Big Data Course

Capstone Project   
Final Report

For students (instructor review required)

ⓒ2023 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

|  |
| --- |
| Building and Optimizing Machine Learning Models |

25/08/2024

Team name: N0. 1

Lê Thị Ngọc Ánh

Phạm Hùng Cường

Nguyễn Phúc Cường

Hà Văn Hậu

Trịnh Ngọc Khiêm

Văn Thanh Lân

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. Simulated Scenario Description

2.2. Datasets Selection and Description

2.3. Models Selection

2.4. Build and train models

2.5. Evaluate and Compare

3. Results

3.1. The f1-score results from models

3.2. Percentage growth after optimizing

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

1.1. Background Information

Building and optimizing machine learning models involves selecting appropriate algorithms, preprocessing data, and fine-tuning model parameters to achieve high accuracy and efficiency. This process is crucial for deriving meaningful insights and predictions from data, enabling better decision-making across various industries. Effective optimization ensures models are not only accurate but also computationally efficient, leading to improved performance in real-world applications.

1.2. Motivation and Objective

- Motivation: Because of the aspiration of making machine learning model, our team endeavour research as much as possible on documents from instructor to create most wonderful outputs to reach goals

- Objective: Successfully build three models to evaluate customer shopping habits and Optimize models to obtain the highest f1-score

1.3. Members and Role Assignments

- Hậu, Khiêm: Build and optimize Random Forest Classifier model.

- Ánh, Hùng Cường: Build and optimize Support Vector Machine model.

- Lân, Phúc Cường: Build and optimize Logistic Regression model.

1.4. Schedule and Milestones

2. Project Execution

2.1. Simulated Scenario Description

To explore customer shopping habits, our models aim to evaluate on the available dataset. The process involves data preprocessing, feature engineering, model selection, hyperparameter tuning, and evaluation. Beside, the most priority is optimizing f1-score while building machine learning models

2.2. Datasets Selection and Description

The data was selected as a suitable dataset for the goal of evaluating the shopping habits of customer. The dataset includes features like "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another. The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session. The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction. The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8. The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

2.3. Model selection

2.3.1. Random Forest Classiffier  
 The Random Forest Classifier is a versatile machine learning algorithm used for classification tasks. It works by creating an ensemble of decision trees, where each tree is trained on a random subset of the data. The final classification is determined by aggregating the predictions of all the trees, usually through a majority vote. This approach improves accuracy, reduces overfitting, and handles large datasets efficiently. Random Forest is commonly used in applications like medical diagnosis, fraud detection, and customer segmentation due to its robustness and ability to handle complex, non-linear relationships in the data.

The reasons make it match the topic requirements:

+ **Handling Complex Relationships:** Random Forest can model complex, non-linear relationships between features, which is essential when dealing with diverse customer behavior.

**+ Robustness:** It reduces the risk of overfitting, providing more reliable predictions even with noisy data.

+ **Feature Importance:** The algorithm can identify which features (e.g., past purchases, demographics) are most influential in predicting purchasing intentions, helping refine marketing strategies.

+ **Scalability:** Random Forest works well with large datasets, making it suitable for analyzing extensive customer data.

2.3.2. Random Vector Machine

Support Vector Machine (SVM) works by finding the optimal hyperplane that separates data points of different classes with the maximum margin. The model aims to create the widest possible gap between different classes, making it robust to overfitting.

We have chosen this model because:

+ **High Accuracy**: SVM is known for its accuracy, particularly in high-dimensional spaces, which is crucial when dealing with complex customer data.

+ **Effective with Limited Data**: SVM performs well even with smaller datasets, making it reliable for predicting purchasing intentions.

+ **Handling Non-linearity**: Through the use of kernels, SVM can handle non-linear relationships in the data, which is often the case in customer behavior analysis.

2.3.3. Logistic Regression

Logistic Regression is a statistical model used for binary classification tasks. It estimates the probability that a given input belongs to a specific class by fitting data to a logistic curve. The model outputs values between 0 and 1, which can be interpreted as probabilities, making it ideal for binary outcomes.

The prevalent about this model:

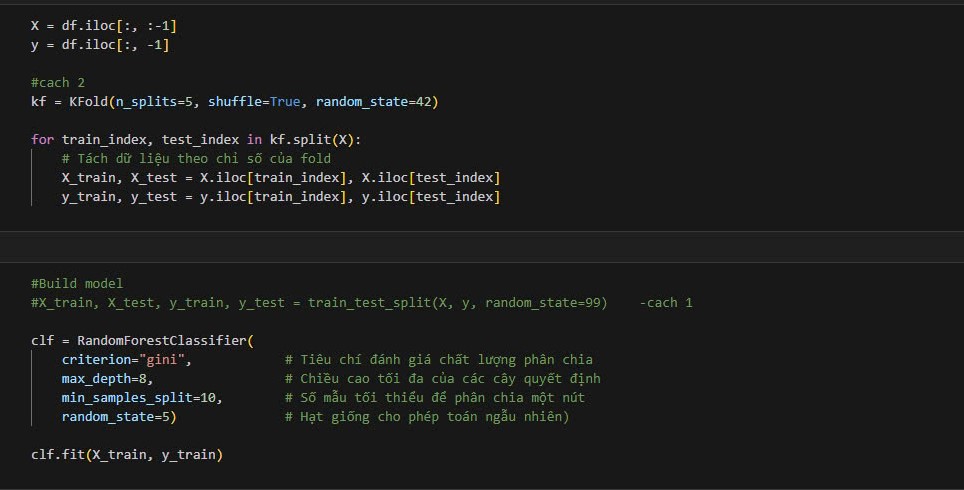
+ **Simplicity and Interpretability**: It’s easy to implement and interpret, providing clear insights into the relationship between features and purchasing intentions.

+ **Efficiency**: Logistic Regression is computationally efficient, making it ideal for large datasets.

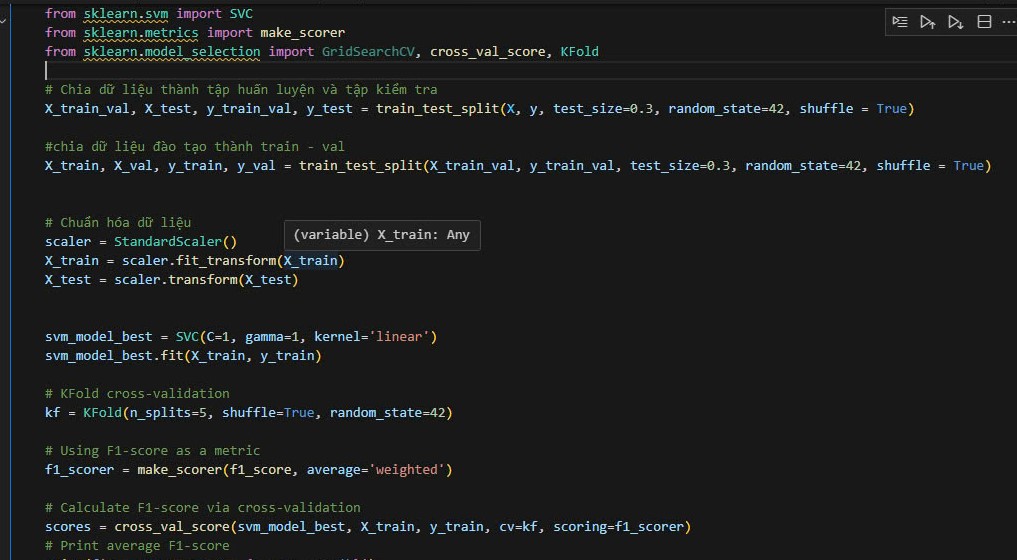
+ **Probabilistic Output**: The model provides probabilities, which can be useful for ranking customers by their likelihood to purchase.

2.4. Build and train models

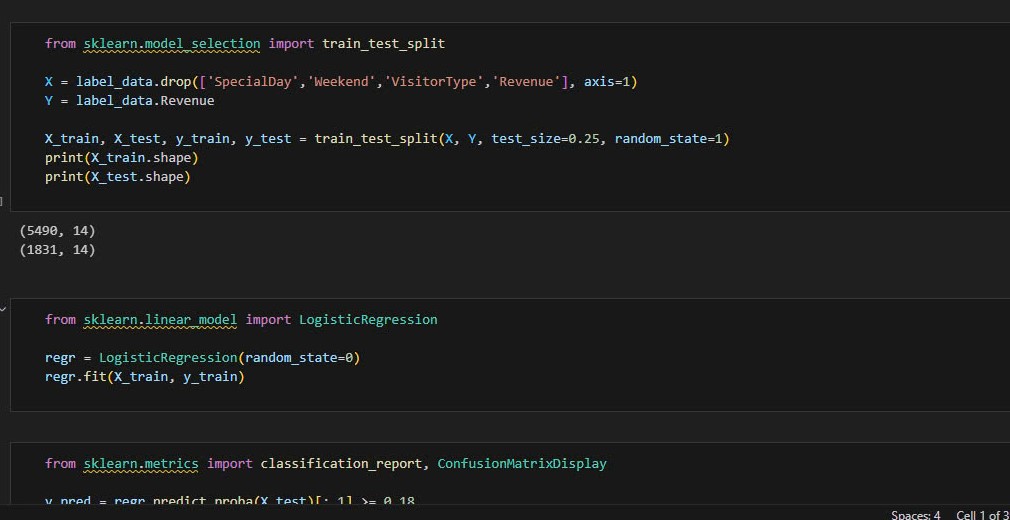
2.4.1. Random Forest Classifier



2.4.2. Support Vector Machine



2.4.3. Logistic Regression



2.5. Evaluate and compare

Eventhough models used the same datasets, the outcomes differentiated among them. Firstly, RandomForestClassifier was the model got the highest f1-score on the first training. Secondly, Support Vector Machine had the lowest f1-score in both pretrain and posttrain. Although observeing f1-score at the middle in rank, Logistic Regression model grew highest in the target score, compared to others, reached at 0.94 in label 1 and 0.72 in label 2. After all, the selected models enhanced in demonstrated statistics after optimizing

3. Results

3.1. The f1-score results from models

- Random Forest Classifier: 0.91 (accuracy), 0.95 (label 1), 0.65 (label 2)

- Support Vector Machine: 0.85 (accuracy), 0.93 (label 1), 0.52 (label 2)

- Logistic Regression: 0.90 (accuracy), 0.93 (label 1), 0.70 (label 2)

3.2. Percentage growth after optimizing

- Random Forest Classifier: 0.92 (accuracy), 0.95 (label 1), 0.7 (label 2)

- Support Vector Machine: 0.89 (accuracy), 0.94 (label 1), 0.6 (label 2)

- Logistic Regression: 0.93 (accuracy), 0.95 (label 1), 0.72 (label 2)

4. Projected Impact

4.1. Accomplishments and Benefits

- Overall, all models have already been improved in f1-score by many ways even getting a tremendous amount of discrepancy on target label.

- Acknowledge solutions which was examples for the projects.

- Trully understand structure of models and the parameters put into it.

- All team members learned how to enhance day-to-day collaboration with each others and how to reach peak performance.

4.2. Future Improvements

- Researching another solutions to optimize models in several aspects

- Develop selected models for getting the most perfect version and sharing with others.

- Learn other ways to visualize the outputs and evaluate datasets and scenarios for the uses in economic.

- Lifelong learning to acknowleadge other solutions (beside machine learning models) in other applications

5. Team Member Review and Comment

|  |
| --- |
| Picture of team members |

|  |  |
| --- | --- |
| NAME | REVIEW and COMMENT |
| Ánh | The F1 score has met the initial target but has not yet reached the higher expectations. |
| Hùng Cường | After the process of exploration and optimization, the results have met the team’s expectations. |
| Phúc Cường | Choosing the appropriate evaluation algorithm for the dataset is crucial. |
| Hậu | Although building the model is easy, optimizing it is difficult. |
| Khiêm | Optimization is not as important as selecting the right model. |
| Lân | Gained a lot of knowledge during the process of completing the project. |

6. Instructor Review and Comment

|  |  |  |
| --- | --- | --- |
| CATEGORY | SCORE | REVIEW and COMMENT |
| IDEA | \_\_/10 |  |
| APPLICATION | \_\_/30 |  |
| RESULT | \_\_/30 |  |
| PROJECT MANAGEMENT | \_\_/10 |  |
| PRESENTATION & REPORT | \_\_/20 |  |
| TOTAL | \_\_/100 |  |