# **EXPERIMENT REPORT**

| **Student Name** | Rusan Vaidya |
| --- | --- |
| **Project Name** | Performing KNeighborClassifier to predict if the customer purchases new car or not |
| **Date** | 27-04-2024 |
| **Deliverables** | 36106-AT2-24886400-experiment-3.ipynb  KNeighborClassifier, SVC  Precision, Recall, Imbalance Data, n\_neighbor |

| 1. **EXPERIMENT BACKGROUND** | | |
| --- | --- | --- |
| This project aims to predict the repurchases of cars by the customer using KNeighborClassifier comparing the accuracy to the SVC model target variable with 97% accuracy. The goal of this experiment is to check if the KNeighborClassifier model could outperform the SVC. | | |
| **1.a. Business Objective** | The goal of this experiment is to overcome the issue from experiment 2 and assess whether the K-Neighbors model can surpass the Support Vector Classifier (SVC) in terms of precision. The objective is to refine the classification of potential customers, achieving higher accuracy while minimizing resource wastage. The business goal is to enhance the effectiveness of customer targeting strategies, thus improving resource allocation and decision-making processes. | |
| **1.b. Hypothesis** | **Hypothesis**: A KNeighbor model outperforms SVC in car repurchase predictions.  **Explanation**: Although SVC achieves 96% accuracy, its precision is only 0.39, resulting in numerous false positive predictions. Conversely, the KNeighbor model utilizes nearby data for predictions, potentially offering higher accuracy and aiding better business decision-making. | |
| **1.c. Experiment Objective** | This experiment expects to predict accurate results and check if the KNeighbor model performs better than the SVC and overcome the issue in experiment 2. The KNeighbor model may perform better in predicting results also using correct hyperparameters. | |

| 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. For the column `***car\_model***`, and ***`car\_segment`***, it is encoded using OneHotEncoding. 3. Dropped the identifier ***`ID`*** from the dataset. 4. Dropped all the duplicate values. 5. The features are scaled using the StandardScaler and Encoded using the ColumnTransformer library. | |
| **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 missing from the dataset. These missing values have been filled with 'Not Specified' to denote unspecified information, suggesting a common characteristic among these individuals. 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 ColumnTransformer and pipeline was used to create a model. 5. **The dataset used for this experiment is imbalanced.** However, for the KNeighbor model, this doesn't matter as much because it looks at nearby data points to make predictions, so the imbalance doesn't affect it much. | |
| **2.c. Modelling** | KNeighborClassifier is a complex model suitable for capturing non-linear relationships between features and labels.  The hyperparameter used for this experiment were:   1. **n\_neighbors** : After tuning, n\_neighbors was set to 5 to 50. Higher values increase precision but decrease recall, risking missed positive cases, crucial in decision-making where recall is vital. | |

| 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** | After applying SVMSMOTE and StandardScaler, precision for class 1 improved from 0.39 to 0.55, enhancing overall accuracy to 97.55% in validation and 97.52% in testing.  KNeighborClassifier with n\_neighbors=5 achieved balanced precision and recall (0.95 and 0.59 respectively), but with higher neighbor values, precision remained high while recall decreased significantly, potentially due to broader decision boundaries.  Though KNeighbors accuracy increased, the average precision score and ROC AUC score of SVC model remain higher. Also, recall for class 1 “Re-purchased” of SVC is greater than the KNeighbors. | |
| **3.b. Business Impact** | The KNeighbors experiments found that as the number of neighbors increased, precision improved but recall dropped. Higher n\_neighbors values result in higher precision but lower recall,potentially leading to missed positive cases and impacting decision-making accuracy, particularly in scenarios where recall is critical.  Since the SVC model predicts much more accurate potential customer than the KNeighbors, KNeighbors is not considered a better model for this experiment. | |
| **3.c. Encountered Issues** | **Issues:**   1. High number of null values in the data 2. Using appropriate scaling for the data. 3. Balancing recall and precision.   **Solution:**   1. Used fillna to overcome the null values. 2. StandardScaler was used for scaling the data. 3. 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** | 1. Effectively enhance model precision, validating further use of data augmentation techniques. 2. Confirmed the strength of SVC in handling complex patterns, suggesting deeper hyperparameter tuning and kernel experimentation. 3. Increasement in neighbors improves precision but reduces recall | |
| **4.b. Suggestions / Recommendations** | For the future experiments,   1. Experiment with different hyperparameters for the SVC model to further improve precision and recall. Expected uplift: Enhanced model accuracy and potentially higher ROC AUC score. 2. Use decision tree classifiers for better prediction. | |