WQD7001 PRINCIPLES OF DATA SCIENCE

GROUP PROJECT PART 2 (G2.9)

PROJECT TITLE:

Analyzing Customer Preferences and Effective Marketing with a Data-Driven Approach to Retail Strategies

Team Member:

No	Matric ID	Name	Roles	E-Portfolio Link
1	S2192852	Jia Hui Wong	Generalist	https://jhwong97.github.io
2	S2191926	Kar Hong Sam	Leader	https://karhong-sam.github.io/
3	22060214	Mei Zhu	Secretary	https://22060214.wixsite.com/christine
4	S2177044	Wing Hong Cheah	Maker	https://winghongjason.github.io/
5	S2158054	Yuan Wei Kam	Oracle	https://yuanweiagatha.wixsite.com/yuan-wei-kam-s-e-por

Table of Contents

1 Executive Summary

The proposal executive report, titled "Analyzing Customer Preferences and Effective Marketing with a Data-Driven Approach to Retail Strategies," aims to help online retailers in understanding customer preferences, optimizing product sales, and developing effective marketing strategies by utilizing data science. The COVID-19 pandemic has resulted in significant growth in the online retail industry, intensifying competition and highlighting the importance of data-driven decision making to remain competitive.

The literature reviews emphasize the value of analyzing customer preferences and market trends in modern retail management. However, many online retailers face challenges in effectively leveraging available data due to limited awareness, proficiency, and financial resources. This project addresses these challenges by applying data science techniques to the superstore dataset, providing actionable insights to overcome the underutilization of e-commerce data.

Following the well-established CRISP-DM methodology, the proposal executive report focuses on conducting a comprehensive exploratory data analysis (EDA) using a superstore dataset. The primary objectives include identifying profitable products, evaluating sales performance across different regions, and developing a machine learning model to predict profitability and improve profit margins. The project encompasses various data-driven strategies, such as sales data analysis, regional sales and profit analysis, and the development of a machine learning model.

Data-driven decision making offers substantial benefits, including informed decision making, identification of business opportunities and risks, and cost savings through optimized operations. By analyzing the superstore dataset, online retailers can gain valuable insights into product sales, regional performance, and potential challenges. These insights enable the development of effective strategies and enhance their competitive position in the market.

In conclusion, the proposal executive report highlights the implementation of the CRISP-DM methodology and the valuable insights gained during the exploratory data analysis (EDA) phase. Key findings include the identification of profitable products, evaluation of regional sales performance, and understanding the contribution of specific customer segments to revenue. The copier from the Technology category was identified as the most profitable item, with the West region showing strong performance and the Central region requiring improvements to boost sales. Moreover, the consumer segment emerged as the primary contributor to overall revenue. The next phase of the project will focus on further developing a predictive model to improve profit margins. By embracing data-driven decision making, online retailers can gain a competitive advantage, make informed decisions, and thrive in the ever-changing online retail landscape.

Furthermore, this report will delve into the development of the proposed model using appropriate analytical tools, the deployment of the solution, and the collection of feedback from end-users.

2 Mechanics - Hardware, Software and Platform Used

In Table, it shows the hardware, software and platform being used to conduct this research project.

Table 1 Hardware, Software and Platform used.

Items	Remarks
Hardware	Personal Computers (PCs)
Software	Visual Studio Code Jupyter Notebook
Platform	Version Control – GitHub Deployment – Streamlit
Coding Language	Python

3 Methodology - Design & Development

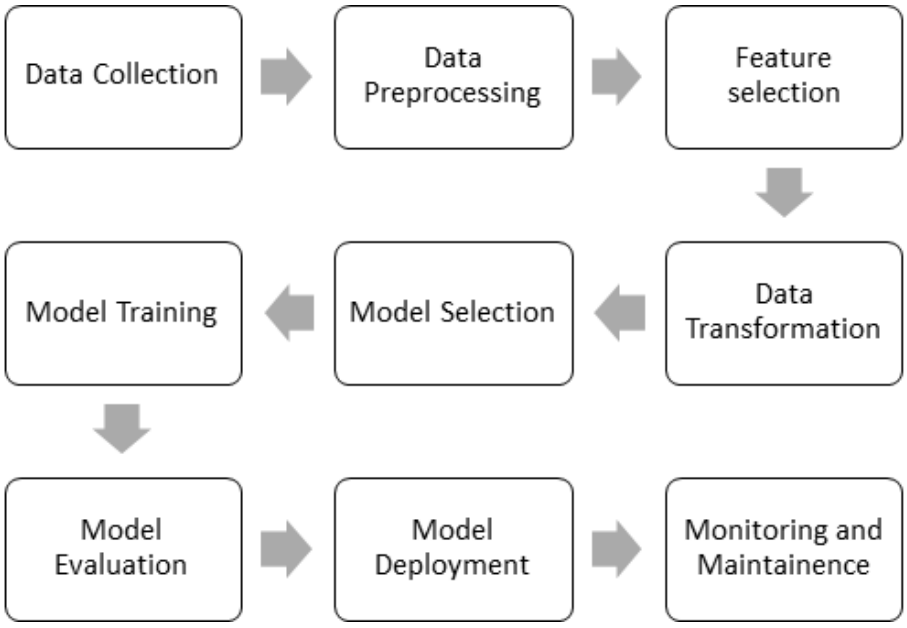

Figure 1: The flow chart of our project methodology.

<u>Data Collection</u>:

The Superstore dataset, which can be found on Kaggle, is a fictitious sales dataset that provides transactional data for a retail company. The dataset consists of 9994 observations and 21 variables, including customer names, product categories, shipping details, order dates, sales amounts, profit margins, quantity sold, discount offered, order priority, shipping mode, region, and sub-category. A new column, named 'Profit Margin (%)" was added into the dataset.

Data Preprocessing:

The cleaned dataset used in the EDA(GP1) will be further used in the data modelling part. The process involved mainly is to remove irrelevant columns and check missing data.

Feature Selection:

To improve the model performance, only the necessary features for classification will be selected. The features selected are market segment, state, region, category, subcategory, sales amount, quantity, and discount. Label encoder will be performed onto the string and category type features.

Data Transformation:

Label encoder will be performed onto the features that consist of string and category type. After the process of label encoder, data correlation and hypothesis testing is carried out to identify the importance of each feature. In addition, a new column called "Gain/Loss" is added. "Gains" is selected if there is a positive profit and "Loss" is selected if there is a negative profit. Both words then map with [0] and [1]. This is done by using the label encoder function.

Model Selection:

In this project, classification type of machine learning algorithms is employed to predict the possibility of gain and loss based on the identified features. This includes the Decision Tree Classifier (DF), Random Forest Classifier (RF), KNN Classifier, Logistic Regression Model (LR), XG Boost (XGB) Classifier and Light GBM Classifier.

Model Training:

The dataset will be split into training and testing dataset, using the ratio of 80:20, where 80 is the training data and 20 is the testing data. Training dataset will be used to train the machine learning model, hyper parameters will be tuned and adjusted to minimize errors and maximize the model's performance.

Model Evaluation:

To evaluate the performance of the model, the machine learning model will be evaluated with accuracy and F1 score.

Deployment:

The selected machine learning model will be deployed into a production environment namely Streamlit. Streamlit is an open-source Python library that allows developers to create interactive web applications for machine learning and data science projects.

Monitoring and Maintenance:

The performance of the deployed model will be consistently monitored to monitor any potential degradation and shifts in data patterns. The model will be updated periodically to ensure that its accuracy and relevance is well maintained.

4 Experiment & Results

4.1 Introduction on Modelling Process

In this project, classification type of machine learning algorithms is employed to predict the possibility of gain and loss based on the identified features. The machine learning algorithms used are shown below:

- Decision Tree Classifier (DF)
- Random Forest Classifier (RF)
- KNN Classifier
- Logistic Regression Model (LR)
- XG Boost (XGB) Classifier
- Light GBM Classifier.

Prior to training the machine learning model, the transformed data is split into training and data using a ratio of 80:20 where 80% is used for training and 20% for testing. A total of six (6) different machine learning algorithms are used to identify the best performance model to be selected for our deployment purposes. The performance evaluation of our model is done by analyzing their accuracy and confusion matrix.

4.2 Results

The performance of each tested model is shown in Table 2. Based on the results, XGB Classifier has the best overall performance in accuracy and F1 score. Therefore, it is selected for our deployment by using a production environment called Streamlit. The differences in performance among those machine learning models could be attributed to several factors such as:

- The Dataset Characteristics used for training and testing the models. The characteristics of the dataset, such as its size, quality, and class distribution, can affect the performance of the models. It is possible that certain models are more suitable for handling the specific characteristics of the dataset, leading to variations in performance.
- Model Complexity. Each of the six models has their own interpretation or methods in capturing the underlying patterns and relationships in the data.

Table 2 Accuracy and F1 score for each tested model.

Model	Accuracy	F1 Score
Decision Tree	0.939470	0.844273
Random Forest	0.947974	0.860963
KNN	0.846923	0.533537
Logestic Regression	0.905953	0.739612
XGBoost	0.949475	0.864793
LightGBM	0.948974	0.862903

5 Plan for Reproducible Research

To reproduce our research results, here are some key expects to be followed.

5.1 Documentation of Research Process

A clear and detailed documentation of our research process following the CRISP-DM is shown in Table 3.

Table 3 Steps to reproduce our research results.

Steps	Elaboration
Data Collection	• The Superstore dataset is obtained from this link: Superstore Dataset | Kaggle
Data Preprocessing	• To remove irrelevant columns and check missing data.
Data Transformation & Feature Selection	• To perform Exploratory Data Analysis • Apply label encoding on features consisting of strings and categorical types. (Non-ordinal data) • To perform Data Correlation and Hypothesis Testing • Select only the relevant features required for classification to enhance model performance. • The selected features include market segment, state, region, category, subcategory, sales amount, quantity, and discount.

	• Added a new column called “Gain/Loss”. “Gains” is selected if there is a positive profit and “Loss” is selected if there is a negative profit. Both words then map with [0] and [1].
Model Selection	• This includes Decision Tree classifier (DT), Random Forest classifier (RF), K-Nearest Neighbors classifier (KNN), Logistic Regression model (LR), XG Boost classifier, and LightGBM classifier.
Model Training	• Split the dataset into training and testing data using an 80:20 ratio, where 80% is used for training and 20% for testing.
Model Evaluation	• Evaluate the performance of the machine learning models using confusion matrix and determine its accuracy and F1 score.
Deployment	• Deploy the selected machine learning model to a production environment, using Streamlit. • Streamlit is an open-source Python library that allows developers to create interactive web applications for machine learning and data science projects.
Monitoring and Maintenance	• Continuously monitor the performance of the deployed model to detect any potential degradation and changes in data patterns. • Regularly update the model to ensure its accuracy and relevance are well-maintained.

5.2 Code Availability

The code used for data analysis, modeling, and other computational aspects of our research can be found in this provided GitHub link (https://github.com/karhong-sam/project-wqd7001-superstore-analysis) which includes scripts and algorithms used specifically for the study. Besides, the provided links also provide the changes made over the code and documentation over the project time which allow users to trace the evolution of the project.

6 Deployment

Streamlit platform was studied for deployment purposes. It is a user-friendly framework that allows for the creation of interactive web applications using Python. It also simplifies the process of building and sharing data science projects.

The project was organized into separate files for better modularity. The visualization part was implemented in the explore.py file, while the machine learning prediction functionality was implemented in the predict.py file. These files were then imported into the app.py file, which serves as the main entry point for running the Streamlit application.

In the explore.py file, there are three tabs available: "Sales," "Profit," and "Comparison." These tabs provide different visualizations and insights related to the data. Additionally, there are

two filters, namely "Year" and "Target," which allow users to customize the displayed data based on their preferences.

On the other hand, the predict.py file contains a form that enables users to select various input parameters related to a product. Users can fill in the form with information such as segment, state, region, category, sub-category, sales, quantity, and discount. Based on these inputs, the machine learning model predicts whether the product will be profitable or not.

To make the project accessible to a wider audience, the deployment was hosted on streamlit.io. This hosting service allows anyone to access and interact with the project without needing to install it locally. By deploying the project on streamlit.io, it becomes readily available for others to try and explore.

Overall, the deployment process involved leveraging the Streamlit platform, organizing the project into separate files, implementing visualizations and prediction functionality, and hosting the project on streamlit.io for easy access and exploration by users.

6.1 End User Feedback

For this project, the concept of our data product is presented to three end users where two are from the financial sector and the other one is from the marketing sector. The feedback from them is summarized in the table below.

Table 4 Summary of end user feedback.

Feedback
Improvement is needed on the prediction part of the data product by providing the users with details about why it is or isn't profitable and could help the user understand what is going on cause on the surface.
Redundancies of information are found such as the "Sales by Category" and "Sales by Sub-Category". Generally, it is sufficient to provide the end user information in the bigger picture.
The formatting of the numbers is inconsistent for all charts or graphs shown in the data product. It is troublesome for a person to look at the financial results.

7 Future Work & Conclusion

7.1 Future Work

Based on the end user feedback, there are several areas for future work to improve the data product:

- **Redundancy Elimination & Formatting Consistent:**
  Eliminating redundancy in the data product, such as merging the "Sales by Category" and "Sales by Sub-Category" sections, to present information in a more concise manner. Utilize data visualization tools that provide formatting options to ensure consistent representation of numbers across different charts and graphs, particularly for financial results, improving readability.

- **Model Interpretability:**
  In business decision-making, the interpretability of a model is crucial. Providing users with detailed information on why certain products are profitable can make it easier to explain to end users. And help users gain a better understanding of the factors influencing profitability, increasing their trust and adoption of the model.

- **Model Optimization:**
  By utilizing techniques such as Bayesian optimization, it is possible to more effectively search for the optimal hyperparameter configurations of a model. This helps in finding the best model structure, learning rate, and other parameters, thereby improving the model's performance. Exploring advanced optimization algorithms such as adaptive learning rate optimization can also be attempted. These algorithms can better adapt to parameter updates and accelerate the model's convergence process.

- **Feature Engineering:**
  Creating new interactive features by combining multiple existing features. These derived features may contain additional useful information, and introducing new features can enhance the model's ability to represent and generalize the data.

- **Data Privacy and Security:**
  Strengthening data privacy and security measures is crucial as data breaches and privacy concerns become increasingly serious. Future work can focus on areas such as data encryption and access control to ensure the confidentiality and integrity of customer data.

In summary, the above content can serve as directions for further development and improvement in the retail business. It aims to enhance the accuracy and profitability of prediction models, enable more personalized and targeted marketing strategies, and maintain competitiveness and sustainable growth.

7.2 Conclusion

The project aims to analyze customer preferences and drive effective marketing strategies in the online retail industry through data-driven approaches. The implementation of the CRISP-DM methodology and exploratory data analysis provides actionable insights. The research findings reveal the profit potential of the printer technology category, the sales advantage in the western region, and the importance of customer segmentation for overall revenue. These insights can be used by online retailers to develop marketing strategies, optimize product portfolios, increase investments in promising areas, and meet the specific needs of targeted customer segments, thereby achieving higher sales performance and profit growth.

The XG Boost (XGB) classifier was selected as the best-performing model for predicting profitability and improving profit margins due to its accuracy and F1 score. It helps online retailers

identify products with potential profitability, avoid low-profit products or ineffective marketing strategies, reduce unnecessary costs, optimize resource allocation, and improve profit margins.

Additionally, a reproducible research plan is provided, including detailed research process documentation and available code, enabling others to replicate the research findings. The availability of research process documentation and code on GitHub ensures research reproducibility. By deploying the findings to a production environment using Streamlit, the research is transformed into an interactive web application, facilitating a better understanding and utilization of the models and analysis results, promoting knowledge sharing, and encouraging innovative development.

In the future, it is necessary to continue monitoring the performance of deployed models, regularly update them, optimize hyperparameters, explore more advanced optimization algorithms, and improve the accuracy of predictive models. This ensures that the models can adapt to constantly changing market demands, provide more reliable decision support for retailers, and reduce subjective guesswork and risks.

8 References & Appendixes – Slide Presentation, Modelling, User Manual etc.

Plotly. Plotly Python Graphing Library. (n.d.). https://plotly.com/python/

Streamlit docs. Streamlit documentation. (n.d.). https://docs.streamlit.io/

Sam, K. H. Analyzing Customer Preferences and Effective Marketing with a Data-Driven Approach to Retail Strategies. https://github.com/karhong-sam/project-wqd7001-superstore-analysis