# **EXPERIMENT REPORT**

| **Student Name** | Rusan Vaidya |
| --- | --- |
| **Project Name** | Performing Logistic Regression to predict if the customer purchases new car or not |
| **Date** | 27-04-2024 |
| **Deliverables** | 36106-AT2-24886400-experiement-1.ipynb  LogisticRegression  Precision, Recall, Imbalance Data |

| 1. **EXPERIMENT BACKGROUND** | | |
| --- | --- | --- |
| This project aims to predict the repurchases of cars by the customer using logistic regression, comparing accuracy to the dumb model, which always predicts the mode of the target variable. This experiment aims to show that LogisticRegression significantly outperforms the simplest model. | | |
| **1.a. Business Objective** | The goal of the project is to refine the car business and identify maximum customer repurchases with minimum waste of resources to target the customer to make marketing efforts and make proper decision-making. Accurate predictions will lead to identifying potential buyers and profitability while inaccuracy of results is a waste of marketing and production of cars. | |
| **1.b. Hypothesis** | **Hypothesis:** A LogisticRegression model outperforms the simplest baseline model in customer behavior in car repurchase.  **Explanation:** A dumb model or the baseline model always predicts the same outcome no matter the actual prediction. The logistic regression is expected to outperform the simplest model and make more accurate predictions which would help in better decision making. | |
| **1.c. Experiment Objective** | This experiment expects to predict accurate results and perform better than the baseline model. If the result from this experiment of the LogisticRegression outperforms the baseline model, a more complex algorithm or tuning of the hyperparameter may perform better in predicting results. If the result is contrary, it doesn’t require any further complex algorithm as the baseline model is the best fit. | |

| 1. **EXPERIMENT DETAILS** | | |
| --- | --- | --- |
| The experiment uses different data preprocessing tools to drop the columns and fill values to deal with the missing data from the dataset and also uses tools such as OneHotEncoding to encode the categorical data used to predict the customer repurchasing behavior. | | |
| **2.a. Data Preparation** | 1. In this experiment, the `***age\_band***`and `***gender***` column from the dataset was dropped from features to train the model. 2. Dropped the identifier ***`ID`*** from the dataset. 3. Dropped all the duplicate values. 4. The column `***car\_model***`, it has four different categories: “LCV”, “Small/Medium”, “Large/SUV”, and “Others” and was encoded. 5. The column ***`car\_segment`***, consists of 19 different models removing the prefix “model\_” and converting it into integer value. | |
| **2.b. Feature Engineering** | 1. More than 85% of the data in column *`****age\_band****`* has and more than half of the data in the *`****gender****`* column are Null from the dataset. Hence the columns were dropped from the dataset to train the model. 2. Encoding numeric value for different values of ***`car\_model`*** and ***`car\_segment`*** as it requires numeric value to train the model. 3. **The dataset used for this experiment is imbalanced,** hence requiring preprocessing where RandomUnderSampler is used for training the model. | |
| **2.c. Modelling** | Logistic regression is considered one of the simpler models among the complex ones and outperforms the dumb model, making a perfect algorithm for this experiment.  The hyperparameters used for this experiment were:   1. **random\_state**: The random\_state was set to 42 to ensure the results remain the same every time the model training is executed. 2. **class\_weight**: The class\_weight is set to ‘balanced’ as it is an imbalanced dataset. 3. **C**: After conducting a series of tuning and analyzing through the Precision-Recall curve and ROC curve, the value of C was set to 1. | |

| 1. **EXPERIMENT RESULTS** | | |
| --- | --- | --- |
| From a technical point of view we evaluate the performance of the LogisticRegression model from accuracy, precision, recall and ROC AUC. While from a business point of view, analyze features and their impact on repurchases of cars. | | |
| **3.a. Technical Performance** | The analysis is a comparison of baseline models that gives a biased prediction towards the majority class of car class 0 i.e. “*Not Re-purchased*”, every time and achieves accuracy of 97%. The precision and recall of class 0 is very high while class 1 is 0 for both precision and recall. While it shows high accuracy it is completely useless for predicting customers who would repurchase which is the main motive of this experiment.  The Logistic Regression model gives a better prediction with lower recall for class 0 (“*Not Re-purchase*”) suggesting it might miss some cases for cars not being re-purchased. For class 1 (“*Re-purchase*”) cases, the precision is 11% on training which is not great as many of the results might be falsely predicted although recall is high at 79% as it accurately predicts most of the actual re-purchases and performing 78.61% on validation and 78.95% on testing. | |
| **3.b. Business Impact** | The experiment intends to accurately identify customers who are likely to “*Re-purchase”* a car. The dumb model predicts the biased result which is the majority of the prediction fails to identify the potential customers who would repurchase cars. Logistic Regression on the other hand though less accurately predicts customers excels in identifying potential customers with high recall score. However, it also has low precision which suggests many customers incorrectly which could lead to waste of resources. | |
| **3.c. Encountered Issues** | **Issues:**   1. Numeric mapping was performed which might not be an appropriate way to encode the categorical values. 2. Two of the features which might be critical for better prediction were dropped due to the number of NaN values present in the data which might have affected the performance. 3. The model performs with a low precision for class 1 (“*Repurchase*”) indicating high number of false positive and very low f1\_score   **Possible Solution:**   1. Performing OneHotEnoder to encode the categorical values. 2. Try to use fillna to handle features with NaN values. 3. Trying more advanced and complex algorithms to improve precision for ‘Repurchase’ class in future experiments. | |

| 1. **FUTURE EXPERIMENT** | | |
| --- | --- | --- |
| This experiment handled the imbalance dataset and predicting customer repurchasing behavior and emphasizing evaluation and business applicability for improved decision making. | | |
| **4.a. Key Learning** | From this experiment, the LogisticRegression performs better in predicting car repurchasing behavior even though it has lower precision for class 1 “*Re-purchasing*” customer than the dumb model which performs really badly for potential customers as it predicts only “*Not Re-purchasing*” customer class 0 every time making it unsuitable for practical applications aimed at customer retention. | |
| **4.b. Suggestions / Recommendations** | For the future experiments,   1. Using functions such as fillna() to overcome losing of data 2. More data refining can be done such as data encoding or handling missing values. 3. Exploring hyperparameters that can make it perform better. 4. Exploring other advanced and complex models which might perform in better precision. | |