# **EXPERIMENT REPORT 4**

|  |  |
| --- | --- |
| **Student Name** | Ngoc Quang Vinh Pham |
| **Project Name** | Predicting Customer Car Repurchase Likelihood |
| **Date** | April 16, 2024 |
| **Deliverables** | 36106-AT2-25100660-experiment-4.ipynb  Random Forest 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** | The hypothesis suggests that optimizing the Random Forest hyperparameters would enhance the prediction accuracy of the model and, as a result, increase its application for marketing campaigns. This includes adjusting settings related to tree depth, leaf size, and the number of trees to optimize both the bias-variance tradeoff and computational efficiency. |
| **1.c. Experiment Objective** | A well-optimized Random Forest model that performs more accurately and efficiently than earlier iterations is the anticipated outcome. To determine the best possible configuration that provides the optimum trade-off between model complexity and performance, this entails experimenting with different hyperparameter choices. |

|  |  |
| --- | --- |
| 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 selection of Random Forest was motivated by its adaptability in handling many data types and its resilience to overfitting, particularly in the presence of imbalanced datasets such as our own. Wide-ranging hyperparameter were explored by the grid search, including:   * Number of Trees (n\_estimators): We tested values of 50, 100, 500, 1000, and 2000. Increasing the number of trees typically improves model accuracy but requires more computational resources. * Tree Depth (max\_depth): We explored depths of 10, 20, and 30 to determine the optimal level that captures enough data detail without falling into overfitting. * Minimum Samples per Split (min\_samples\_split) and Leaf (min\_samples\_leaf): We set values for splits at 2, 5, 10 and for leaf nodes at 1, 2, 4. These parameters help regulate the growth of trees, ensuring they are detailed enough to be effective but not so much that they model the noise in the data. * Bootstrap: We experimented with both using (True) and not using (False) bootstrap samples. This choice affects how samples are drawn for building trees, thus influencing the model's bias and variance.   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** | |
| 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 Random Forest tuning significantly enhanced the model's performance. Before tuning, the F3-score was 76.2, indicating an acceptable balance between precision and recall, but with potential room for improvement in recognizing minority classes. After tuning, the best configuration yielded an F3-score of 79.55. This improvement was achieved with the following hyperparameters: {'bootstrap': False, 'max\_depth': 30, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 500}. These settings allow the model to construct deeper trees without resampling, enabling a more detailed learning from the data without significant overfitting. |
| **3.b. Business Impact** | The business's objective of precisely identifying returning clients corresponds with the Random Forest model's improvement, which may greatly maximize marketing resources and boost campaign effectiveness. Risks, however, include the possibility that the model would perform poorly when exposed to new or unexpected data, which might result in lost opportunities or inefficient utilization of marketing resources. |
| **3.c. Encountered Issues** | Problems:   * The primary challenge was balancing model complexity with prediction accuracy and computational resources. Higher tree depths and lower node size thresholds sometimes led to overfitting, especially noticeable in smaller validation datasets. * Resource Allocation: The Random Forest's extensive parameter tuning required substantial computational resources, which exceeded the capabilities of my available hardware. * Time Limits: Training the model took longer and required more effort due to the wider parameter range.   Solutions:   * Utilizing Kaggle's GPUs helped reducing this issue. * 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** | The experiment demonstrated that precise tuning of Random Forest parameters can considerably improve the model's ability to predict customer repurchase likelihood, particularly by reducing errors in minority class predictions. The significant parameters in enhancing the model were max\_depth and bootstrap, where allowing deeper trees without bootstrap sampling proved most effective. It indicates that further research and development into this model would be beneficial in order to achieve even higher performance improvements. The continuous investigation of Random Forest model to improve prediction accuracy in the marketing strategies we employ is helped by these findings. |
| **4.b. Suggestions / Recommendations** | * Advanced Hyperparameter Tuning: To effectively explore the hyperparameter range, future research will build on the existing grid search methodology by using advanced approaches like randomized searches or genetic algorithms. * Model Simplification: Look for methods to use feature engineering or pruning approaches to make the model simpler so that training time may be reduced without significantly losing performance. * Scalability Testing: To make sure the improved Random Forest model maintains performance when implemented, scale it under various operating conditions. |