# **EXPERIMENT REPORT**

| **Student Name** | Rusan Vaidya |
| --- | --- |
| **Project Name** | Performing DecisionTreeClassifier to predict if the customer purchases new car or not |
| **Date** | 27-04-2024 |
| **Deliverables** | 36106-AT2-24886400-experiment-4.ipynb  KNeighborClassifier, SVC, DecisionTreeClassifier  Precision, Recall, Imbalance Data, n\_neighbor, max\_depth, pipeline, ColumnTransformer |

| 1. **EXPERIMENT BACKGROUND** | | |
| --- | --- | --- |
| This project aims to forecast car repurchases by customers using DecisionTreeClassifier, contrasting its accuracy with the KNeighborClassifier model, which achieves 98% accuracy and a precision of 0.55. The objective is to check if the DecisionTreeClassifier can surpass the SVC performance. | | |
| **1.a. Business Objective** | The aim of this experiment is to address the challenges identified in experiment 3 by evaluating whether the DecisionTreeClassifier can achieve a higher number of potential customers with minimum wastage of resources compared to the KNeighbor model and SVC. Additionally, it seeks to determine the potential impact of greater wastage due to errors, considering their implications for business outcomes. | |
| **1.b. Hypothesis** | **Hypothesis:** The DecisionTreeClassifier model surpasses KNeighbor in car repurchase predictions.  **Explanation:** Despite KNeighbor's high 98% accuracy, its precision is superior to previous models', but the decrease in recall leads to numerous true negative predictions. DecisionTree model being tree based classifier may have better accuracy and balance in recall and precision. | |
| **1.c. Experiment Objective** | This experiment aims to achieve accurate predictions and assess whether the DecisionTreeClassifier outperforms SVC, addressing the shortcomings identified in experiment 3. Despite SVC's strong accuracy, its lower precision led to errors, examining whether the DecisionTreeClassifier can improve predictions. | |

| 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 was filled with “Not Specified”. 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. 5. The features are scaled using the StandardScaler | |
| **2.b. Feature Engineering** | 1. Over 85% of the entries in the ***`age\_band`*** column and over 50% of the entries in the ***`gender`*** column are missing from the dataset. These missing values have been filled with the term "Not Specified" to analyze whether this particular categorization leads to improved performance in the model. 2. OneHotEncoding of ***`car\_model`*** and ***`car\_segment`*** as its essential step as it is not an ordinal category. 3. Scaling is performed in order to maintain scaling across features. 4. Standardization and Encoding was performed using sklearn instead of an imblearn tool named Pipeline. 5. **The dataset used for this experiment is imbalanced.** However, for the DecisionTreeClassifier model, this doesn't matter as much because it looks at trees to make predictions, so the imbalance doesn't affect it much. | |
| **2.c. Modelling** | The DecisionTreeClassifier utilizes a hierarchical structure to model relationships between features and labels, making it effective for capturing non-linear patterns in data.  The hyperparameter used for this experiment were:   1. **max\_depth** : The max\_depth was set to 10, preventing overfitting and enhancing generalization performance, resulting in improved predictions. | |

| 1. **EXPERIMENT RESULTS** | | |
| --- | --- | --- |
| In terms of technical evaluation, the DecisionTreeClassifier model's effectiveness is evaluated through metrics like accuracy, precision, recall, and ROC AUC. Meanwhile, from a business standpoint, it is checked how different features impact car repurchase decisions. | | |
| **3.a. Technical Performance** | For SVC with C=1 and max\_depth=-1, precision and recall were 0.55 and 0.94 for class 1, respectively, showing higher number of correct predictions of individuals who are potential buyers and making errors on predicting individuals as buyers who won’t buy leading to wastage of resources.  While tuning, max\_depth=10 has a precision and recall of 0.95 and 0.81 respectively, indicating balanced performance with high accuracy. For max\_depth=20, precision and recall reached 100%, suggesting potential overfitting as the model may struggle with unseen data despite high accuracy. | |
| **3.b. Business Impact** | For DecisionTreeClassifier, it gives a more balanced prediction as it makes minimum errors with a higher number of actual potential buyers which helps in maintaining resources as well as serving potential customers. This balanced accuracy suggests reliable predictions for business decisions, making it a preferable choice over the SVC. | |
| **3.c. Encountered Issues** | **Issues:**   1. Using appropriate scaling for the data. 2. Balancing recall and precision.   **Solution:**   1. StandardScaler was used for scaling the data. 2. Analyzing through AveragePrecision and ROC curves and visualization. | |

| 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** | The experiment displayed DecisionTreeClassifier model's potential in handling imbalanced data, offering a promising outcome for improving prediction accuracy in car repurchase scenarios. Further exploration and fine-tuning of the model's parameters could unlock even better performance and decision-making support. | |
| **4.b. Suggestions / Recommendations** | For the future experiments,   1. Optimize Decision Tree parameters for better performance. 2. Use ensembles for better prediction. | |