TERM PROJECT – TEAM 13

Section: B Team: 13

Names: Eesha Akula(ea218), Muhammad Adil(ma609), Sulaiman Khan(srk63), Shiman Xu(sx97), Andrea Li(ml756)

Business Understanding

The primary business problem for this **UK-based online retailer** is to optimize sales forecasting, to improve customer retention and accurately assess customer lifetime value (CLV) to better equip to ensure customer satisfaction and maximize profitability by predicting sales patterns.

1. **Customer Segmentation**: To develop personalized marketing strategies, it's critical to segment customers based on purchasing behavior. Segmentation helps identify high-value customers, at-risk customers, and those with the potential for growth. By targeting these segments appropriately, the business can tailor promotions, loyalty programs, and retention strategies to the needs of each group.
2. **Sales Forecasting**: Accurate sales forecasting is fundamental for managing inventory, setting revenue targets, and planning marketing campaigns. Predicting sales on a weekly basis allows for better resource allocation, including optimizing stock levels, avoiding overstocking or understocking, and preparing for high-demand periods (such as holidays).
3. **Customer Lifetime Value (CLV) Prediction**: CLV helps the business estimate the total revenue a customer is likely to generate over the course of their relationship with the company. By accurately predicting CLV, the business can allocate resources more effectively, focusing on customers who will provide long-term value.
4. **Churn Prediction**: Churn prediction enables the business to identify customers likely to stop purchasing, giving them an opportunity to intervene with personalized offers, incentives, or re-engagement campaigns. This proactive approach minimizes revenue loss from customer churn, which is particularly important in an industry with high customer acquisition costs.

The goal of the data mining solutions presented here is to help the retailer increase retention, improve targeting efficiency, and ultimately drive revenue growth.

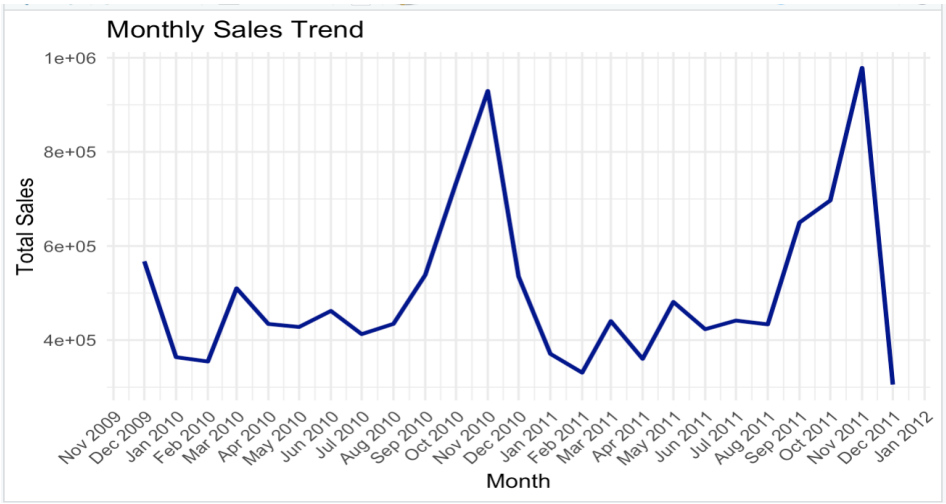
Data Understanding

The dataset includes **1,067,371 transactions** from a UK-based online retailer between 2010 and 2011. Each row represents a purchase for each product, and the key attributes include:

- **Invoice**: A unique identifier for each transaction.
- **StockCode**: Product code for items sold.
- **Description**: A brief description of the product.
- **Quantity**: Number of units purchased in a transaction.
- **InvoiceDate**: Date and time of the purchase.
- **Price**: Price per unit of the product.
- **Customer ID**: Unique identifier for customers (some missing values, ~45% of customers).
- **Country**: Customer's country (with 92% of transactions being from the UK).

As we can observe, the number of columns that we have are not quite extensive, despite this one of the major reasons for selecting the data is that we wanted ***to replicate the real-world scenario*** *where readily access to customer demographics data is a privilege*. As mentioned earlier, the data is for the online sales, hence we do not see any other features such as store no., etc. Also, since it is a UK based retailer, we could observe that more than 94% of the data pertains to UK and the rest of it belongs to the countries where the retailer operates.

Additionally, since it was retail data, we did perform an EDA after data cleaning (summarized in next step) to understand the underlying patterns (seasonality) and identify the months of increased sales as visualized below:



Data Preparation

The preparation involved several steps of data cleaning:

1. **Data Cleaning (1) – "StockCode" :** Rows where the **StockCode** contains only letters were filtered out, as these likely represent non-transactional or irrelevant entries. This step ensures that the data used for analysis includes valid product codes, facilitating better segmentation and inventory tracking.
2. **Data Cleaning (2) - Null Value Treatment for "Description":** We identified missing **Description** values and replaced them with descriptions corresponding to the same **StockCode** when available. Rows that still had missing descriptions after this process (360 rows) were dropped to maintain data integrity for product analysis.
3. **Data Cleaning (3) - Null Value Treatment for "Customer_ID":** Rows with missing **Customer_ID** were categorized under the label "Unknown," since we had significant number of rows and these would be beneficial for sales forecasting. This allowed us to retain the rows while acknowledging incomplete customer data, maintaining consistency across the dataset.
4. **Data Cleaning (4) - Outlier Treatment:** Outliers in **Price** and **Quantity** were capped at the 95th percentile, replacing extreme values with more reasonable figures. This step helped to prevent skewed analysis and modeling outcomes caused by outlier transactions.
5. **Data Cleaning (5) - Miscellaneous Cleaning:** We trimmed white spaces in the **Description** column to ensure consistency in product descriptions. This step was essential for grouping similar products correctly and enhancing the accuracy of further analyses.

   94% of the data comes from the UK, which could lead to geographic bias. To avoid skewing our models toward other regions, only UK data was retained for further analysis. After cleaning, **944,581 rows** of UK data remained, which was sufficient for building robust models.

6. **Data Cleaning (6) - Duplicate Rows:** We identified and removed **34,272 duplicate rows** from the dataset, retaining only the first occurrence of duplicated transactions. This reduced redundancy and ensured that each transaction in the dataset was unique, resulting in more accurate analysis.

7. **Data Cleaning (7) - Text Removal from StockCode:** StockCodes ending with letters (often representing product variations) were standardized by assigning a unique number to each variation. This step made it easier to perform numeric operations and maintain consistency in product identification across the dataset.

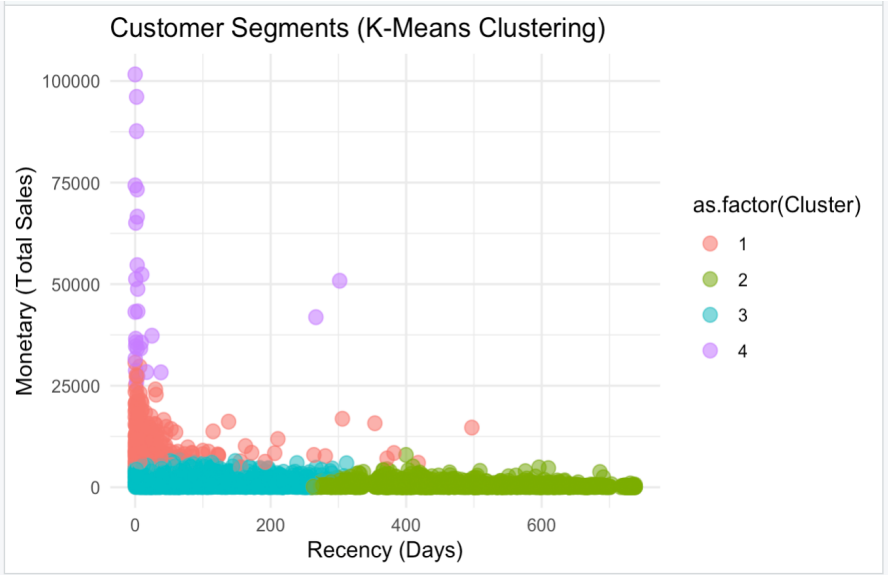
Modeling

1. **Customer Segmentation (RFM Analysis)**: Using **Recency, Frequency, and Monetary** (RFM) metrics, we applied **K-Means clustering** to group customers into four distinct clusters. These clusters help us understand customer behavior better, facilitating targeted marketing and retention strategies.

    - **Cluster 1: Engaged and High-Spenders** – Customers with frequent purchases and high monetary value.
    - **Cluster 2: At-Risk Customers** – Customers with high recency, indicating they haven't purchased recently.
    - **Cluster 3: Potential Loyal Customers** – Moderate frequency and monetary value but engaged customers.
    - **Cluster 4: Recently High Spending Customers** – Customers with a sudden increase in spending.

    Elbow Method was used to determine the optimal number of clusters (K=4). The segmentation helped in understanding different customer personas, which could be directly used for targeted marketing campaigns. We started with 5 clusters, however the clusters that were made then had very minimal difference in their cluster centers, hence we proceeded with having 4 clusters.

Visualization/Output:



2. **Sales Forecasting**: We built sales forecasting models using the following models:

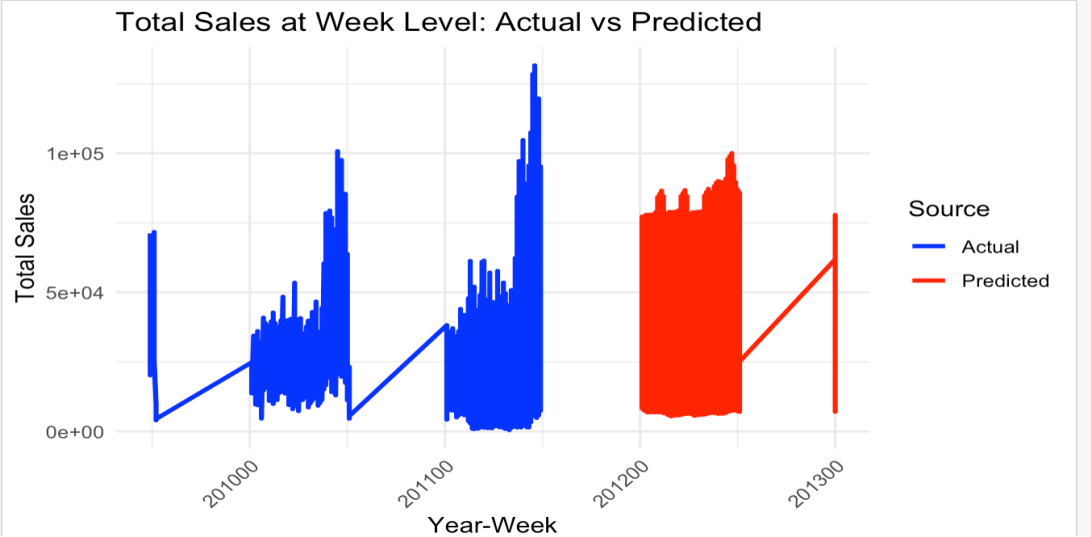
   Specific Data Preparation: Since we did not have sufficient transactional data or a missing customer ID for all customers, we had to omit some of the customers while clustering, hence they did not have any cluster mapped against every customer. However, while predicting sales having such entries is important in eliminating bias and predicting the sales coherently. Hence, we have calculated the ratio of clusters before any data transformation followed by replacing all NA clusters with corresponding clusters without impacting the overall ratio of customers across each cluster.

   - **Linear Regression**: Initially, the linear regression model provided an **R-squared** of around **0.77**, indicating that 77% of the variability in sales was explained by the model. We used time features such as **Year, Month, Week**, and **Cluster** for prediction along with interactions and polynomial variables.
   - **Random Forest Model**: After improving the model with lag features (lag of 1 week, 4 weeks, and 52 weeks), a **Random Forest** model was trained, achieving a higher **R-squared** of **0.83**. Random Forest was chosen due to its ability to capture complex relationships between the features.
   - **Cross-Validation**: We employed 5-fold cross-validation to prevent overfitting. The final model had an **RMSE of 14,000** and a **R-squared** of 0.84 for the Random Forest model

with cross-validation indicates that our model captures a large portion of the variance in the sales data. The RMSE of 14,000 is acceptable given the scale of the sales data.

Output: As you can infer from below from the output of **Random Forest with 5 fold cross validation**, the model captures the trend of increased amount of sales from year 2010 to 2011 and to finally the predicted year of 2012. However, if we look closely, we are able to understand that the seasonality spike is not captured in a precise manner. The major reason for this being the limited number of columns that we have had, with additional data about online browsing patters, customer demographics the existing model should be able to capture all such patterns much more precisely

Learning: One of the most important learnings that we could deduce from this sales forecasting was that having a target variable of total sales (which is a product of Price and quantity) it is not ideal to use the variables Price and Quantity since it will lead to a perfect multicollinearity. Hence, we have used the other relevant metrics along with the metric – "Cluster" which is a direct combination of Recency, Frequency and Monetary.



3. **Customer Lifetime Value (CLV) Prediction**: CLV was modeled using **linear regression** with features such as the number of purchases, recency, and average price/quantity. The CLV model achieved an **R-squared of 0.98**, demonstrating high accuracy in predicting future customer revenue.

   Evaluation: The model was evaluated using **RMSE (1892.28)** and **MAE (762.29)**. These metrics were used to understand how closely the predicted CLV matched the actual CLV.

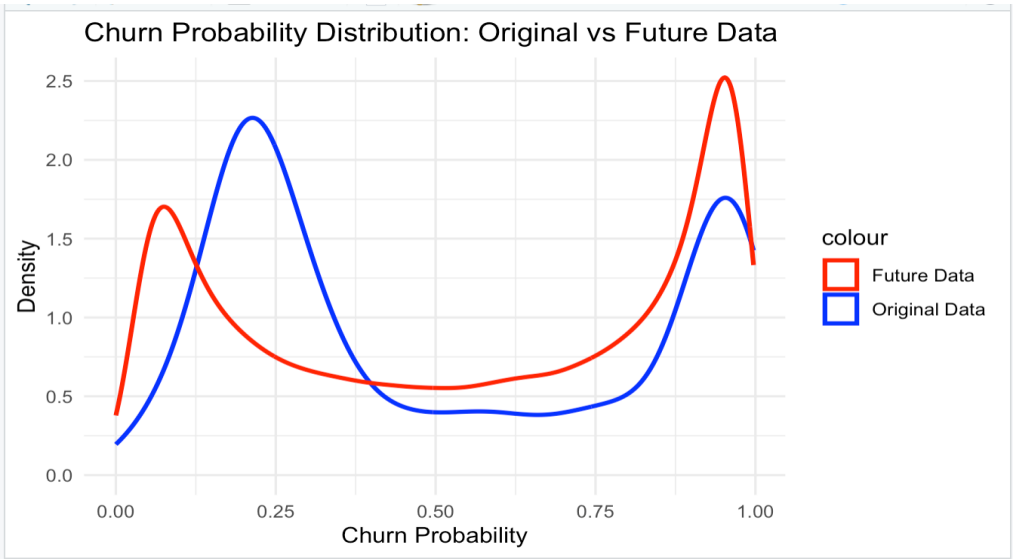
4. **Churn Prediction:** Using **logistic regression** and **Ridge regularization**, we predicted customer churn based on **Recency, Frequency, and Monetary** features. The **AUC** for the model was **0.998**, suggesting a highly accurate churn prediction model. Some of the main reasons why this is possible is summarized below –

    - For churn prediction, we had to remove the customers who did not have a customer ID, leading to removal of a significant number of such customers hence a smaller data set to train and test on.
    - Additionally, the lack of customer behavior data further adds to increased dependency of the prediction on only a small set of variables, causing the data to overfit for any other data set provided.

    Model Choice: Ridge regression was selected to reduce overfitting and penalize complex models. This improved the model's generalization to new data.

    Output:

        - ROC Curve illustrating the model's classification ability.
        - Confusion matrix showing the true positive, false positive, and false negative rates for churn prediction.
        - Also, we did plot a trend chart to see the Probability distribution of the original and the future data (random generated data)



As observed above, we can see that the original data is more distributed between 0 and 1, however the future data has more bias towards 1 and 0s.

Deployment

Below we have tried to summarize each of the challenge in deployment and its solution:

1. **Data Integration and Scalability**

- **Challenge**: A major hurdle in deploying machine learning models is smoothly integrating them with existing data infrastructure. Retail data often comes from multiple sources, such as POS systems, inventory databases, and CRM tools. Ensuring that models have access to real-time or near-real-time data is critical for accurate predictions and timely insights.
- **Scalability**: As the business expands, the amount of data will increase. The models need to be able to scale effectively to handle this growth without sacrificing performance. Systems like batch processing, real-time streaming, and cloud platforms (like AWS or Google Cloud) will be crucial.
- **Solution**: Building a strong data pipeline to handle the extraction, transformation, and loading (ETL) processes is key. Using cloud storage solutions like Amazon S3 or Google Cloud can help manage large amounts of data efficiently.

2. **Model Generalization and Maintenance**

- **Challenge**: While a model may perform well during development, it may struggle to maintain its accuracy in a live environment, where it encounters new, unseen data. Customer behavior in retail can shift rapidly due to seasonal changes, trends, and unexpected external factors like economic shifts or holidays.
- **Model Drift**: Over time, as customer behavior changes, models can experience drift, becoming less accurate. For example, a churn prediction model might start incorrectly flagging loyal customers as at risk of leaving.
- **Solution**: It's essential to continuously monitor model performance. Setting up a feedback loop to evaluate metrics like RMSE for sales forecasting or AUC for churn prediction will help catch performance drops early. Regularly retraining the models with updated data and using automated pipelines can keep models fresh and relevant.

3. **Real-World Changes and Unforeseen Events**

- **Challenge**: External factors, such as shifts in market conditions, new regulations, or unexpected events like the COVID-19 pandemic, can heavily impact model performance. For instance, sudden changes in consumer behavior or supply chain disruptions can make sales forecasts less reliable.

- **Solution**: Develop flexible models that can adapt or be retrained when needed. For unpredictable situations, using hybrid forecasting—where expert judgment is combined with model predictions—can reduce the risk of poor predictions.

4. **Ethical and Privacy Concerns**

- **Challenge**: Although the current model only uses transactional data, a more robust model would likely require access to customer demographics and behavioral data. However, this raises significant ethical concerns around privacy, data security, and potential bias in predictions.
- **Bias and Fairness**: There's a risk that models could unintentionally reinforce biases. For example, if certain demographic groups are underrepresented in the data, the churn model might unfairly classify them, leading to less marketing outreach or fewer retention efforts directed at them.
- **Solution**: Regularly audit models for fairness and evaluate performance across different demographic groups. It's also critical to comply with privacy regulations such as GDPR and CCPA, which protect customers' data rights. Using anonymization techniques and ensuring data is securely encrypted can help mitigate privacy risks.

Appendix

Data Source - https://www.kaggle.com/datasets/shashanks1202/retail-transactions-online-sales-dataset?resource=download

Information Sources

- https://towardsdatascience.com

- https://machinelearningmastery.com/planning-your-data-science-project/

Work Division:

- **Data Cleaning & EDA** – Sulaiman
- **Customer Segmentation** – Sulaiman & Eesha
- **Sales Forecasting** – Sulaiman & Adil
- **CLV** – Sulaiman & Andrea
- **Churn** – Sulaiman & Shiman