**2024S-T2 BDM 2203 - Big Data Visualization for Business Communications 01 (DSMM Group 1)**

**Project Report**

**Total Revenue Prediction of Electronic Sales**

**---------------------------------------------------------------**

**Submitted by:**

Jobina Joy - C0924759

Anand Chathananickal Sajeevan - C0928396

Abhishek Mungath - C0928517

Jeffin Kochurani Ravi - C0925106

**Submitted to:**

Prof. Ishant Gupta

Lambton College Mississauga

Date : August 16, 2024

**Introduction**

Predicting revenue for a global electronics retailer is vital for optimizing inventory management, pricing strategies, and marketing initiatives. This report details the analytical process used to forecast revenue from a complex dataset encompassing sales transactions, product details, customer demographics, and store information. By integrating and analyzing these data sources, we aim to develop a robust model for revenue prediction.

**Data Preparation**

The initial step in our analysis involved importing data from multiple CSV files, each representing different facets of the retail business. The sales data file contained information about individual transactions, including order numbers, quantities, and unit prices. We supplemented this with product data, which detailed product names, brands, and categories. Additionally, customer data provided demographic information such as location, and store data included attributes related to store locations.

To create a unified dataset, we merged these individual files. The sales data was combined with product data using the 'ProductKey', which enabled us to integrate transaction details with product attributes. Similarly, we merged sales with customer data using the 'CustomerKey' and with store data using the 'StoreKey'. This comprehensive dataset provided a holistic view of the sales transactions within the context of product details, customer demographics, and store attributes.

Data cleaning was a crucial step in preparing the dataset for analysis. Missing values were addressed systematically: for example, the 'Delivery Date' field, which was often missing, was imputed using the average time between order and delivery dates. Other columns with missing values, such as 'State Code', were filled with placeholder values or handled by removing incomplete rows where necessary. Outliers in the 'Quantity' column were managed by capping extreme values based on the interquartile range, thus preventing skewed results.

Feature engineering was employed to enhance the dataset's usability. We calculated the 'Order Processing Time' by determining the difference between 'Delivery Date' and 'Order Date', which provided insights into order fulfillment efficiency. Temporal features such as 'Order Weekday', 'Order Month', and 'Order Year' were extracted to capture seasonal and time-based patterns. Additionally, 'Total Revenue' was computed as the product of 'Quantity' and 'Unit Price USD', reflecting the revenue generated from each transaction.

Data encoding and scaling were performed to prepare the dataset for machine learning algorithms. Categorical features, including 'Product Name', 'Brand', and 'Category', were converted into numeric values using label encoding. This transformation is essential for algorithms that require numerical inputs. To ensure all features contributed equally to the model, numerical features were standardized using StandardScaler, which prevents discrepancies due to differing scales.

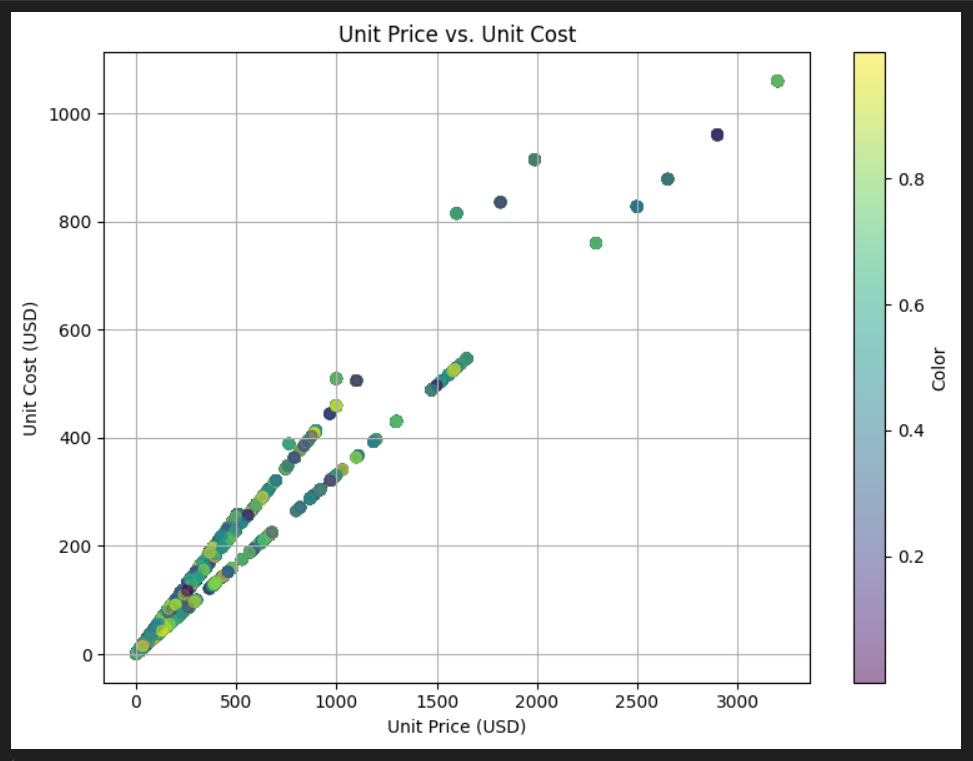
**Data Analysis & Visualization**

Descriptive statistics provided an initial understanding of the dataset. Summary statistics for numerical columns revealed the central tendencies and variability within the data. This analysis included measures such as mean, median, standard deviation, and quartiles, which offered insights into the distribution of transaction values, quantities, and revenue.

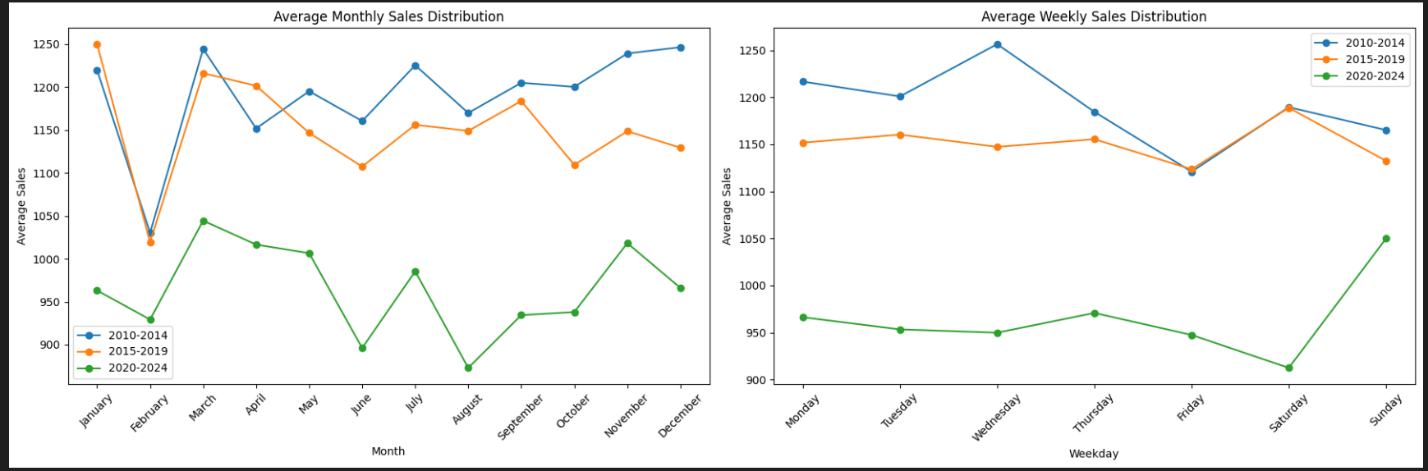
Data visualization was used to explore various aspects of the dataset. For instance, we examined the number of orders per month to identify seasonal trends and variations. This analysis highlighted peak periods and trends in customer purchasing behavior. The distribution of products across different categories was visualized to determine the popularity of various product categories, aiding in inventory management and marketing strategy formulation.

We also explored customer distribution by country to assess market reach and identify key regions. This geographic analysis provided valuable information on where the retailer has a strong customer base and where there might be opportunities for expansion. Additionally, the relationship between unit price and unit cost was examined to understand pricing strategies and cost management.

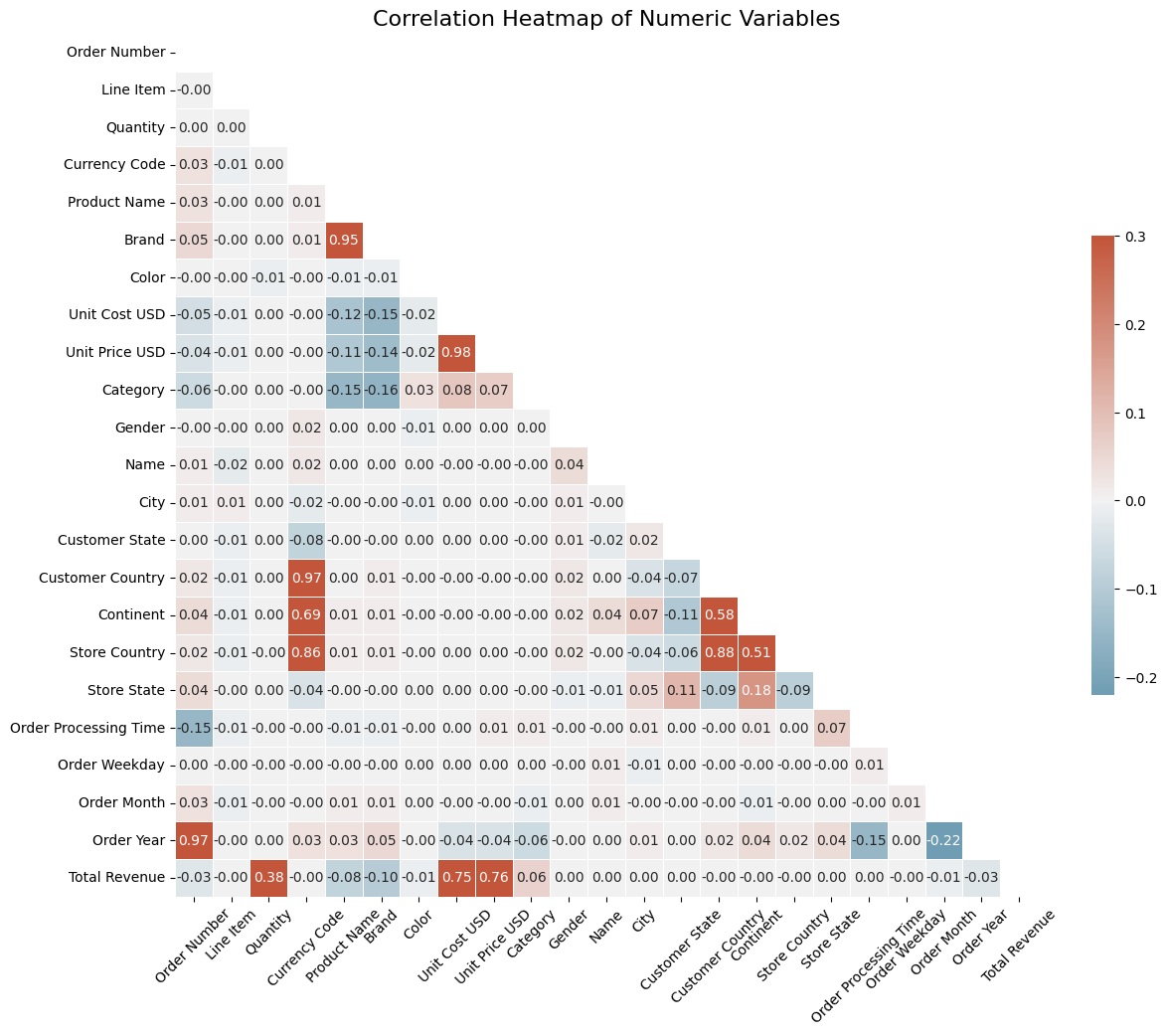
Sales trends were analyzed both monthly and weekly to observe patterns and anomalies. This analysis helped identify recurring trends and potential outliers in sales performance, which can inform future sales strategies and operational adjustments.



* The scatter plot shows a positive linear relationship between Unit Price and Unit Cost. As the cost increases, the price tends to increase as well, which is expected.
* There is some variance in the color coding (possibly indicating another variable such as product type or profitability), but the overall trend is consistent across different colors.



* **Monthly Sales**:
* Sales have declined over time, with the highest sales observed in the period 2010-2014, followed by 2015-2019, and the lowest in 2020-2024. This decline could be due to various factors, including market saturation, increased competition, or economic downturns.
* January shows a significant spike in the earlier years (2010-2014) but decreases over time, while February consistently has lower sales, possibly due to post-holiday season effects.
* **Weekly Sales**:
* Similar to monthly sales, weekly sales have declined over the years.
* Saturday generally shows a peak in sales across all years, which is typical as it falls during the weekend. However, sales on Sunday have significantly increased in the 2020-2024 period compared to earlier years, indicating a possible shift in consumer behavior.
* Sales are lower on weekdays, with a slight dip on Thursday and Friday, suggesting these might be less favorable days for consumers to make purchases.



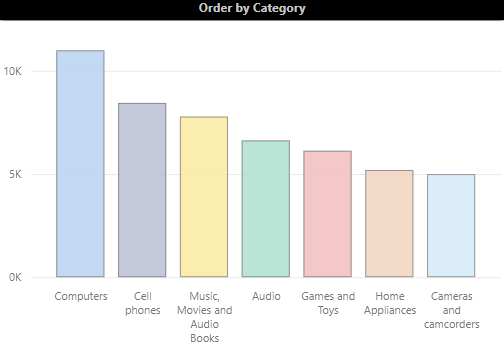
**Correlation Heatmap of Numeric Variables**

* **Strong Correlations**:
  + Product Name and Brand show a very high correlation (0.95), indicating that these two variables are likely linked, possibly because certain brands are consistently associated with specific product names.
  + Unit Price USD and Unit Cost USD have a very strong correlation (0.98), which is expected as the cost often influences the price.
  + Order Year and Order Number have a strong positive correlation (0.97), suggesting that the order number increments with the year.
* **Moderate Correlations**:
  + Customer Country and Store Country have a correlation of 0.88, indicating that customers tend to purchase from stores in their own country.
  + Continent and Customer Country have a moderate correlation (0.69), which might suggest regional purchasing behaviors.
* **Negative Correlations**:
  + Category and Unit Price USD (-0.16) indicate that different categories might have distinct price ranges.
  + Order Year and Category (-0.06) suggest that product categories may have changed over the years.

**Power Bi**

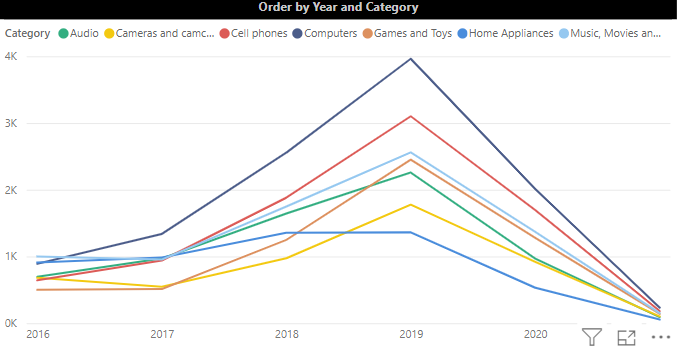
**------------**

**1. Order by Category:**



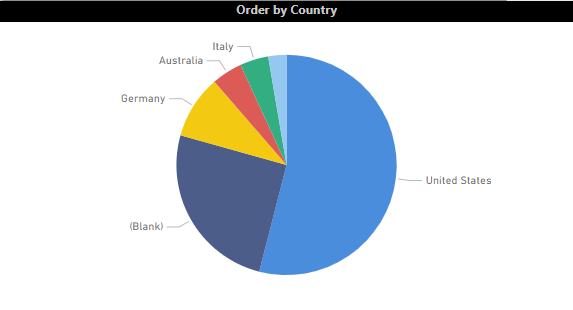
* **Top Categories:** Computers and Cell Phones have the highest order counts, indicating strong demand for these products.
* **Moderate Categories:** Music, Movies, and Audio Books, Audio equipment, and Games and Toys also show substantial orders but are behind Computers and Cell Phones.
* **Lowest Categories:** Cameras and Camcorders, along with Home Appliances, have the lowest order counts among the listed categories.

**2. Order by Year and Category:**



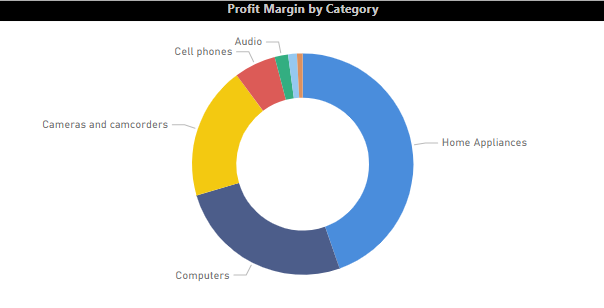
* **Trends Over Time:**
  + There is a noticeable spike in orders across all categories in 2019, followed by a decline in 2020 and 2021.
  + This could be attributed to specific events like product launches, promotions, or broader market trends.
* **Consistent Growth:** Despite the overall trend, categories like Cell Phones and Computers show more consistent growth compared to others.

**3. Order by Country:**



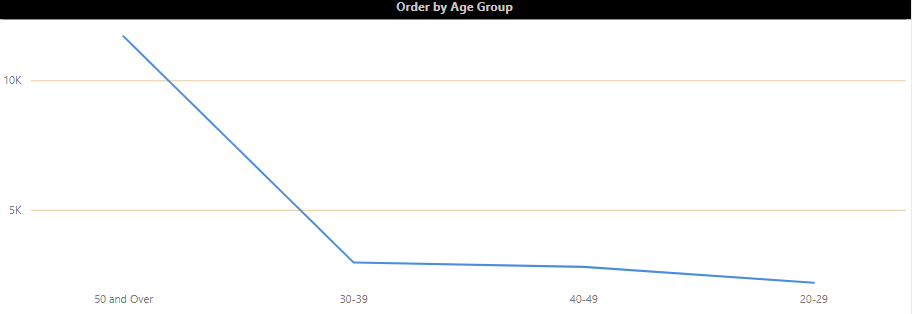
* **Dominant Market:** The United States accounts for the majority of orders, significantly outpacing other countries like Germany, Australia, and Italy.
* **Blank Entries:** The presence of a "Blank" category indicates possible data issues or missing information that might need addressing.

**4. Profit Margin by Category:**



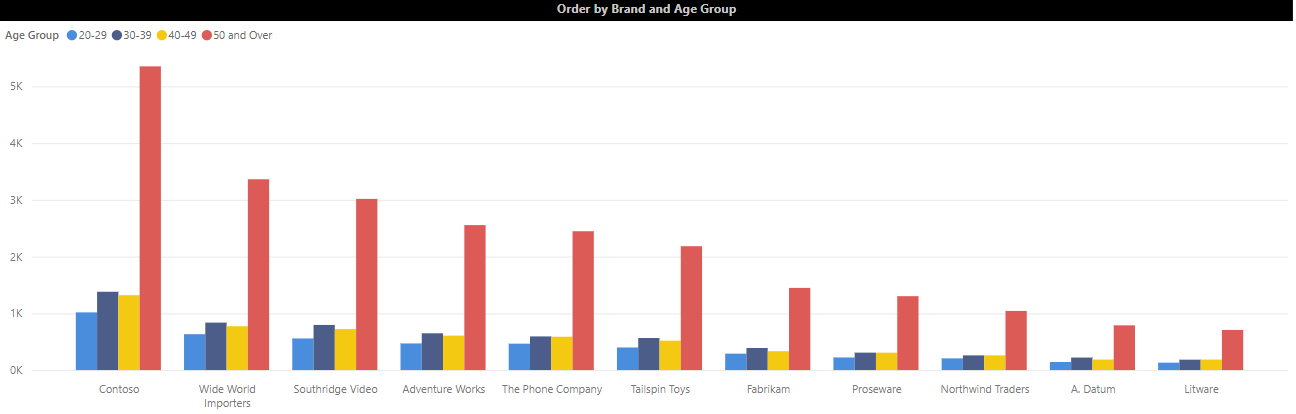
* **Top Profitable Categories:**
  + Home Appliances and Computers have the largest slices, suggesting they offer the highest profit margins.
* **Lower Profitability:** Cameras and Camcorders, along with Audio and Cell Phones, have smaller profit margins, indicating lower profitability despite possibly high sales volumes.

**5. Order by Age Group:**



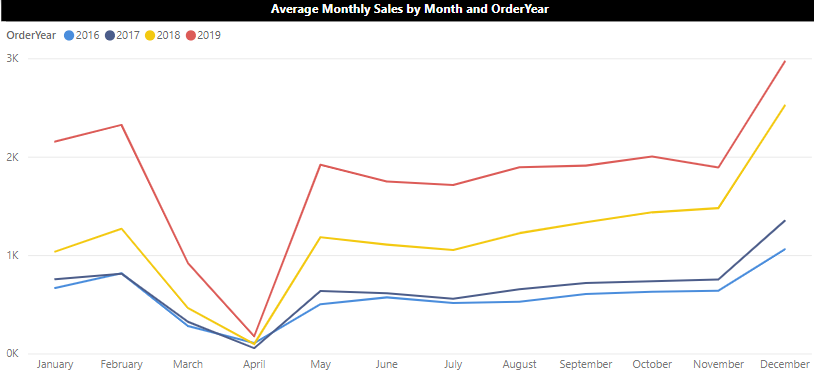
* **Leading Age Group:** Individuals aged 50 and over have the highest order count, significantly more than younger age groups.
* **Trend:** There is a declining trend in orders as age decreases, with the 20-29 age group having the lowest order count.

**6. Order by Brand and Age Group:**



* **Brand Popularity Among Older Age Groups:**
  + Older customers (50 and over) seem to prefer brands like Contoso, Southridge Video, and Adventure Works.
* **Brand Diversification:** Younger age groups (20-29) have a relatively even distribution across various brands, but their order counts are low compared to the older demographic.

**7. Average Monthly and Weekly Sales:**



* **Monthly Sales Pattern:**
  + Sales tend to peak around the beginning (January) and end of the year (November and December), possibly due to holiday shopping seasons.
  + A notable dip is observed around April each year, which might be a post-holiday sales slump.
* **Weekly Sales Pattern:**
  + Sales gradually decrease throughout the week, with the lowest activity on Sunday.
  + This suggests that customers are more likely to place orders during the workweek, particularly at the beginning.

**Predictive Modeling**

In developing predictive models for revenue, several regression techniques were employed. Linear regression served as a baseline model, providing a straightforward approach to understand the relationship between features and revenue. The model's performance was evaluated using metrics such as Root Mean Squared Error (RMSE) and R-squared (R²), which measure prediction accuracy and goodness of fit.

Elastic Net regression was applied to combine penalties from both Lasso and Ridge regression. This technique helps in feature selection and managing multicollinearity, offering a balanced approach to model complexity. The performance of the Elastic Net model was compared to that of the Linear Regression model to assess improvements in prediction accuracy.

Lasso regression was used to enforce sparsity by penalizing less important features, leading to a simplified model with improved interpretability. Ridge regression, on the other hand, was applied to handle multicollinearity by adding a penalty term, which stabilizes the regression coefficients. Each model's effectiveness was evaluated using RMSE and R² metrics, highlighting the relative strengths and weaknesses of each approach.

* **Linear Regression**

RMSE: 829.83

R²: 0.72

* **ElasticNet Regression**

RMSE: 829.83

R²: 0.72

* **Lasso Regression**

RMSE: 829.80

R²: 0.72

* **Ridge Regression**

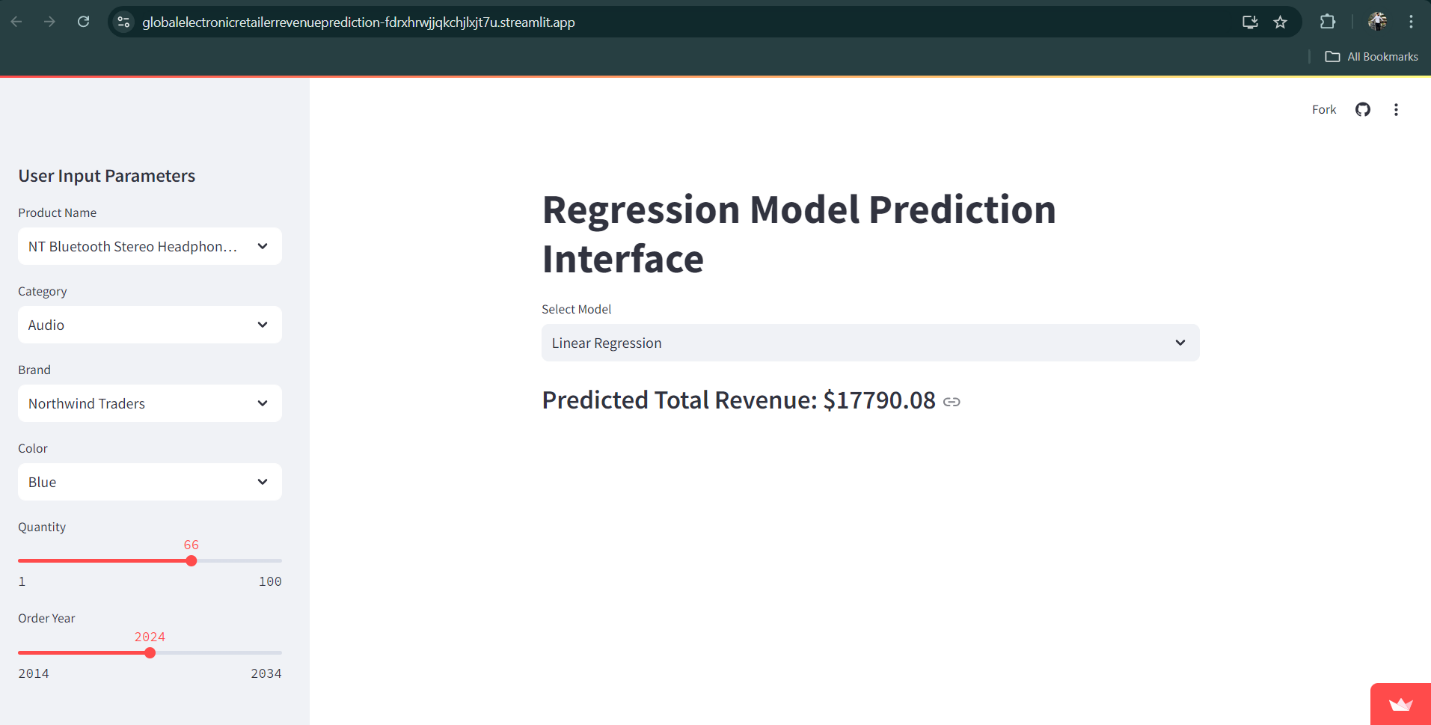
RMSE: 829.83

R²: 0.72

These results indicate that all models performed similarly, with Lasso Regression slightly outperforming others in terms of RMSE, though the difference is minimal.

The comparison of these models provided insights into their relative performance and suitability for revenue prediction. Although all models performed similarly, the Lasso Regression model showed a slightly better predictive accuracy based on RMSE, making it the best-performing model in this analysis. However, the overall fit, represented by R², was consistent across all models.

**Front End: -**

****

**Application Overview:**

* **Saved Models**: We’ve previously trained and saved several regression models. These models are specifically designed to predict outcomes such as sales forecasts or product demand based on various input features, including product details, historical sales data, and other relevant business metrics. These trained models are securely stored and can be easily accessed by the application when needed.
* **app.py File**: We created an app.py file, which acts as the main script for our Streamlit application. This script is the heart of our app, responsible for loading the saved models, setting up the user interface, processing user inputs, filtering data in real-time, and generating predictions using the selected models.
* **Streamlit Sidebar Inputs**: In the sidebar of our Streamlit app, we’ve set up input fields that allow users to select or enter specific details such as Product Name, Category, Brand, Color, Quantity, and Order Year. These inputs are crucial because they help filter the data and serve as features for the predictive models, tailoring the predictions to the user’s needs.
* **Real-Time Filtering**: As users make selections in the sidebar, our application dynamically filters the relevant product details from the dataset based on those inputs. This real-time filtering ensures that the data displayed or used for predictions is always relevant to the user’s current selections. The real-time feedback enhances the user experience by making the application responsive and interactive.
* **Model Selection**: We’ve incorporated a feature that allows users to choose from multiple pre-trained regression models. Users can select models based on criteria such as the type of model, its accuracy, or its suitability for the specific prediction task at hand. Once a model is selected, it processes the filtered data and generates predictions. This feature empowers users to experiment with different models and choose the one that best fits their needs.

**Model Deployment:-**

**1. Saved Models and app.py Checked into GitHub**

* Saved Models: We developed predictive models, likely using libraries such as Scikit-learn, TensorFlow, or others, and trained them on relevant data. After training, we saved these models to files, which might be in formats like .pkl for Scikit-learn, .h5 for TensorFlow/Keras, or .pt for PyTorch.
* app.py: This is our Python script that contains the code to load the saved models, preprocess input data, and serve predictions through a web interface. Additionally, it might handle data visualization, user interaction, and displaying results.
* GitHub Repository: We stored both the saved models and the app.py file in a GitHub repository named Global\_Electronic\_Retailer\_Revenue\_Prediction. This allows us to version control our code and models, collaborate with others, and easily deploy the application.

Repository Link: [jobinajoy/Global\_Electronic\_Retailer\_Revenue\_Prediction](https://github.com/jobinajoy/Global_Electronic_Retailer_Revenue_Prediction)

**2. Connected GitHub Repo to Streamlit App**

* Streamlit: Streamlit is an open-source app framework that enables us to create and deploy machine learning and data science web apps with minimal effort. It allows us to write an interactive web app using pure Python.
* Connecting GitHub to Streamlit: We connected our GitHub repository to a Streamlit deployment. This integration ensures that any updates we push to the repository are reflected in the deployed Streamlit app. Streamlit directly accesses the code and models from the GitHub repository to serve our app.

**3. Deploying the Model in Streamlit App**

* Deployment: We deployed our model live, making it accessible via a web interface. When users visit the deployed URL, they can interact with the Streamlit app, which uses our app.py script to run the model and return predictions or insights based on user inputs.
* App Functionality: Our app may include features such as accepting user input for model predictions (e.g., entering data about retail sales), visualizing predictions, displaying model metrics, and more.

Deployed App URL: [Global Electronic Retailer Revenue Prediction](https://globalelectronicretailerrevenueprediction-fdrxhrwjjqkchjlxjt7u.streamlit.app)

**4. Explanation in Detail**

* Predictive Model: We developed a model to predict the revenue of a global electronic retailer, potentially using historical sales data, economic indicators, marketing spend, etc.
* Model Saving: After training the model, we saved it in a way that allows easy loading and use in production environments.
* App.py: This script is the backend of our Streamlit app, responsible for loading the saved model, processing any input from users, making predictions using the model, and returning the results.
* GitHub: By pushing our code and models to GitHub, we ensure that our work is backed up, shareable, and deployable.
* Streamlit Deployment: Hosting the app on Streamlit provides an easy-to-use web interface, allowing anyone to interact with our predictive model without needing to run the code locally.

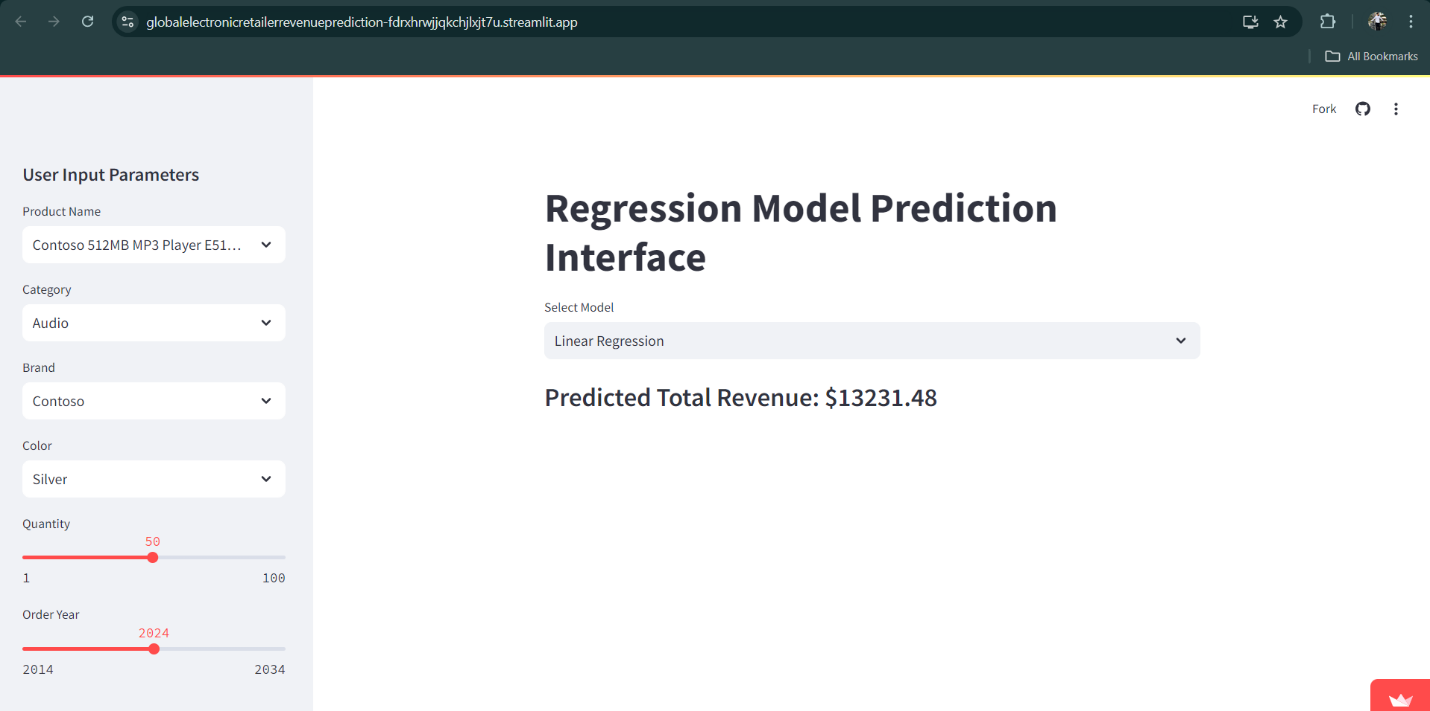
**Testing**

**----------**

**Case 1:**  
Model - Linear Regression   
User Input Parameters:  
Product Name - Contoso 512MB MP3 Player E51 Silver

Category – Audio  
Brand - Contoso

Color – Silver  
Quantity – 50  
Order Year - 2024

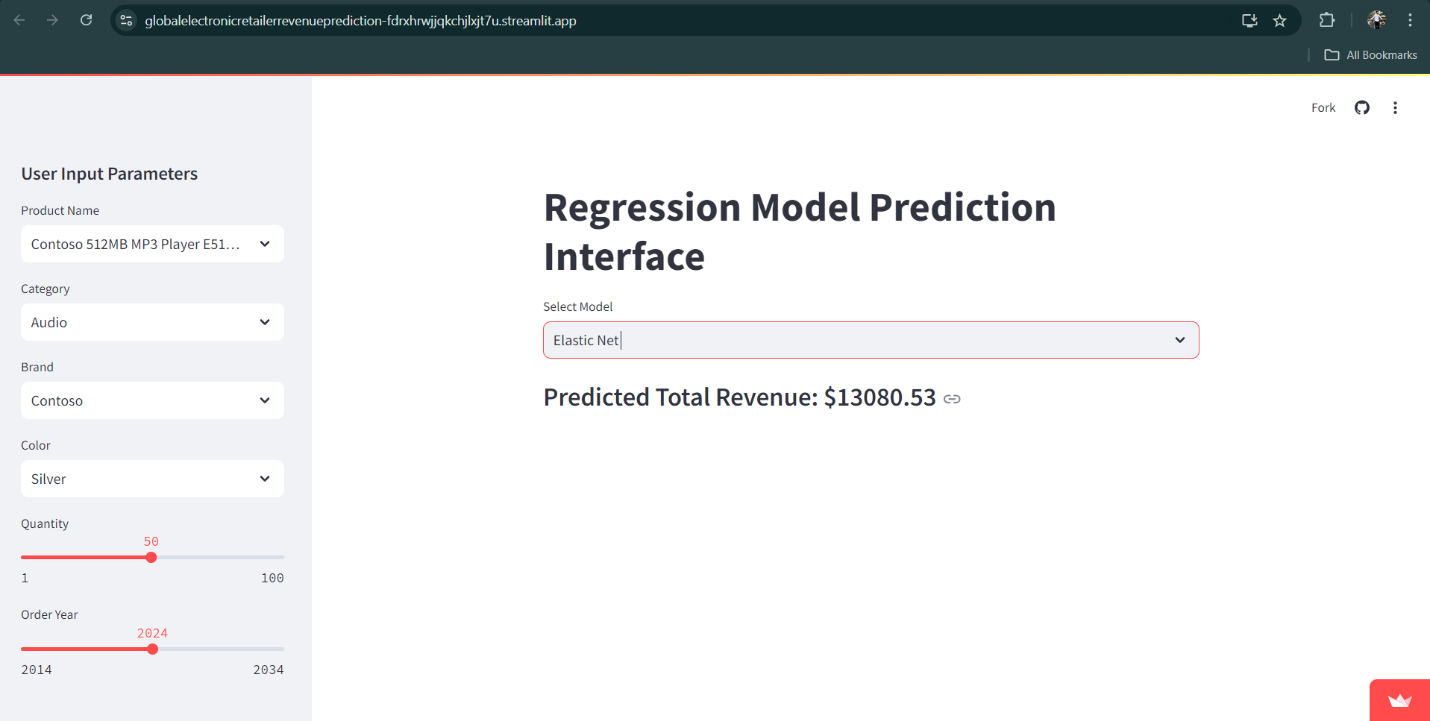
****

Predicted Total Revenue - $13231.48

**Case 2:**  
Model – Elastic Net   
User Input Parameters:  
Product Name - Contoso 512MB MP3 Player E51 Silver

Category – Audio  
Brand - Contoso

Color – Silver  
Quantity – 50  
Order Year - 2024

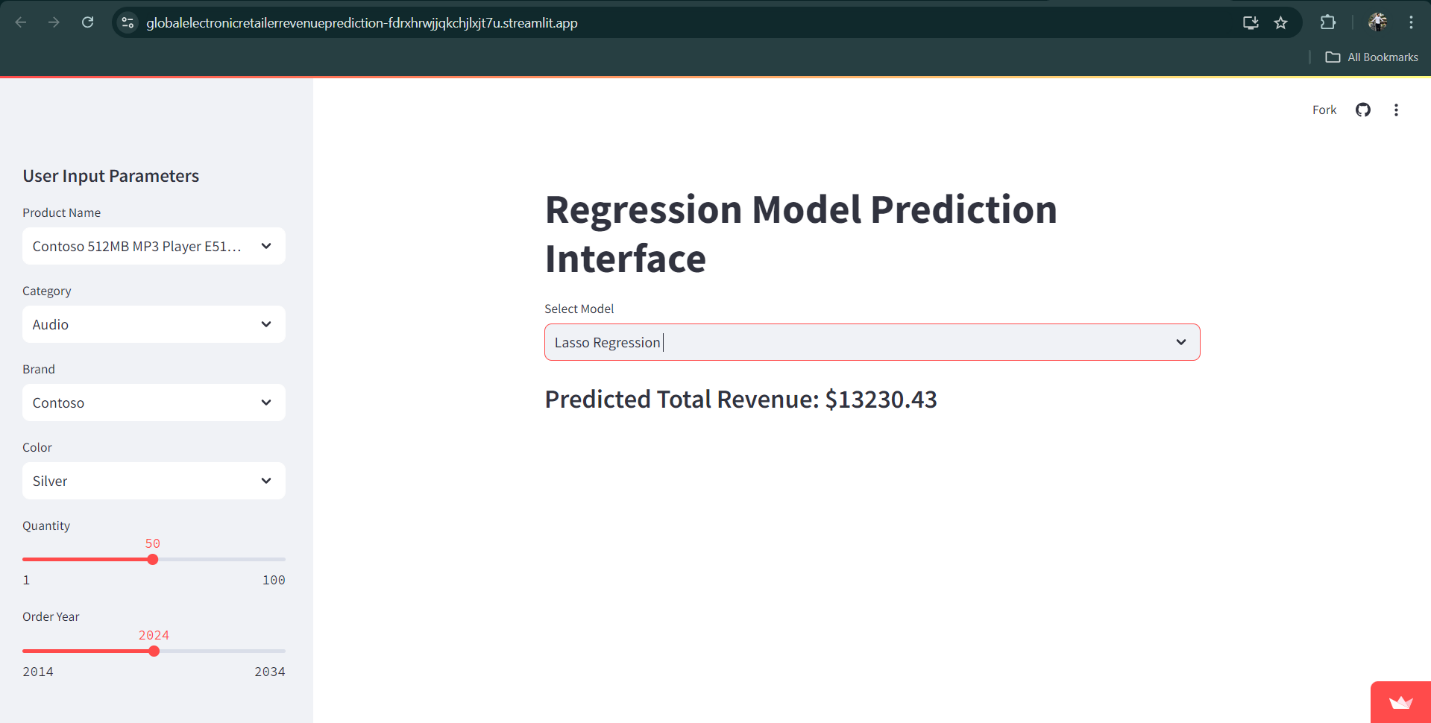


Predicted Total Revenue - $13080.53

**Case 3:**  
Model – Lasso Regression  
User Input Parameters:  
Product Name - Contoso 512MB MP3 Player E51 Silver

Category – Audio  
Brand - Contoso

Color – Silver  
Quantity – 50  
Order Year - 2024

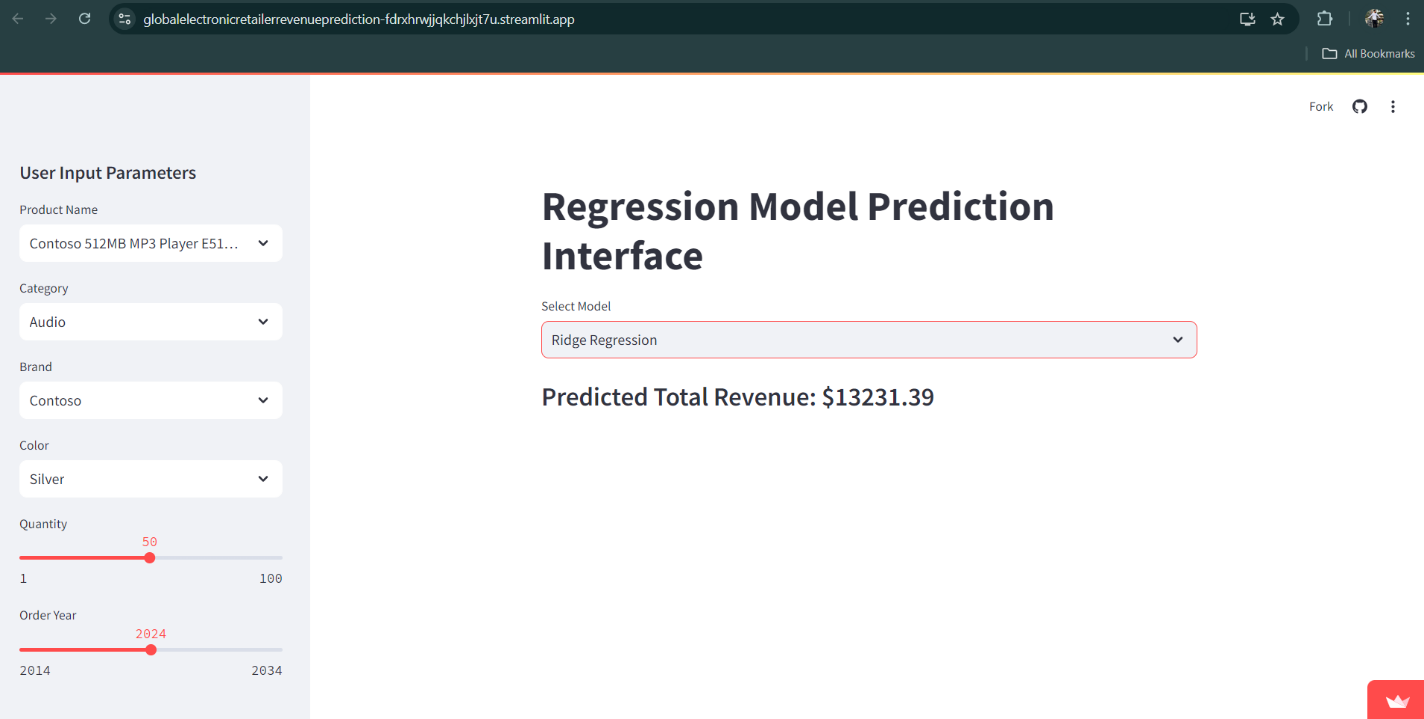
****

Predicted Total Revenue - $13230.43

**Case 4:**  
Model – Ridge Regression  
User Input Parameters:  
Product Name - Contoso 512MB MP3 Player E51 Silver

Category – Audio  
Brand - Contoso

Color – Silver  
Quantity – 50  
Order Year - 2024

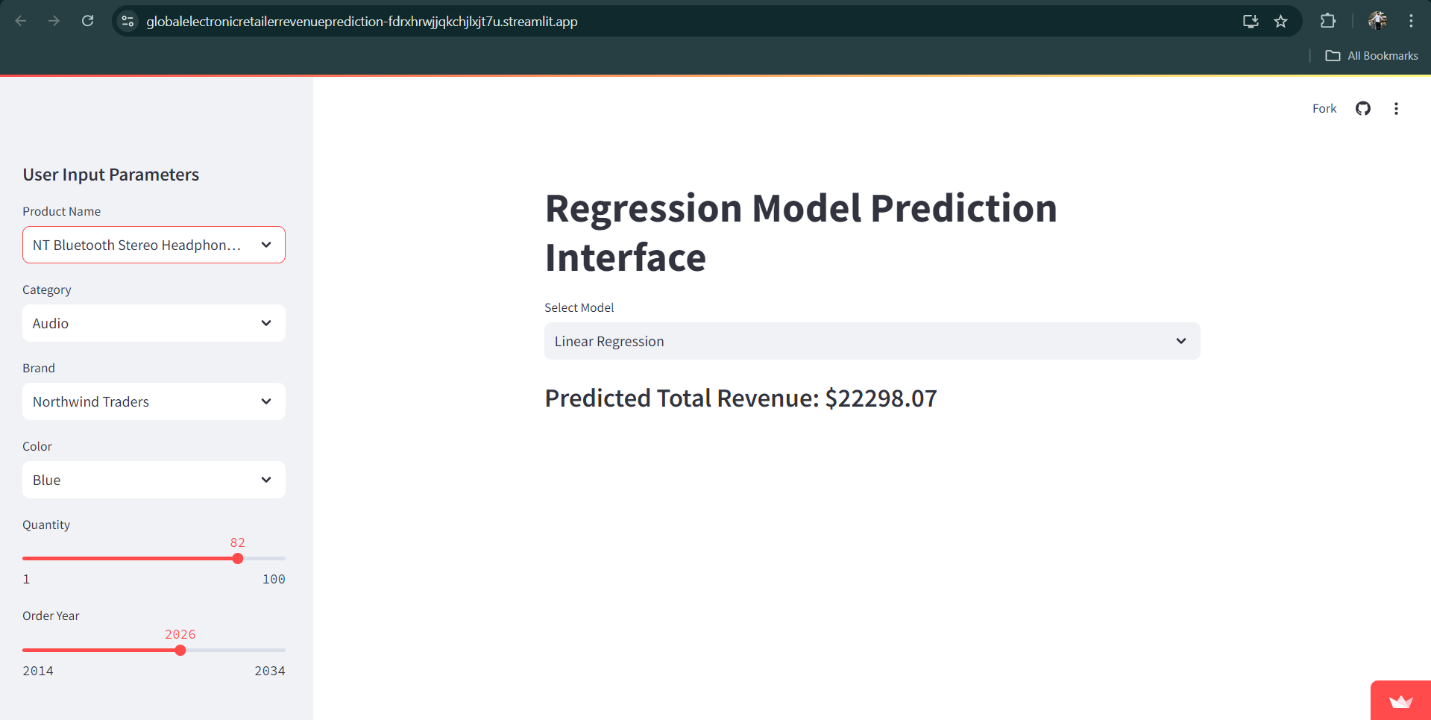
****

Predicted Total Revenue - $13231.39

**Case 5:**  
Model – Linear Regression  
User Input Parameters:  
Product Name – NT Bluetooth Stereo Headphones E52 Blue

Category – Audio  
Brand – Northwind Traders

Color – Blue  
Quantity – 82  
Order Year – 2026



Predicted Total Revenue - $22298.07

**Conclusion**

The analysis successfully integrated and prepared the dataset, addressing missing values, outliers, and feature engineering. Various regression models, including Linear Regression, Elastic Net, Lasso, and Ridge, were developed and evaluated to predict revenue. Each model provided valuable insights into the factors influencing revenue and demonstrated different strengths in terms of predictive power and interpretability.

Based on the evaluation, recommendations include further refinement of model hyperparameters and the exploration of additional features to enhance accuracy. Deploying the best-performing model in real-time systems for revenue forecasting could significantly benefit the retailer's strategic planning and operational efficiency.