# **EXPERIMENT REPORT 2**

|  |  |
| --- | --- |
| **Student Name** | Ngoc Quang Vinh Pham |
| **Project Name** | Predicting Customer Car Repurchase Likelihood |
| **Date** | April 10, 2024 |
| **Deliverables** | 36106-AT2-25100660-experiment-2.ipynb  Model Comparison & Synthetic Over-sampling Application |

|  |  |
| --- | --- |
| 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 is that Synthetic Minority Over-sampling Techniques (SMOTE) can improve the prediction accuracy of customer repurchase likelihood by addressing the imbalance in the dataset. This could lead to better model performance, particularly in detecting the less frequent positive class (customers likely to repurchase). |
| **1.c. Experiment Objective** | The objective is to evaluate how various SMOTE techniques affect the performance of the top-performing models from the previous experiment. We aim to improve the models' ability to predict minority class instances and achieve better balance in class representation. The anticipated outcome includes improved F1-scores and ROC-AUC scores for models using SMOTE variants compared to the models without SMOTE. |

|  |  |
| --- | --- |
| 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. |
| **2.c. Modelling** | Applied SMOTE techniques to the three highest-performing models from the previous experiment: Random Forest, XGBoost, and Decision Tree. Various SMOTE techniques tested include:   * Standard SMOTE * Borderline SMOTE * SVM SMOTE * ADASYN * SMOTEENN * SMOTETomek   Each model was evaluated using the same default hyperparameters as in the previous experiment, focusing on tuning aspects directly affected by class distribution changes due to SMOTE. |

|  |  |
| --- | --- |
| 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** | The impact of each SMOTE technique on the model's minority class classification performance varied:   * The best F1-scores were obtained by Standard SMOTE and SMOTETomek, suggesting a superior trade-off between accuracy and recall. * Additionally, the F1-scores of ADASYN and SVM SMOTE increased, improving the sensitivity and specificity of the model. * Although they showed modest gains, borderline SMOTE and SMOTEENN outperformed the others by a little margin.   AUC-ROC values were close to 0.50, indicating that while the F1-score increased, the models' capacity to distinguish between classes at various thresholds did not show a statistically significant improvement. |
| **3.b. Business Impact** | Enhancing the F1-score indicates that the model may be better at identifying potential repurchasers, which corresponds directly to more efficient resource management in marketing campaigns and perhaps higher conversion rates. Nevertheless, the little fluctuation in AUC-ROC values suggests that model selection and training may still be enhanced. It is evident that there is potential for improvement since the assessment was conducted using the default machine learning model without any parameter tuning. |
| **3.c. Encountered Issues** | Some SMOTE techniques, particularly those involving complex algorithms like SVM SMOTE, required substantial computational resources.  There was a risk of model’s overfitting the minority class due to the increased presence of synthetic samples.  Solutions:  We are using synthetic sampling for only the training data, not the validation and testing set which can enhance the complexity and refinement of the machine learning models before meeting real data with high imbalance to see how the model perform. |

|  |  |
| --- | --- |
| 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 how crucial it is to use sophisticated resampling strategies to overcome class imbalance. By increasing the F1-score, SMOTE variants—particularly Standard SMOTE and SMOTETomek—were successful in increasing model performance. Standard SMOTE will thus be used in the parameter tuning process. Nevertheless, the evaluation metrics can still be modified for the benefit of the model’s performance and credibility, f-beta should be taken into consideration. |
| **4.b. Suggestions / Recommendations** | Better results could be obtained from further tests using various SMOTE approaches and hybrid models. Nonetheless, it works best when using the SMOTE technique that yielded the highest score (Standard SMOTE) and concentrating on fine-tuning each model's hyperparameters.  Create a plan for deploying the models that perform the best and keep an eye on their progress to make sure they adjust to new information.  Insights from this thorough analysis may direct future efforts to enhance predictive modeling in marketing analytics, especially in targeted tactics for acquiring and retaining customers. |