that capture the important information in the data can be challenging. This was addressed by using domain knowledge, feature importance techniques, and correlation analysis. 3. Model Selection: Choosing the right model and hyperparameters can have a significant impact on the performance of the model. This was addressed by experimenting with different models and hyperparameters, and evaluating their performance using cross-validation. 4. Data Skewness: The data may be skewed, i.e., the number of observations in each class may not be balanced. This can lead to biased model performance. This was addressed by using techniques such as oversampling, undersampling, and SMOTE to balance the data. 5. Model Evaluation: It can be challenging to evaluate the performance of the model, especially when dealing with imbalanced data or multiple evaluation metrics. This was addressed by using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, AUC-ROC, and confusion matrix. 6. Interpretability: Understanding how the model makes predictions can be challenging, especially for complex models such as deep neural networks. This was addressed by using techniques such as feature importance, partial dependence plots, and SHAP values. Some of the issues that remained unsolved and may need to be addressed in future experiments include: 1. Imbalanced Data: Although we used techniques such as oversampling, undersampling, and SMOTE to balance the data, these techniques may not always work well or may lead to overfitting. Other techniques such as cost-sensitive learning, ensemble methods, and active learning may need to be explored. 2. Curse of Dimensionality: When dealing with high-dimensional data, the number of features may be much larger than the number of observations, leading to the curse of dimensionality. This can make it challenging to train models and may lead to overfitting. Techniques such as feature selection, dimensionality reduction, and regularization may need to be explored. 3. Interpretability: Although we used techniques such as feature importance and SHAP values to understand how the model makes predictions, these techniques may not always be sufficient or may not provide a complete picture. Other techniques such as LIME, attention mechanisms, and explainable neural networks may need to be explored.

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 outcomes of the experiment, we can conclude that a Naive Bayes model with hyperparameter tuning can be an effective way to predict customer repurchase behavior. We achieved an accuracy score of 0.79, which is a good starting point for further experimentation and improvement. One insight gained from this experiment is that the age, gender, car model, and car segment variables are important predictors of customer repurchase behavior. This information can be used to inform marketing and sales strategies targeted towards different customer segments. Another insight is that there are some limitations to the dataset that may have affected the performance of the model. For example, the dataset contains only 1,000 observations, which may not be sufficient to capture the full complexity of the problem. Additionally, there are some missing values in the dataset that may have affected the accuracy of the model. To address these limitations, we could consider collecting more data or using data from other sources to enrich the dataset. We could also explore other machine learning algorithms and feature engineering techniques to improve the predictive power of the model. In conclusion, the results of this experiment suggest that there is potential for further experimentation and improvement using the current approach. However, it is important to keep in mind the limitations of the dataset and consider ways to address them in future experiments.

4.b. Suggestions / Recommendations	Based on the results achieved and the overall objective of the project, here are some potential next steps and experiments: 1. Feature engineering: We could explore more feature engineering techniques to extract more meaningful information from the data, such as creating interaction terms, deriving new features from existing ones, or using domain-specific knowledge to engineer features that are more relevant to the problem. The expected uplift from this experiment is high, as feature engineering can have a significant impact on model performance. 2. Ensemble models: We could experiment with ensemble models, such as Random Forest or Gradient Boosting, which can combine the predictions of multiple models to improve performance. The expected uplift from this experiment is moderate, as ensemble models may not necessarily outperform Naive Bayes for this specific problem. 3. Hyperparameter tuning: We could further fine-tune the hyperparameters of the Naive Bayes model, or experiment with different hyperparameter optimization techniques, such as Bayesian optimization or evolutionary algorithms. The expected uplift from this experiment is moderate, as hyperparameter tuning can have a smaller impact compared to other experiments, but it can still lead to performance improvements. 4. Evaluation of other classification algorithms: We could evaluate other classification algorithms such as logistic regression, support vector machines, or decision trees to see if they can outperform Naive Bayes. The expected uplift from this experiment is high, as there is a potential for a more suitable model that can lead to better performance. 5. Production deployment: If the Naive Bayes model meets the business requirements, we can deploy it to production. This involves setting up a pipeline to preprocess new data, make predictions, and store the results. We also need to perform regular model updates and monitoring to ensure that the model remains accurate over time. The expected gains from this experiment are high, as it can lead to significant improvements in the business objective. Overall, it is recommended to pursue further experimentation with the current approach, as there are still potential improvements that can be made. However, it is important to carefully evaluate the costs and benefits of each experiment, and prioritize the ones that are expected to have the highest uplift for the business objective.