EXPERIMENT REPORT

Student Name	Shaqran Bin Saleh
Project Name	Classification Models
Date	26/4/2024
Deliverables	Notebook Name : 36106_AT2_25010238_experiment_2.ipynb model name : Support Vector Classifier

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	The goal of this project is to build a predictive model with accurate likeliness of repurchasing a car for customers. The model's predictions can be used in a myriad of ways. The model can be used to segment customers by repurchase likelihood, focusing marketing efforts on high-potential features of the dataset for efficient resource use. Insights from the data and trained model can also enhance customer retention strategies and optimize inventory management. Additionally, the results would provide feedback to refine sales tactics and customer interactions, continually improving approaches to meet customer needs effectively. Accurate predictions would mean enhance efficiency in targeting the right customers in reducing marketing efforts and higher sales. Through proper data exploration and trained model insights proper understanding of customer needs and help improve customer satisfaction. Inaccurate repurchase likelihood prediction by the model might result in diminished business opportunities, reduced ROI, wasted marketing budget, and unhappy clients. Customers may become disconnected from a brand due to improper targeting, which could be perceived as spam and harm the company's reputation.
1.b. Hypothesis	I wish to test the hypothesis that several characteristics: Gender, The model of vehicle, The type of vehicle, Age of their last vehicle, Number of scheduled services used under warranty, Number of non-scheduled services used under warranty, Amount paid for scheduled services, Amount paid for nonscheduled services, Amount paid in total for services, Total number of services, The number of months since the last service, Annualized vehicle mileage, Number of different dealers visited for servicing, Number of services had at the same dealer where the vehicle was purchased have significant effects on the customers' likeliness of repurchasing a vehicle. I like to get the answers of few questions like "What are the primary repurchasing variables for customers and how do they interact to influence their repurchasing likelihood?" By testing this hypothesis and answering the accompanying question, we can acquire a deeper understanding of the factors that influence customers' likelihood of repurchasing.

1.c. Experiment Objective	From this experiment we will be deploying a predictive model using SVC (Support Vector Classifier) model that accurately handles the data and make accurate predictions. The goal is to identifying key variable effecting the customers' likelihood of repurchasing, developing an accurate model, apply it on the dataset and find the accuracy of the model basing upon the different hyperparameter tuning.

2. 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

For proper data preparation, the following steps were taken:

1. The assigned .csv file was loaded into the notebook as a dataframe for the experiment.
2. The dimensions (rows, columns) of the data loaded from the .csv file was found. This step was performed so that we can properly understand the amount of data we are working with. It also gives us information about the number of features present in the data.
3. We have 17 features among which 1 is our target variable which is “Target”. We explored all the features of the dataset.
4. We check all the features if there are any null values and handle the situation.
5. We check the statistics (count, mean, standard deviation, minimum values, maximum value) of the features.
6. We drop the rows containing missing values from the dataset.
7. We drop the rows containing duplicate data.
8. Since we have 4 features containing categorical data. We introduced OneHotEncoder through which we transform the data and concatenate the new columns into our dataset. And finally remove the previous columns on which the transformation was done.
9. The feature “Target” is our target variable so we separate it as our ‘y’ variable and the rest were are considering our independent variable which
10. Then we split the data to three categories. Which are training data, testing data and validation data.
11. We distribute the data firstly into two segments Train + Validation and Test sets. The training + validation set would hold 80% of the data and rest 20% is testing data.
12. Then we again split the Train + Validation to separate Training and Validation sets. The validation would hold 30% of the Train + Validation data, meaning it hold 24% of the original data. The rest of the 56% of the original data is reserved for the training dataset.
13. And finally we performed feature scaling

All the above steps were vital in this experiment is also vital for the next experiments.

2.b. Feature Engineering

While performing data preparation, we found out that the features : 'age_band', 'gender', 'car_model', 'car_segment' are categorical and are all important in the influence of the target variable. So, we had to encode these features. We achieved this encoding using the OneHotEncoder()

As we concatenated the new columns, we made sure to drop the columns on which the transformation was performed.

2.c. Modelling

For the experiment to predict customer repurchase likelihood, the Support Vector Classifier (SVC) was chosen.

SVC is straightforward to implement and understand, SVC works effectively in high-dimensional spaces.

- One of SVC's most powerful features is the kernel trick. It allows the model to handle non-linear decision boundaries by transforming data into higher dimensions where it is possible to find a linear separator. Through the choice of a kernel function (linear, polynomial, radial basis function (RBF), sigmoid), SVC can be adapted to different data structures and complexities.
- The regularization parameter (C) in SVC provides a way to control the trade-off between achieving a low error on the training data and minimizing the model complexity for better generalization. By adjusting C, you can reduce the risk of overfitting, especially when the dataset is noisy or when there is not a clear margin of separation between classes. making it an accessible choice for initial modeling efforts. It is well-known for its effectiveness in classification scenarios, especially when the relationship between feature variables and the class is not linear.

The decision to use SVC should be aligned with the specific characteristics of the data and the business or research objectives. Its powerful capabilities in handling complex, high-dimensional, and imbalanced data make it a robust choice for many classification challenges.

While performing the experiment hyperparameters tuning was done quite some time. The primary hyperparameter which is the regularization parameter C was tuned. And the following values were tested: 2, 0.4, 0.05.

The regularization parameter C significantly impacts by controlling the trade-off between achieving a low error on the training data and minimizing the model complexity for better generalization to new data.

For future we plan on exploring Decision Tree, Random Forest and also Extra Trees.

14. EXPERIMENT RESULTS

Analyse 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 performance metrics used in this experiment are :
1. Accuracy Score
2. Precision Score
2. F1 Score
3. Confusion Matrix

We used the SVC model and trained it multiple times as we performed multiple Hyperparameter tuning.

The results of the metrics according to different hyperparameters are as follow:

1. For default hyperparameters:
   - Accuracy Scores:
     a. On Training set: 0.9980487804878049
     b. On Validation set: 0.9952207555757852
     c. On Testing set: 0.9967231021299836
   - F1 Scores:
     a. On Training set: 0.8979591836734694
     b. On Validation set: 0.7407407407407407
     c. On Testing set: 0.8378378378378378
   - Precision Score:
     a. On Training set: 0.967032967032967
     b. On Validation set: 0.8571428571428571
     c. On Testing set: 1.0
   - Confusion Matrix:

```
[[3619    0]
 [  12   31]]
```

2. For Hyperparameter C=2:
   - Accuracy Scores:
     a. On Training set: 0.9988292682926829
     b. On Validation set: 0.9959035047792444
     c. On Testing set: 0.9975423265974878
   - Precision Score:
     a. On Training set: 0.979381443298969
     b. On Validation set: 0.8333333333333334
     c. On Testing set: 0.9722222222222222
   - F1 Scores:
     a. On Training set: 0.9405940594059405
     b. On Validation set: 0.7954545454545455
     c. On Testing set: 0.8860759493670887
   - Confusion Matrix:

```
[[3618    1]
 [  8   35]]
```

3. For Hyperparameters C=0.4:

- Accuracy Scores:
  a. On Training set: 0.9940487804878049
  b. On Validation set: 0.9915794264906691
  c. On Testing set: 0.9926269797924632
- Precision Score:
  a. On Training set: 1.0
  b. On Validation set: 0.8461538461538461
  c. On Testing set: 1.0
- F1 Scores:
  a. On Training set: 0.5906040268456376
  b. On Validation set: 0.3728813559322034
  c. On Testing set: 0.5423728813559322
- Confusion Matrix:

```
[[3619    0]
 [  27   16]]
```

4. For Hyperparameters C=0.05:

- Accuracy Scores:
  a. On Training set: 0.9940487804878049
  b. On Validation set: 0.9915794264906691
  c. On Testing set: 0.9926269797924632
- Precision Score:
  a. On Training set: 1.0
  b. On Validation set: 0.8461538461538461
  c. On Testing set: 1.0
- F1 Scores:
  a. On Training set: 0.5906040268456376
  b. On Validation set: 0.3728813559322034
  c. On Testing set: 0.5423728813559322
- Confusion Matrix:

```
[[3619    0]
 [  27   16]]
```

Main Issues and Causes:

- The F1 scores vary significantly between different configurations, indicating a fluctuating balance between precision and recall. This fluctuation is especially notable when adjusting the regularization parameter. This can happen due to class imbalance.
- For lower C values, the model exhibits a higher number of false negatives, which significantly impacts the model's ability to detect positive cases.
- While all configurations maintain relatively high accuracy, there's a noticeable variability in F1 scores and precision-recall balance across training, validation, and testing sets.
- With higher values of hyperparameter C (like 2), there is less regularization, allowing the model to fit more closely to the training data, potentially capturing more complexities of the minority class. Conversely, lower values of C (like 0.4 and 0.05) increase regularization, which may make the model too simple to capture the nuances necessary to identify the minority class effectively, thus reducing recall.

3.b. Business Impact

According to our objective which we set earlier which was to deploy a predictive SVC model that performs well. We can come the following conclusions about the results of this experiment:

- All the tested configurations show high accuracy on training, validation, and testing sets. This suggests that the model is generally reliable for making predictions across various data samples. High accuracy is desirable.

	• Higher precision with C=2 and lower precision with C=0.4 and C=0.05 on testing sets indicates fluctuating reliability in predicting actual customers. A precision of 1.0 at C=2 on the testing set means no false positives—every predicted customer actually repurchased, which is ideal for targeted interventions. • The F1 scores, particularly in lower C values, show some variability and lower performance on the validation and testing sets. This suggests difficulties in balancing precision and recall, which are crucial for targeting the right customers for retention efforts. • The configurations with higher false negatives (especially at lower C values) suggest missed opportunities to identify potential repurchasing customers. And the impacts the business might face due to imperfect results would be: • Missed Revenue Opportunities due to inability to identify customers (for low recall for the minority class) likely to repurchase may lead to missed chances for effective targeted marketing and, reducing potential revenue gains. • Inefficient Resource Allocation as model misclassifications can cause misallocation of marketing resources, wasting efforts on less likely customers rather than those predisposed to repurchase. • Customer Relationship Management as inaccurate predictions, which may happen if Bias towards the majority class happens as it fails to capture the diverse behaviors and needs of all customers, can degrade customer experience by misaligning marketing efforts with actual customer behavior and preferences.
3.c. Encountered Issues	As we performed this experiment, we were encountered with the following issues: • Class Imbalance was found significantly and it can lead to models that are biased towards the majority class, as indicated by high accuracy but low recall and F1 scores. • Hyperparameter Optimization was difficult as it was difficult in choosing the optimal C value, which affects model performance. • Feature Engineering was important as ineffective feature engineering can limit the model's ability to learn complex patterns.

---

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	Reflecting on the outcomes and performance of the SVC model experiments for predicting customer repurchase likelihood, several key insights emerge. • The choice of hyperparameters, specifically the regularization parameter C in an SVC model, significantly affects the model's performance, particularly in balancing precision and recall. This highlights the necessity of careful hyperparameter tuning to optimize model performance for specific business needs. • The F1 score serves as a critical metric when dealing with imbalanced classes, as it balances precision and recall in a single measure, helping assess the model's overall effectiveness in identifying minority class (likely to repurchase) instances. • The setting of C=2 provides the best balance between accuracy and F1 score, indicating that moderate regularization might be optimal for this scenario. This setting minimized false negatives without introducing many false positives • Class imbalance challenge. The class imbalance significantly affects the model's ability to predict less frequent outcomes accurately. Techniques to handle class imbalance need

to be integral to the modeling process to improve minority class predictions.

- The business impact of false negatives (missing out on identifying likely repurchasers) can be more severe than false positives (incorrectly identifying non-churners as churners). This prioritization affects how a model should be tuned, emphasizing the need to minimize false negatives to prevent loss of revenue.

Given the insights and the issues encountered, here are reasons to continue experimenting with the SVC approach:

- The SVC model demonstrated high precision in several configurations, particularly with higher C values. This indicates its effectiveness in correctly identifying the positive class (likely repurchasing customers) in imbalanced datasets.
- SVC's inherent regularization feature is beneficial for preventing the model from overfitting, especially important when dealing with high-dimensional data. Experimenting with different regularization strengths could further enhance the model’s generalizability.
- Further development and optimization of features could improve the model's predictive power, especially with domain-specific insights.

Depending on continued lack of improvement in handling the minority class, exploring other models like decision trees, ensemble methods, or even neural networks might be warranted. These models might offer better feature handling and intrinsic class balance mechanisms.

4.b. Suggestions / Recommendations

The experimentation with the SVC for predicting customer repurchase likelihood identified several pathways for further improvement. These refinements can be categorized into two main phases: model optimization and deployment preparation.

1 Model Optimization:

- Consider implementing resampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or under-sampling the majority class to balance the dataset. This can help improve the model's ability to generalize and perform better in predicting the minority class.

- We will investigate more methods such as randomized search, grid search, or Bayesian optimization to fine-tune the SVC model's hyperparameters.

- If the SVC model's performance reaches a plateau, we will consider employing alternative algorithms like Random Forest, Extra Trees. These models may be better suited to capture non-linear relationships and feature interactions within the data, potentially leading to enhanced prediction accuracy.

2 Deployment Preparation:

Following the optimization phase, a series of steps will ensure a successful deployment of the chosen model:

1. Once all testing and validation efforts are complete, the model configuration demonstrating the best performance will be selected for deployment.

2. A comprehensive plan outlining the integration of the chosen model into the existing IT infrastructure should be formulated.

3. A monitoring system should be established to track the model's performance in a production environment using real-world data. This will be accompanied by a maintenance schedule for periodically updating the model to maintain its effectiveness.

	4. Training materials and comprehensive documentation should be developed to educate end-users on the model's functionality, its intended use within business processes, and its limitations. 5. The deployment process should adhere to all relevant compliance regulations and data privacy rules, ensuring the integrity and security of customer data. By prioritizing model optimization followed by a well-defined deployment plan, we aim to maximize the predictive power of the customer repurchase prediction model while guaranteeing a smooth and effective transition to real-world applications.