**Statistical Methods for Data Science**

**Final Project Report**

**Team 5**

Team Members : Gan-Erdene Benderiya & Yash Malviya

**Business Proposition**

The company is based in the United Kingdom and specializes in all occasion gifts to consumers online. Most of the customers are wholesale small or mid companies who may not have an efficient direct-to-consumer delivery model that allows them to sell their goods. Our business proposition is to develop a more robust recommendation system by understanding a user’s patterns and the patterns of the general user base of the company. With a refined analysis of specific products and transaction history, we can analyze when, where, and how people order gifts and suggest relevant products to them at a given time of day or week. If we can predict a person's preference and suggest other items that they may want or as a substitute, we can have more items in their cart along with what they were originally intended to buy, therefore increasing profit for us and the retailers that use our platform.

**Our Motivations**

People are busier than ever, juggling work, family commitments, travel, and personal errands. This often leaves little time for shopping, especially when it comes to choosing the perfect gift. Navigating crowded stores or spending hours online trying to find just the right item can be frustrating and time-consuming. Moreover, many people struggle with selecting the best gift, which can make the shopping experience even longer.

To address these challenges, we propose a multi-level recommendation system that aims to streamline the shopping process and reduce the stress associated with finding gifts. Our solution will use advanced machine learning to recommend products based on individual preferences and shopping patterns, helping users find gifts faster and with less hassle. Additionally, we'll incorporate deep learning to predict optimal times for product launches and promotions, allowing users to make informed shopping decisions.

By implementing this system, we hope to save people time and effort, providing them with a more efficient and enjoyable shopping experience. This will not only make it easier for users to find the right gifts but also help businesses improve customer relationships by offering a more personalized and convenient service.

Our goal is to create a recommendation system that anticipates users' needs, making shopping more efficient and enjoyable, and ultimately leading to happier customers who can spend more time on what truly matters to them.

**Conjectures and Data Analysis**

**Steps of Performed Data Analysis:**

To gather the dataset, we accessed Kaggle.com, a popular resource for data and machine learning enthusiasts. For this project, we chose a comprehensive e-commerce dataset featuring actual transaction data from the UK, which provided the necessary level of detail for our analysis.

**Dataset Overview**

**Variables:** 8

**Observations:** 541,909 (figure 1)

**Key Variables in the Dataset**

The following are some of the primary variables in the dataset and their descriptions:

* **InvoiceNo:** A 6-digit number uniquely identifying each transaction. The prefix 'C' indicates a cancellation.
* **Description:** The name of the product purchased.
* **Quantity:** The quantity of products per transaction.
* **UnitPrice:** The price per unit, in British pounds.
* **StockCode:** A unique 5-digit number for each product.
* **CustomerID:** A unique identifier for each customer.
* **Country:** The country where the customer resides.
* **InvoiceDate:** The date and time of the transaction.

Data Analysis Considerations

The dataset contains a rich set of information that allows us to explore various aspects of customer behavior, product sales, and transactional trends. Here are some key aspects of our data analysis:

Data Preparation: The initial step involved cleaning the data, addressing missing values, and handling duplicates. Although about 1% of the observations were found to be duplicated, this did not impede our exploratory data analysis (EDA).

Exploratory Data Analysis (EDA): We used data visualization and statistical techniques to understand the data. This involved examining patterns in user behavior, analyzing product popularity, and assessing sales trends over time.

Customer Behavior Analysis: We looked at customer patterns, including purchasing frequency, preferences, and behaviors based on country. This helped in identifying customer segments and forming conjectures about purchasing trends.

Product Trends Analysis: By analyzing product descriptions, stock codes, and unit prices, we explored which products were most popular and how they were priced. This informed decisions about product recommendations and potential promotions.

Time-Based Analysis: The InvoiceDate variable allowed us to examine sales trends over time. This was useful for identifying peak sales periods, seasonal trends, and other time-related patterns.

A screenshot of a computer

Description automatically generated

Figure 1. Data mapping. There are 8 total features and at least more than half a million data points (products) ordered.

In our exploratory data analysis (EDA), we uncovered key metrics that provide a preliminary overview of the dataset and its composition. Below are the findings:

Unique Orders: We identified a total of 25,900 unique orders in the dataset, indicating the scope of transactions analyzed.

Unique Users: The dataset comprises 4,373 distinct users, providing a solid foundation for understanding customer behavior and segmentation.

Unique Products: There are 4,224 unique products across all transactions. This diversity allows us to explore a wide range of product categories and preferences.

Countries: The dataset encompasses sales across 38 different countries, offering a global perspective on e-commerce activities and allowing for geographic segmentation.

These preliminary values give us a sense of the scale and variety within the dataset, serving as a foundation for deeper analysis. By examining these unique identifiers—orders, users, products, and countries—we can derive insights into customer purchasing patterns, product popularity, and geographical trends, setting the stage for more advanced modeling and analysis.

A graph showing the number of invoices

Description automatically generated

In Figure 2, the frequency of orders is shown in a bar chart - most people prefer to submit orders on weekdays (especially even days). Based on the shown image, we also assume that during Saturday, the company doesn’t receive any requests.

A graph showing a distribution of events

Description automatically generated In figure 3, the frequency of orders is shown in a bar chart – the people tend to buy more after the mid of the day until work end hours.

A graph of a number of invoices

Description automatically generated

Figure. 4. Frequency of orders per day through month. It shows that people are being more generous between the 5th to 10th of each month. So, if you were born during these days, congratulations, you are really lucky to receive more gifts than people who were born at the end of each month.

A graph of sales

Description automatically generated

Figure. 4. Histogram shows the top 20 sold products. Depending on the popular products we wanted to see more demographic data of the user, unfortunately, we do not have any demographic data, so we decided to analyze the products that we would be working with to draw some correlations between the behaviors of users and products.

A black background with colorful text

Description automatically generated

In Figure 5, a deeper analysis was made for products to highlight what kind of things customers prefer to give to others as presents.

**Conjectures**

1. There could be a significant increase in sales activity towards the end of the year due to the holiday season?
2. Sales are consistently high throughout the day with no significant hourly fluctuations?

Conjecture: these conjectures are explained below with supporting figures forecasted Sales quantity by an XGboost Model.

**Techniques and Methods**

We employed a combination of machine learning and deep learning to achieve our goals. Here's the breakdown of our approach:

**Regression and Classification:**

Regression for forecasting e-commerce sales and determining optimal product selling times.

Classification for generating personalized product recommendations based on user behavior.

**Methods**:

**t-Distributed Stochastic Neighbor Embedding (t-SNE):** For high-dimensional data visualization and to identify consumer behavior patterns.

**Random Forest Algorithm:** Predicted sales trends and uncover factors influencing transactions.

**Deep Learning Models or Ensemble Techniques Like XGboost or LightGBM :** Assessed their effectiveness in improving prediction accuracy and recommendation system performance.

**Product Recommendation System Development:** Created personalized product suggestions based on historical purchase data using machine learning.

**NLP technique:** Created World clouds to investigate the popular items/words/ within the products

**Model**:

Product Recommendation System In this model, the fully mapped data frame was processed using the scipy.sparse csr\_matrix. This creates the user-item matrix to show likelihood of products that they have purchased in the past. After passing through the csr\_matrix, a second method took the outputs of the csr\_matrix method and was used to generate a similar product list. K-Nearest-Neighbors were used to find similar products, thus making it easy to recommend other products. 5 nearest neighbors were used with the algorithm being a brute force and using a cosine metric. Similar product names to the product ID provided would generate a similar list of product IDs, and would then be mapped to a product name.

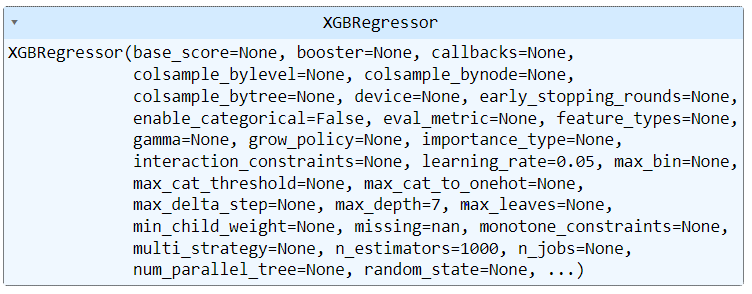
A screen shot of a computer

Description automatically generated

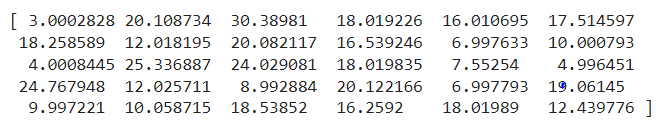
Results are shown in figure 12. Figure. 12. Recommendation model.

Generated model:

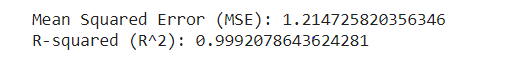
We tried to use a Deep learning Sequential model initially the results were looking good, but as we tried to increase the number of Epochs to capture more complexity the code crash, and it became unresponsive which is why we shifted from a sequential deep learning model to an ensemble Learning such as XGBoost.



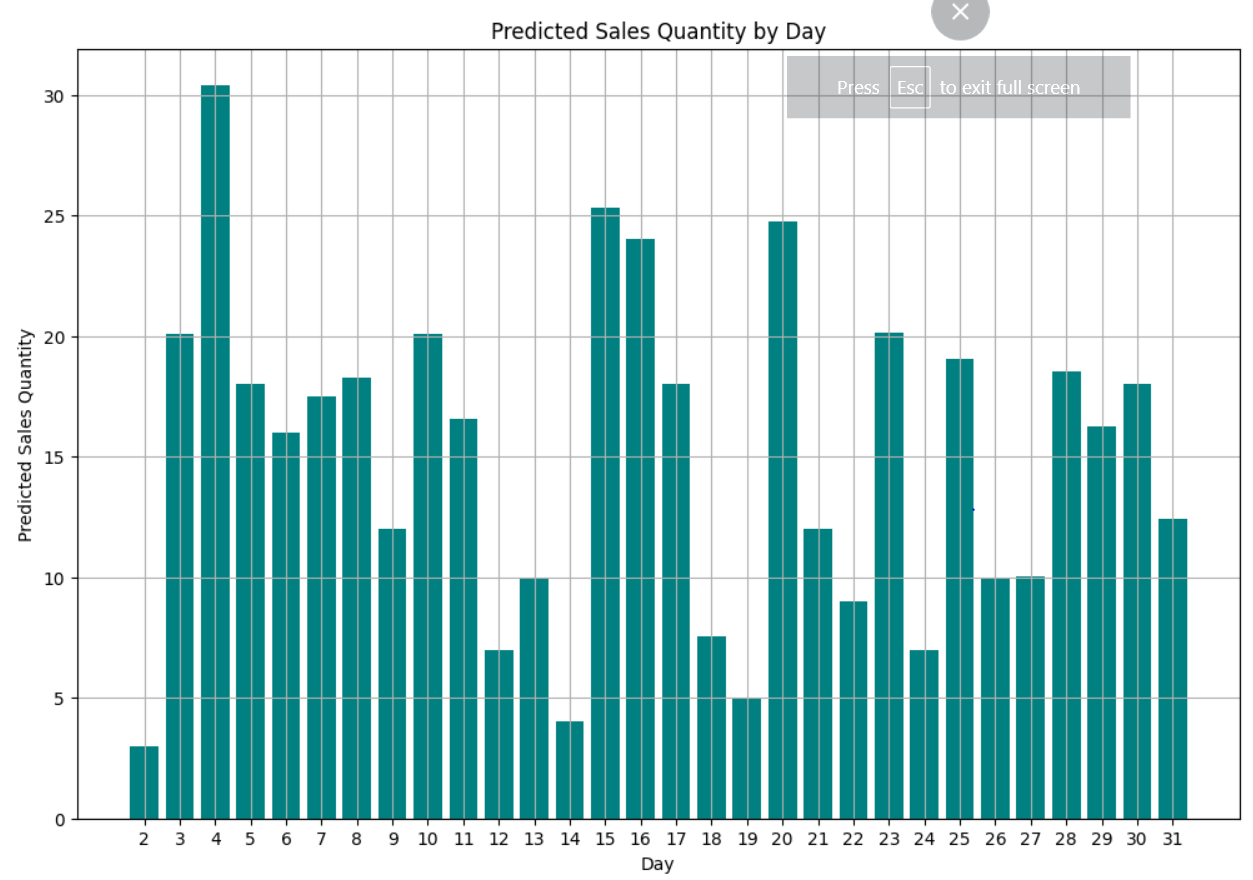
Predicted Quantities :



Random Forest Regressor: We used a Random forest regressor for predicting the same quantities, how every the results were not promising as compared to XGBoost.

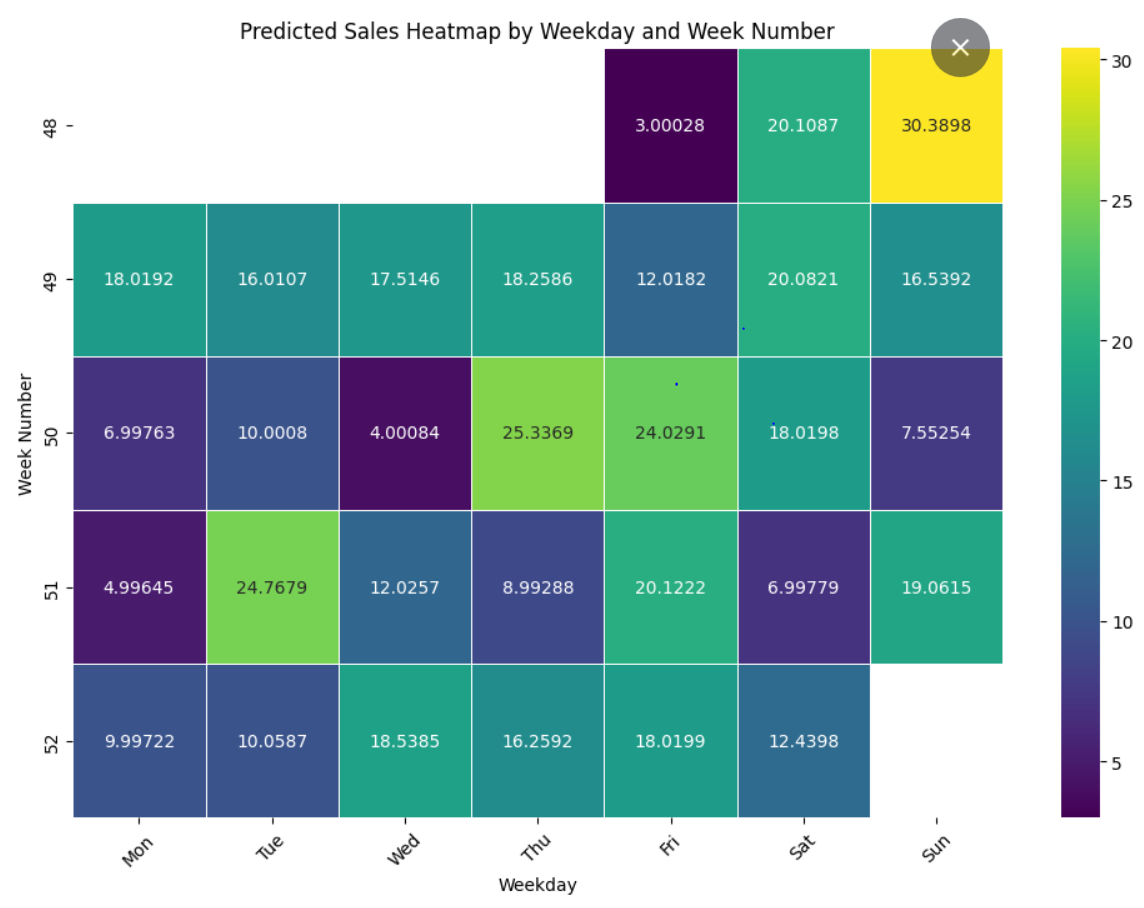


**There could be a significant increase in sales activity towards the end of the year due to the holiday season?**



1. Variability in Daily Sales: The graph shows noticeable fluctuations in the predicted sales quantities across different days of the month. This suggests that sales volume is not consistent but varies significantly from day to day.
2. Peaks in Sales Quantities: There are several days, such as days 2, 8, 16, 18, and 23, where the predicted sales quantities are notably higher than on other days. These peaks could indicate days when promotional activities are in effect, or they might correspond to specific market or consumer behaviors that trigger increased buying activity.
3. Lower Sales on Certain Days: Conversely, there are days like 4, 9, 11, 15, 21, and 29 where the sales predictions are relatively low. These dips could be due to various factors such as weekdays versus weekends, customer payday cycles, or perhaps negative external market influences.
4. Mid-Month Sales Increase: There appears to be a trend where sales increase around the middle of the month, peaking around the 16th to the 18th before gradually tapering off. This pattern could align with mid-month marketing strategies or shopping habits.
5. End-of-Month Variation: Towards the end of the month, the variability in sales predictions increases, with peaks and troughs more pronounced. This could be influenced by month-end financial activities of consumers or stock adjustments by the retailer. so this conjecture came out to be true from the above statements.

**Optimal Selling TImes Based on Week Numbers** :



1. Weekend Sales: There's a noticeable trend where weekends (specifically Sundays) tend to show higher sales in some weeks. For instance, in weeks 49 and 52, Sunday shows significantly higher sales compared to other weekdays. This could be due to weekend shopping habits when consumers typically have more free time for shopping.
2. Weekday Variation: Across various weeks, Wednesday and Friday also show relatively higher sales quantities in some instances (e.g., week 50 for Wednesday and week 51 for Friday). This might indicate mid-week and end-of-week surges possibly due to payday effects or promotional activities typically centered around these days.
3. Consistency Across Weeks: Some days consistently show lower sales across multiple weeks, such as Monday in weeks 48 and 52. This could indicate a typical trend where sales dip at the beginning of the week.
4. High Variability: Some weeks like weeks 50 and 52 exhibit high variability in sales across different days. This could be influenced by specific events, promotions, or other market activities impacting consumer purchasing behavior.

**How do the generated models tie in with our business proposition?**

Our business proposition relies on leveraging machine learning models to optimize the shopping experience and enhance customer satisfaction. Here's how the generated models support this proposition:

**Predictive Analysis for Customer Behavior:** Our machine-learning models are designed to analyze user behavior and transactional history. By examining patterns in past transactions, we can predict when a user is likely to order specific products. This predictive capability helps us anticipate customer needs, allowing us to offer timely suggestions or promotions.

**Recommendation System for Alternative Products:** If a desired item is out of stock, our recommendation system uses machine learning to suggest suitable alternatives. This minimizes the time users spend searching for replacement products and reduces frustration due to product unavailability. By providing these immediate solutions, we create a more seamless purchasing experience, encouraging users to complete their transactions without interruption.

**Streamlined User Experience:** The integration of these models into our business proposition reduces the friction that customers often experience when shopping online. By providing relevant product suggestions and personalized recommendations, we guide users toward products they are likely to enjoy, reducing the need for prolonged browsing and enhancing the overall shopping experience.

**Strengthening Customer Relationships:** Offering personalized recommendations and proactive solutions builds customer trust and fosters a stronger relationship with our brand. As customers find value in our tailored suggestions, they are more likely to return for future purchases, leading to a higher rate of repeat orders through our app or website.

**Increased Customer Satisfaction and Business Profitability:** By delivering a smoother and more efficient shopping experience, we boost customer satisfaction. Satisfied customers are more likely to recommend our service to others and continue shopping with us. This contributes to a higher customer retention rate, ultimately driving profitability for the business.

**How did we work with each other throughout this process?**

To execute our project, we adopted the Agile methodology, focusing on an iterative approach that breaks down complex goals into smaller, manageable tasks. This strategy allowed us to manage time and resources effectively, leading to a streamlined workflow. We divided the overall project into discrete tasks such as identifying datasets, exploring business ideas, learning new algorithms, and testing different concepts. This modular approach not only made our work more efficient but also gave us the flexibility to respond quickly to changes in the project's scope or direction.

To foster inclusive collaboration, we held weekly online meetings accessible to all team members, providing a platform for collective input on agenda and scheduling. These sessions encouraged active participation and allowed each team member to bring their unique ideas and insights to the table. We used various communication tools to stay connected and organized: Google Drive for sharing documents, email for formal correspondence, and WhatsApp for quick, informal communications. Team members conducted individual research and shared their findings via email, which we then discussed and decided upon during our online meetings. Through this collaborative process, we ultimately chose the E-Commerce dataset for our project.

Regular communication was key to our success. We maintained a high level of transparency by frequently reporting progress, sharing results, and seeking help from team members when needed. The Agile approach gave us the flexibility to adapt quickly to changes, allowing us to pivot or adjust our plans as needed to keep the project on track. This flexibility was especially beneficial when we encountered challenges, such as limited computational resources due to using advanced techniques like deep learning and t-SNE, the scarcity of unique business ideas in freely available datasets, and time conflicts among team members.

Collaboration extended beyond formal meetings. Team members actively supported each other by providing feedback, suggesting improvements, debugging, and fixing code. Despite the challenges we faced, the Agile methodology fostered a cooperative and adaptable environment that facilitated problem-solving and innovation. The iterative nature of the methodology allowed us to address issues swiftly and continue progressing toward our project goals, ensuring a successful outcome.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

References [1] https://www.kaggle.com/datasets/carrie1/ecommerce-data