# **EXPERIMENT REPORT 1**

|  |  |
| --- | --- |
| **Student Name** | Ngoc Quang Vinh Pham |
| **Project Name** | Predicting Customer Car Repurchase Likelihood |
| **Date** | April 4, 2024 |
| **Deliverables** | 36106-AT2-25100660-experiment-1.ipynb  Model Comparison with Standardized Features |

|  |  |
| --- | --- |
| 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** | Our hypothesis is that machine learning algorithms can use past customer data to accurately forecast the chance of a repurchase. Our goal is to determine which algorithm best captures the patterns in our data that are associated with a customer's choice to repurchase by comparing several models. |
| **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 baseline 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%)   This experiment will determine the optimum feature scaling approach for the dataset. The techniques under review are min\_max\_scaler, max\_abs\_scaler, standard\_scaler, robust\_scaler. |
| **2.c. Modelling** | Given the imbalance, accuracy score is not a reliable metric. On the other hand, focusing on metrics that provide insights into the performance on the minority class will be more effective for the predictive models:   * F1 Score: The weighted average of Precision and Recall. This is useful when you seek a balance between Precision and Recall. * ROC-AUC: Area Under the Receiver Operating Characteristic Curve. A higher AUC indicates a better model performance.   Generate a baseline model performance based on the most frequency class in the target variable.  F1\_score of baseline model performance on training set: 0.0  F1\_score of baseline model performance on validation set: 0.0  F1\_score of baseline model performance on test set: 0.0  For the first experiment, we tend to evaluate 8 machine learning models using Kfold from sklearn.model\_selection library to compare their performance in predicting a binary outcome using F1 Score and ROC-AUC as the evaluation metrics. The machine learning models are: Logistic Regression, DecisionTree Classifier, RandomForest Classifier, XGBoost Classifier, KnearestNeighbour Classifier, Support Vector Machine. |

|  |  |
| --- | --- |
| 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 f1-score and standard deviation of the machine learning model above are shown in the metrics below.  LR\_model:0.3184040754354613(0.015763525652172008)  DT\_model:0.7769255304038888(0.015342083339219652)  RF\_model:0.8524664719593253(0.006907857712558807)  XGB\_model:0.8740457966092746(0.011865063531119515)  KNN\_model:0.6241727581242291(0.01290486624554512)  SVM\_model:0.6736390759206159(0.016083993838863386)   * High Performers: The data's intricate patterns could be effectively captured by XGBoost and Random Forest, as seen by their outstanding performance. * Moderate Performers: SVM and KNN showed mediocre performance, pointing to room for improvement in terms of feature scaling and parameter adjustment. * Poor Performer: Because it was unable to manage the non-linear connections in the severely unbalanced data, Logistic Regression performed poorly.   Key Issues:  Because of their simplicity compared to the complexity of the data, models such as Logistic Regression underperformed.  Performance of the model was greatly impacted by class imbalance, especially for simpler algorithms.  Here's a brief analysis and some considerations for choosing the best scaler after examining 4 tests with various feature scaling methods on your dataset with the Random Forest Classifier.  Scaling Results Analysis   * MinMax Scaler: F1\_score = 0.841 * MaxAbs Scaler: F1\_score = 0.845 * Standard Scaler: F1\_score = 0.845 * Robust Scaler: F1\_score = 0.844   Given the narrow range in performance across the scalers, there's only a marginal difference among them. However, the Standard Scaler, which achieved an F1-score of 0.845, is my preferred choice for this project. This decision is based on its ability to standardize features by removing the mean and scaling to unit variance, which generally ensures that the model's performance is not skewed by the presence of outliers. |
| **3.b. Business Impact** | Efficient Targeting: By precisely focusing on prospective repurchasers, high-performing models such as XGBoost have the ability to improve marketing campaigns and boost sales. Misallocation of Resources: Targeting using lower-performing models may result in inefficient use of marketing resources, as well as the omission of important client categories. |
| **3.c. Encountered Issues** | Class Imbalance: Considerable difficulties because of the target variable's skewed distribution, which had an impact on certain models' performance.  Needs for Feature Engineering: The predictive ability of simpler models could have been hampered by a lack of feature engineering.  The pace at which models are trained and optimized was affected by the large amount of computer resources needed for high-performance models such as XGBoost.  Solutions:  Resampling methods: In order to balance the dataset in subsequent tests, use methods like SMOTE. Advanced Feature Selection: To improve model performance, use more complex feature selection techniques. Efficiency Enhancement: Look at techniques like dimensionality reduction or easier ensemble approaches to lower computing burden without sacrificing model accuracy. |

|  |  |
| --- | --- |
| 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** | This experiment confirmed that the use of ensemble techniques is efficient when handling complicated datasets. Moreover, applying advanced feature engineering will offer a great chance to boost model performance. As the result has shown, the adaptability of models like XGBoost with imbalanced dataset reveal potential improvement of gradient boosting in predictive analytics. |
| **4.b. Suggestions / Recommendations** | Improve upon top-performing models (XGBoost and Random Forest in particular) by fine-tuning hyperparameters and doing cross-validation. However, less effective models should still be applied parameter-tuning process for comparison purposes.  For the following experiment, a frequent report of the retraining and updating of models in response to incoming data in order to ensure that they fulfill the business requirements.  Consider evaluating other algorithms, such as cost-sensitive learning models or various ensemble methods, that could be well suited for unbalanced datasets. |