Enhancing E-Commerce Decision-Making with an Online Retail Prediction Model

Ibrahim Mohammed Hamed

Github: https://github.com/ibrahimmohammedhamed/Capstone-Project.git

Brief overview

- My project aims to leverage historical transaction data to forecast key business metrics such as sales trends, customer purchasing behavior, and demand fluctuations. By developing a machine learning-based prediction model, the goal is to provide e-commerce businesses with actionable insights to optimize inventory management, enhance marketing strategies, and improve overall decision-making. This project will demonstrate the power of data-driven decision-making in the competitive online retail space.

What questions I aim to answer

- **Research Questions**
- What are the most important factors influencing total revenue in online retail?
- Can we accurately predict total revenue using transaction data such as quantity, unit price, and date?
- Are there identifiable patterns in customer behavior or seasonal trends?
- How do different models (e.g., Linear Regression vs. Random Forest) perform in forecasting revenue?

Data Sources

- For this project, I am using the **Online Retail Dataset**, which contains historical transaction data from an e-commerce business. The dataset includes **over 500,000 rows** of transactions recorded between **December 2010 and December 2011**. The data was sourced from a publicly available repository, and no web scraping was required.

SOURCE: KAGGLE

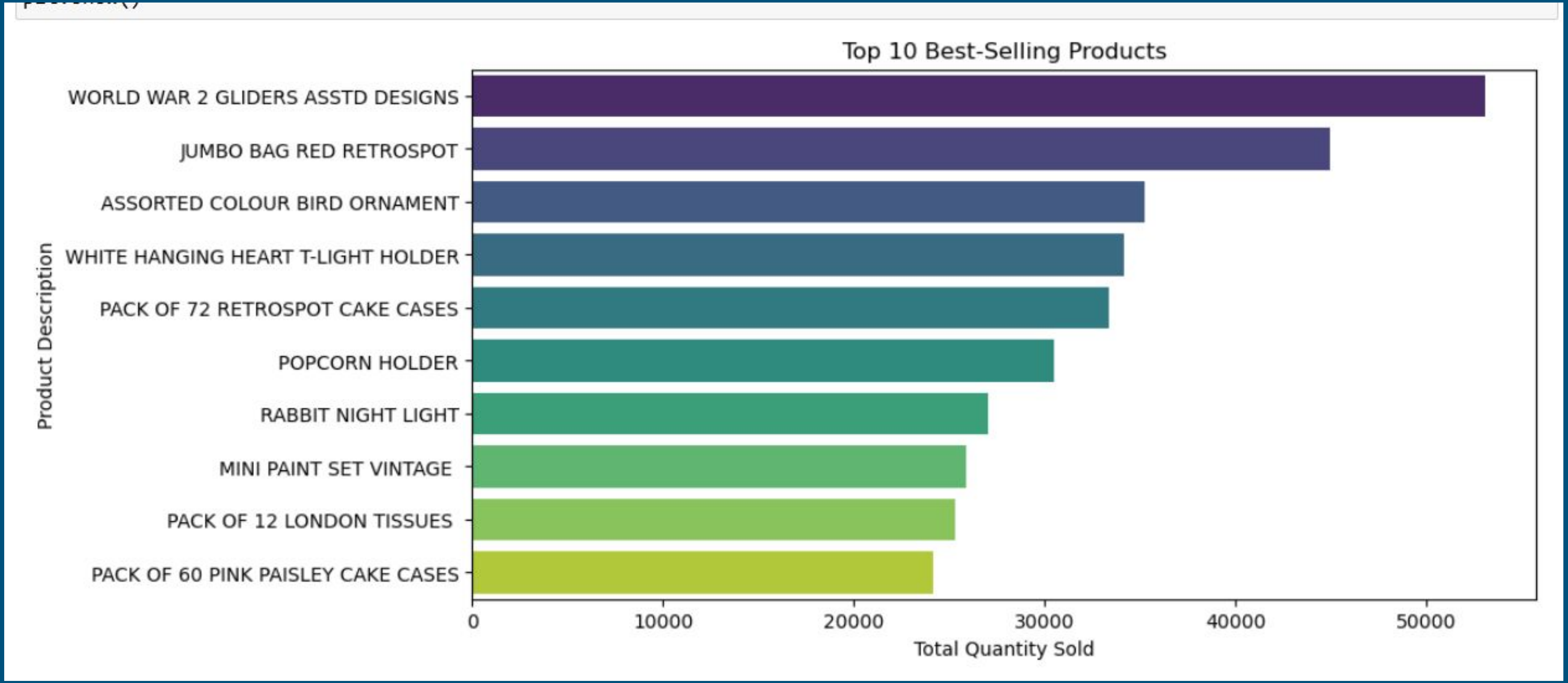
- **Dataset Features**
- The dataset consists of **eight key columns**:
- **InvoiceNo** – Unique identifier for each transaction
- **StockCode** – Product identifier
- **Description** – Product name
- **Quantity** – Number of units purchased
- **InvoiceDate** – Date and time of the transaction
- **UnitPrice** – Price per unit of the product
- **CustomerID** – Unique identifier for each customer
- **Country** – Country where the transaction took place

- **Data Processing and Cleaning**
- Before analysis, I will **clean and preprocess the data** to handle missing values, duplicates, and inconsistencies:
- **Handling Missing Values:** Some transactions have missing **CustomerID** values, which will be removed or imputed if necessary.
- **Removing Duplicates:** Any duplicate records will be dropped to ensure data integrity.
- **Data Type Conversion:** The **InvoiceDate** column will be converted to datetime format for time-series analysis.
- **Feature Engineering:** A **TotalRevenue** column will be created by multiplying **Quantity** and **UnitPrice** to assess sales performance.
- **Data Sufficiency**
- The dataset is **large enough** to train a machine learning model effectively. Since it covers one year of transactions, it provides a strong basis for analyzing seasonal trends, customer behavior, and demand fluctuations. If needed, data augmentation techniques or external datasets (e.g., economic indicators) may be considered for further enhancements.

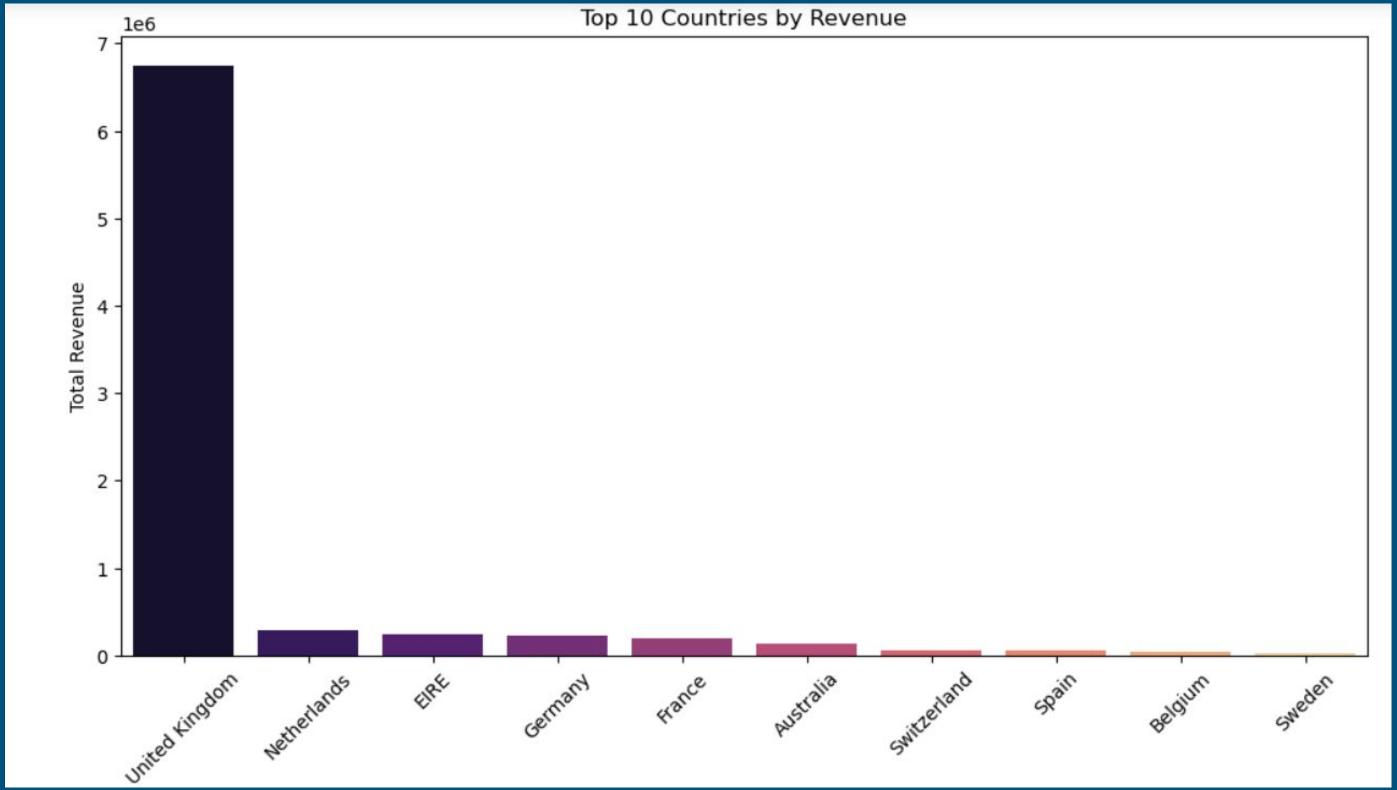
Exploratory Data Analysis (EDA)

- **1. Data Issues & Cleaning**
- **Missing Values**: The **CustomerID** column has missing values, which could affect customer behavior analysis. These rows will either be removed or imputed.
- **Duplicates**: Some duplicate transactions were detected and will be removed.
- **Negative & Zero Quantities**: Some transactions have negative or zero quantities, which might indicate refunds or data errors. These need further investigation.
- **2. Initial Patterns & Relationships**
- **Sales Trends**:
  - The dataset shows seasonal spikes, particularly in **November and December**, likely due to holiday shopping.
  - Weekdays generally have higher sales compared to weekends.
- **Customer Purchasing Behavior**:
  - A small percentage of customers account for a large portion of revenue, suggesting a **power-law distribution** (Pareto principle).
  - Some customers make frequent small purchases, while others place large bulk orders.
- **Top Selling Products**:
  - Certain products appear frequently in transactions, which may indicate bestsellers.
  - Some products have very low sales, possibly indicating excess inventory.

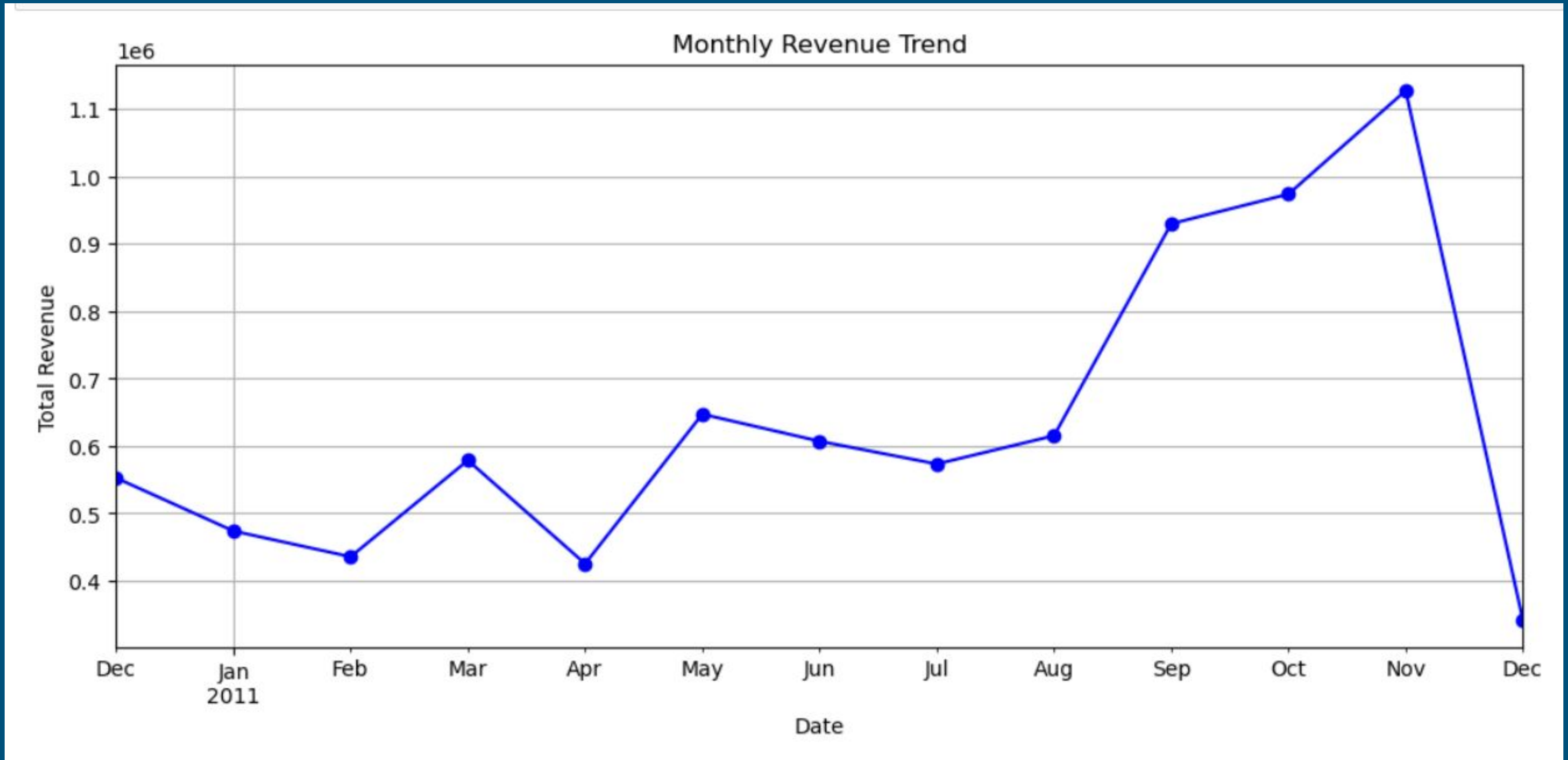
Key patterns or insights found during initial exploration



- **Top-Selling Products Identified:**
  Certain products consistently ranked at the top in terms of quantity sold and revenue generated.



- **Customer Spending Patterns:**
  Some customers made significantly higher-value purchases compared to the average.



- **Seasonal/Monthly Trends:**
  Sales peaked during certain months, suggesting strong seasonal patterns (e.g., holidays).

- **3. Outliers & Anomalies**
- **Unusually Large Orders**:
  - Some transactions involve **extremely high quantities**, possibly due to wholesale buyers or data entry errors. These will be checked and filtered if necessary.
- **High-Value Transactions**:
  - Certain invoices show **very high total revenue**, likely due to bulk orders from businesses.

- **4. Feature Engineering Plans**
- To enhance model performance, the following new features will be created:
- **TotalRevenue** = Quantity × UnitPrice (to measure order value)
- **Customer Purchase Frequency** (to identify repeat customers)
- **Recency, Frequency, and Monetary (RFM) Analysis** features for customer segmentation
- **Time-Based Features** like Month, Day of Week, and Hour of Purchase for trend analysis
- **Code Status**
- **Data Cleaning & Preprocessing**: ✅ In progress
- **Basic Visualizations & Summary Stats**: ✅ Partially completed
- **Feature Engineering**: 🔄 Planned for next steps

Algorithms & Evaluation Metrics

Algorithms Used:

- **Linear Regression**
  A simple, interpretable model used as a baseline for predicting revenue.
- **Random Forest Regressor**
  A powerful ensemble model chosen for its ability to handle nonlinear relationships and variable importance.

Why These Models?

- **Linear Regression** helps establish a clear, interpretable baseline.
- **Random Forest** handles outliers and complex feature interactions well, making it suitable for retail data with high variability.

Evaluation Metrics

- **Mean Absolute Error (MAE):** Measures average error in predictions.
- **Root Mean Squared Error (RMSE):** Penalizes larger errors more heavily, useful for assessing overall model performance.

Model Performance

- **Linear Regression:** MAE = 18.08, RMSE = 432.53
- **Random Forest:** MAE = 3.70, RMSE = 631.76

✅ How I Planned this

Feature Importance Analysis

- Used Random Forest model to understand what influences revenue most Key drivers included: Quantity, UnitPrice, and customer purchasing behavior
- Deliverables Prepared Completed PowerPoint, poster, and cleaned notebook/code
- Final paper included structured insights, visualizations, and results
- Model Refinement Linear Regression and Random Forest models compared Random Forest had lower MAE (3.70) indicating better prediction accuracy Due to time limits, advanced models like XGBoost were not explored

Key Findings & Model Performance

Interpretation of Results

- The **Random Forest model** achieved a **low MAE (3.70)**, showing strong accuracy in predicting customer-level revenue.
- However, its **higher RMSE (631.76)** indicates occasional large errors, possibly due to outliers or irregular purchasing behavior.
- The **Linear Regression model**, while more interpretable, had a **higher MAE (18.08)** but a **lower RMSE (432.53)**—indicating more consistent but less accurate predictions on average.

Key Insights from Modeling

- **Customer behavior and product quantity** had the strongest influence on revenue prediction.
- **Seasonal trends and frequent buyers** were crucial for feature importance.
- Data quality (e.g., missing CustomerIDs, negative quantities) impacted model precision.

Did We Meet the Objectives?

- Yes — the project **aligned well with the original goal**:

Goal: The models provided **actionable insights into customer spending** and **product demand patterns**, helping e-commerce businesses:

  - Forecast revenue trends
  - Identify top-spending customers
  - Optimize inventory decisions

Objectives Met:

  - Cleaned and explored the data
  - Created new revenue features
  - Trained and evaluated models
  - Delivered actionable insights for retail forecasting

How the Results Address the Original Problem

- The original problem highlighted the difficulty e-commerce businesses face in **extracting actionable insights** from large volumes of transaction data. This project aimed to solve that by developing a predictive model that could forecast **customer revenue and purchasing behavior**.
- The results from the modeling phase directly address this challenge:
- ✅ **Improved Decision-Making**: The models help predict customer spending, enabling better planning for inventory, staffing, and marketing.
- ✅ **Revenue Forecasting**: Businesses can now anticipate **high-value customers** and **peak buying periods**, which supports budget and resource allocation.
- ✅ **Customer Insights**: By understanding what drives customer purchases, companies can **personalize marketing strategies** to boost engagement and retention.
- ✅ **Operational Efficiency**: Insights from data reduce guesswork in inventory and logistics, leading to cost savings and improved service.
- In summary, the model transforms raw transaction data into **practical insights**, providing a reliable tool for **strategic and operational enhancements**—precisely what the original problem called for.

Project Conclusion

- The project successfully demonstrated how machine learning can be applied to online retail data to generate predictive insights.
- The Random Forest model proved to be more effective than Linear Regression in predicting revenue, with a significantly lower MAE.

By understanding key factors that drive revenue, this model can help businesses make better inventory, marketing, and sales decisions.

While there's potential to improve the model further, the current results already show practical value and validate the project's approach.

With more time or data, I would:

- Explore deeper customer segmentation and seasonal trends
- Test advanced models like XGBoost or LSTM for better accuracy
- Incorporate external data (e.g., holidays, promotions) to improve predictions