# **EXPERIMENT REPORT**

| **Student Name** | Rusan Vaidya |
| --- | --- |
| **Project Name** | Performing Support Vector Machine to predict if the customer purchases new car or not |
| **Date** | 27-04-2024 |
| **Deliverables** | 36106-AT2-24886400-experiment-2.ipynb  LogisticRegression, SVC  Precision, Recall, Imbalance Data |

| 1. **EXPERIMENT BACKGROUND** | | |
| --- | --- | --- |
| This project aims to predict the repurchases of cars by the customer using SVC comparing the accuracy to the Logistic Regression model target variable with 79% accuracy. The goal of this experiment is to check if the SVC model could outperform the LogisticRegression. | | |
| **1.a. Business Objective** | This experiment aims to enhance the accuracy and address challenges encountered in Experiment 1, where the Logistic Regression model exhibited some of the potential customers, but resulting in a high rate of errors in identifying customers leading to huge waste of resources.  Additionally, this study will evaluate whether the Support Vector Classifier (SVC) can outperform the existing model. The objective is to maximize customer targeting efficiency with minimal resource wastage, thereby optimizing marketing efforts and improving decision-making processes. | |
| **1.b. Hypothesis** | **Hypothesis:** A Support Vector Machine model makes more accurate car repurchase prediction in comparison to LogisticRegression.  **Explanation:** A LogisticRegression prediction is 79% accurate but the precision is 0.11 making many false positive predictions. The SVC model is a more advanced model and is expected to perform better and give a more accurate outcome which would help in better decision making. | |
| **1.c. Experiment Objective** | This experiment expects to predict accurate results and perform better than the LogisticRegression and overcome the issue in experiment 1. The SVC model, is a more complex algorithm or tuning of the hyperparameter may perform better in predicting results also using correct Sampling as the dataset is imbalanced and if the outcome is not better than LogisticRegression then there is no requirement for further tuning. | |

| 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 use 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 had many missing values which were filled by “Not Specified”. This label distinguishes individuals who chose not to disclose their identity or demographic information, potentially indicating shared characteristics among this group. 2. Dropped the identifier ***`ID`*** from the dataset. 3. Dropped all the duplicate values. 4. For the column `***car\_model***`, and ***`car\_segment`***, it is encoded using OneHotEncoding. | |
| **2.b. Feature Engineering** | 1. More than 85% of the data in column *`****age\_band****`* has and more than half of data in *`****gender****`* column are Null from the dataset. Hence the columns were dropped from the dataset to train the model. 2. OneHotEncoding of ***`car\_model`*** and ***`car\_segment`*** as its essential step as it is not an ordinal category. 3. **The dataset used for this experiment is imbalanced,** hence requires preprocessing where SMOTE is used for training the model. | |
| **2.c. Modelling** | SVC is among the complex ones and it is also suitable for non linear relation between features and labels.  The hyperparameter 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. **C :** After conducting a couple of tuning C was set to 1 as it perform better than 0.1 3. **max\_iter** : The max\_iter was experimented to 100 but it performed better in default setting i.e -1 which is no limit. | |

| 1. **EXPERIMENT RESULTS** | | |
| --- | --- | --- |
| From a technical point of view we evaluate the performance of the SVC model from accuracy, precision, recall and ROC AUC. While, from a business point of view, analyze features and its impact on repurchases of cars. | | |
| **3.a. Technical Performance** | The analysis is a comparison of LogisticRegression models taken as a baseline model for this experiment gives a prediction of 79% accuracy with low precision score of 0.11 for for class 1(“Re-purchase”) using RandomUnderSampler in experiment 1 and For experiment 2, using SMOTE technique there was increase in accuracy of 81% and 80.80% and 81.01%.  The SVC model was used and gave a better prediction after tuning with a better precision of 0.39 for class 1 (“*Re-purchase*”) suggesting the model being capable of making more accurate predictions and not mislabel 0 as 1 with a score of 97%. | |
| **3.b. Business Impact** | The experiment intends to identify customers who are likely to “*Re-purchase”* a car. The Logistic Regression accurately predicts who won’t be purchasing the second car and identifies potential customers and also labels them as customers who probably won’t buy the second car.  The SVC model performs higher precision compared to the Logistic Regression model, as it captures individuals who are unlikely to purchase a second car. This enhanced precision helps in marketing efforts by minimizing resource wastage, making the SVC model more effective than Logistic Regression in targeting potential customers. | |
| **3.c. Encountered Issues** | **Issues:**   1. The dataset is imbalanced. 2. The model training consumed more time than Logistic Regression.   **Solution:**   1. SMOTE sampling technique was used for training the data. 2. Try different models or different settings. | |

| 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** | This experiment indicates that the SVC performs better in accurately predicting car repurchasing behavior and is suitable for practical applications aimed at potential and customer retention, thereby optimizing resource allocation in marketing efforts. This insight warrants further investigation into advanced modeling techniques. | |
| **4.b. Suggestions / Recommendations** | For the future experiments,   1. Exploring hyperparameters that can make it perform better and does not overfit. 2. Exploring advanced models that can give better outcomes than SVC as it still makes errors on accurately predicting results. | |