# **EXPERIMENT REPORT 5**

|  |  |
| --- | --- |
| **Student Name** | Ngoc Quang Vinh Pham |
| **Project Name** | Predicting Customer Car Repurchase Likelihood |
| **Date** | April 20, 2024 |
| **Deliverables** | 36106-AT2-25100660-experiment-5.ipynb  Extreme Gradient Boosting Parameter Tuning |

|  |  |
| --- | --- |
| 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** | This study seeks to predict which consumers are most likely to make another automobile purchase, which will improve the accuracy of focused marketing activities. By concentrating efforts on clients who have a greater tendency to purchase, accurate forecasts may dramatically save marketing expenses and boost sales efficiency. On the other side, inaccurate outcomes might result in lost opportunities or resource waste. |
| **1.b. Hypothesis** | We hypothesize that optimizing the hyperparameters of the Extreme Gradient Boosting model, specifically focusing on the learning rate, number of trees, max depth, and min child weight, will significantly improve the model's ability to predict customer repurchase behavior. These adjustments are expected to enhance the model's sensitivity to complex patterns in the data, thereby increasing both the precision and recall of predictions. The hypothesis is grounded in the belief that a more finely tuned model will more accurately identify customers with a high propensity to repurchase, thus optimizing marketing strategies and resource allocation. |
| **1.c. Experiment Objective** | This experiment aims to improve the prediction performance of the Extreme Gradient Boosting model, particularly for identifying possible repeat consumers, by adjusting hyperparameters. In order to increase the F3 score, we want to determine the optimal values for the learning rate, number of trees, maximum depth, and minimum child weight. This will make it possible for marketing strategies to target customers more effectively, maximizing resource allocation and boosting return on investment. |

|  |  |
| --- | --- |
| 1. **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** | Gender Binary Encoding: We encoded the gender variable as binary (Male = 1, Female = 0), facilitating its use in our models. To preserve the integrity of the dataset and prevent bias introduction, missing values in gender were imputed using the current distribution (58% Male).  Dropping ‘age\_band’: This feature was removed due to the high prevalence of missing values in ‘age\_band’ (85%). The rationale for this was that imputing such a vast amount of missing data could lead to some significant inaccuracies and unreliable predictions.  One-hot encoding: Since there is only 4 unique values in ‘car-segment’, I decided to encode ‘car-segment’ into multiple binary variables. This method is required to handle categorical data, ensuring that the model correctly interprets these categories as distinct without any inherent order.  ‘Car\_model’ Frequency Encoding: To handle the large correlation of this feature, car\_model received frequency encoding. Compared to one-hot encoding, this method reduces dimensionality while preserving information regarding model popularity, which may be indicative of repurchase behavior. |
| **2.b. Feature Engineering** | In this stage, we did not add new characteristics to our dataset. This choice was made to first evaluate the baseline predictive potential of existing characteristics without adding complexity. Features that have more than 60% null values will be removed from the dataset due to the possibility of inaccurate and incorrect imputation of these characteristics.  In the future, demographic information and history purchasing behavior interaction aspects may provide insights. Such characteristics may reveal patterns that individual features cannot. To evaluate non-linear correlations between characteristics and the target variable, polynomial features for continuous variables might be examined.  Segmenting clients by their marketing campaign interactions may reveal tendencies that affect repurchase probability.  By applying train\_test\_split from library sklearn, I have split the data into 3 different datasets:   * Training dataset: 91935 records (70%) * Validation dataset: 19701 records (15%) * Testing dataset: 19701 records (15%)   The scaling method selected for this project is the Standard Scaler, based on the results from the previous experiment.  The Standard SMOTE technique is implemented in our training dataset in order to enhance the complexity of the models by synthetically over-sampling minority data and then apply the trained model to the unseen validation and testing dataset with a highly imbalanced distribution. |
| **2.c. Modelling** | The Extreme Gradient Boosting (XGBoost) model was chosen for this experiment because of its strong performance in classification tasks, particularly in situations with complicated and unbalanced datasets. With a wide range of hyperparameters to adjust, XGBoost is popular for its efficacy and efficiency in handling various sorts of data and its capacity to control overfitting. To carefully adjust the model's performance, we used grid search to thoroughly optimize a number of important parameters:  The tuning process involved:  Step 1: Training with default settings to establish a baseline.  Step 2: Optimizing 'max\_depth' and 'min\_child\_weight' to control overfitting.  Step 3: Tuning 'gamma' to manage regularization.  Step 4: Adjusting 'subsample' and 'colsample\_bytree' for better model performance.  Step 5: Tuning the regularization parameter 'lambda' to add another level of complexity control.  Step 6: Final adjustments on 'n\_estimators' and reducing the 'learning\_rate' to converge to the optimal solution more effectively.  Wide-ranging hyperparameter were explored by the grid search, including:   * n\_estimators: Raised to 2000 at a greater computational cost in order to increase model adaptability. * max\_depth: Limited to 5 to avoid overfitting and balance model complexity. * min\_child\_weight: Set at 1 to regulate the model's sensitivity to changes in data. * gamma: Adjusted to 0.5 to manage leaf node creation, adding regularization. * subsample and colsample\_bytree: Both set at 0.9 to reduce overfitting by randomly sampling data points and features. * reg\_alpha: A small L1 regularization term of 0.001 was used for feature selection. * objective: Configured to "binary:logistic" for binary classification tasks. * scale\_pos\_weight: Kept at 1 to balance class weights. * seed: Fixed at 27 for reproducibility.   We selected the F3-score as the primary evaluation metric for this tuning process. The F3-score emphasizes recall significantly more than precision, which is critical in our context where identifying potential repeat customers is more valuable than avoiding false positives. This measure makes sure that our algorithm correctly identifies all potential positive instances (repurchases), which is in line with our business objective of maximizing the effectiveness and impact of marketing. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT RESULTS** | |
| Analyze 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** | Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes. |
| **3.b. Business Impact** | By providing more precise predictions by enhancing XGBoost model, customized client interaction methods are also supported in terms of customer retention. The total efficacy of marketing initiatives might be impacted by inaccurate model forecasts, which could result in the misallocation of resources or the passing up of chances. These insights support strategic decision-making by highlighting how crucial it is to match model performance to company goals in order to optimize effect. |
| **3.c. Encountered Issues** | Problems:   * Computational Restrictions: The first parameter grid was excessively large, resulting in lengthy training durations that were unworkable on local computers. * Complexity management: Larger parameter spaces and deeper trees sometimes caused overfitting. It was difficult to manage this complexity given the limitations of our processing power.   Solutions:   * I got around this by using Kaggle's GPU capabilities, which freed us from the time frame limits I had at home and enabled us to carry out more thorough grid searches. * Same as the Random Forest model, in order to manage these time limits, we had to change the way we approached grid searching. |

|  |  |
| --- | --- |
| 1. **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** | Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end. |
| **4.b. Suggestions / Recommendations** | * Optimization using GPU Acceleration: Since Kaggle's GPU works so well for training complicated models, more extensive hyperparameter grids will be investigated in further studies. It will be possible to go through more settings more quickly by using cloud resources. * Deployment Considerations: After the model achieves the required level of accuracy, the next stage will be to get it ready for use by making sure it works well with the operational infrastructure of the company. |