# **CAPSTONE PROJECT**

# **ENHANCING SUPPLY CHAIN EFFICIENCY AND CUSTOMER INSIGHTS AT DATACO GLOBAL**

# **Introduction**:

“DataCo Global” conducted an extensive analysis utilizing a comprehensive dataset encompassing various facets of the supply chain, including provisioning, production, sales, and commercial distribution. This dataset uniquely facilitates the correlation between structured and unstructured data, thereby yielding invaluable insights. The dataset comprises products categorized into three distinct classes: clothing, sports items, and electronic supplies.

# **Problem statement**:

The problem statement revolves around utilizing “DataCo Global” dataset of supply chains to address two key areas of improvement:

* **Sales Forecasting & Optimization**: The project involves the development of robust regression models for predicting future sales across Clothing, Sports, and Electronic Supplies. The primary objective is to attain precise sales forecasts, enabling the optimization of inventory management and production planning. This, in turn, contributes to an elevated level of customer satisfaction.
* **Customer Segmentation & Product Recommendation**: The initiative entails the utilization of classification algorithms for the purpose of customer segmentation, leveraging behavioral and purchasing patterns derived from the existing dataset. This undertaking aims to establish a sophisticated product recommendation system, harnessing clickstream data to furnish individualized product suggestions to each customer. The objective is to improve customer experiences, increase retention, and drive overall business growth.

# **Overview of the dataset:**

The attainment of the specified problem statements hinges upon the utilization of two discrete datasets. The first dataset has been leveraged to address ***sales forecasting and customer segmentation*** objectives. It encompasses pertinent details encompassing sales, inventory, production, and customer conduct within the realms of Clothing, Sports, and Electronic Supplies. This dataset forms the foundation for the development of sales forecasting regression models and customer segmentation classification algorithms, achieved through meticulous analysis.

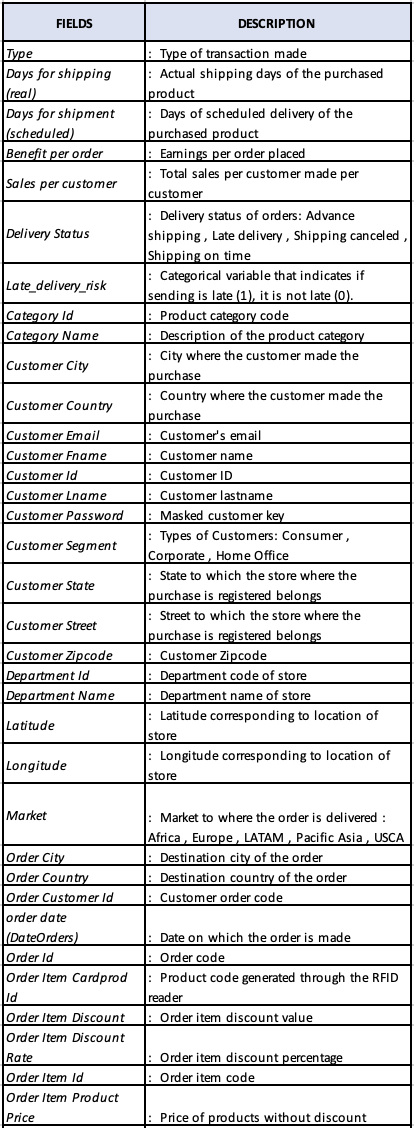




Figure : Description of the dataset (The dataset that use to build models and customer segmentation)

On the other hand, to construct an effective ***product recommendation system***, the second dataset has been employed. This dataset comprises clickstream data, capturing essential customer interactions, preferences, and purchase histories. By leveraging this dataset, the product recommendation system is established, enabling personalized product suggestions to customers based on their past behavior and interests.

A list of items with text

Description automatically generated

Figure : Description of the dataset (The dataset that use to build product recommendation system)

# **Data preparation:**

1. **Dataset with product price and sales details:**

Table : Data preprocessing stage (Dataset 01)

|  |  |  |
| --- | --- | --- |
| Number of data records | 180519 | |
| Number of missing values | ***Variable*** | ***No of missing values*** |
| Customer Lname | 8 |
| Customer Zip code | 3 |
| Order Zip code | 155679 |
| Product Description | 180519 |
| Number of duplicates | 0 | |
| Number of unwanted columns | 30 | |
| Number of data in the training data | 144415 | |
| Number of data in the test data | 36104 | |

* The dataset consists of 180519 rows and 53 columns, focusing on various attributes of the product selling. To maintain data integrity and accuracy, 180519 rows with missing values in the " Product Description " column, 155679 rows with missing values in the " Order Zip code " column, 3 rows with missing values in the " Customer Zip code " column, and 8 rows with missing values in the " Customer Lname " column, were removed.
* Duplicate records were checked, and fortunately, no duplicates were found. Different ID columns, which serves as a unique identifier, was dropped from the dataset since it doesn't contribute to the analysis.
* To prepare the dataset for modeling, label encoding was applied to the categorical columns. This process transformed categorical variables into numerical representations, enabling effective use in machine learning algorithms.
* For model evaluation and generalization testing, the dataset was split into two sets: the training set (80% of the data) and the test set (20% of the data). This division is essential to assess the model's performance on unseen data and ensure its ability to generalize well.

1. **Dataset that uses to develop the product recommendation system:**

Table : Data preprocessing stage (Dataset 02)

|  |  |
| --- | --- |
| Number of data records | 469977 |
| Number of missing values | 0 |
| Number of duplicates | 469901 |
| Number of unwanted columns | 6 |

This dataset has been used to develop a product recommendation system. Therefore, all the duplicate values in the dataset were removed. Before removing the duplicates, the text data was cleaned for the “product” column by lowercasing, removing special characters, stop word removal, lemmatization, and tokenization.

# **Descriptive Analysis:**

**Product price and sales dataset**

A graph showing a bar graph

Description automatically generated with medium confidence

Figure : Total sales for all markets

Europe takes the lead in total sales, representing a thriving market with strong economic activity and diverse consumer preferences. The region's well-developed infrastructure, advanced technology, and stable economies contribute to its prominence in the global sales landscape. LATAM comes in second, indicating a substantial market presence in the Latin American countries. The region's dynamic culture, growing middle class, and increasing urbanization have fostered a burgeoning consumer market, attracting businesses to invest in this vibrant and promising market. Following LATAM, Pacific Asia emerges as the third-largest sales region. This expansive area includes countries like China, Japan, South Korea, and others, boasting a massive population and tremendous economic growth. As a result, businesses have capitalized on the region's immense potential and consumer demand, resulting in significant sales figures. Next in line is USCA, comprising the United States and Canada. These two North American economic powerhouses demonstrate consistent and robust consumer spending patterns. The region's affluence, high standard of living, and substantial disposable incomes has enabled businesses to thrive and achieve considerable sales figures. Finally, Africa completes the list, encompassing a vast and diverse continent with numerous emerging markets. Although characterized by various challenges, such as infrastructural limitations and political instability in some regions, Africa holds tremendous untapped potential for growth. The continent's youthful population and increasing urbanization make it an attractive destination for businesses seeking new opportunities.

A graph of different colored bars

Description automatically generated with medium confidence

Figure : Total sales for all regions

In terms of total sales across different regions, the rankings are as follows: Western Europe takes the lead, representing the most lucrative market in the region. Its high levels of economic development, strong consumer purchasing power, and stable business environments make it a preferred destination for companies seeking expansion and growth. Central America follows closely behind, featuring a region with a mix of economies at different stages of development. Despite its smaller size compared to other regions, Central America offers diverse consumer markets and increasing opportunities, attracting businesses looking to establish a presence in this vibrant and geographically strategic area. South America ranks third in total sales for all regions. This vast and diverse continent offers a mix of rapidly developing economies, along with established markets. The region's rich natural resources, growing middle class, and increasing urbanization contribute to its importance in the global sales landscape.

This pattern indicates that, the total sales across markets and regions present a diverse and evolving global sales landscape. Businesses looking to expand and succeed in these markets must consider the unique characteristics and consumer behaviors of each region to make informed strategic decisions and maximize their sales potential.

A screen shot of a graph

Description automatically generated

Figure : Different types of payment methods used in all regions

Debit cards are the most preferred payment method across all regions due to their convenience and security. With their widespread acceptance, they offer a quick and hassle-free way to make transactions, both online and offline. Credit cards and mobile payment options also rank high, providing added benefits like rewards and enhanced safety. In contrast, cash payments are the least favored choice. Handling physical money can be burdensome and poses security risks. As digitization and technology continue to advance, the global trend towards electronic payments strengthens. While preferences may slightly differ by region, the overall trend remains consistent, with debit cards dominating as the preferred payment method.

A graph of a bar graph

Description automatically generated with medium confidence

Figure : Top 10 product with most loss

A graph of a bar chart

Description automatically generated with medium confidence

Figure : Top 10 region with most loss

The products that are generating a negative benefit per order, resulting in a loss of revenue for the company, are Cleats and Men’s Footwear. The largest loss sales are associated with the Cleats category, followed by Men’s Footwear. The regions with the most significant lost sales are Central America and Western Europe. These losses could potentially be attributed to various factors, including suspected frauds or late deliveries. Analyzing the payment methods used in these transactions may help identify patterns related to fraud. By understanding which payment methods are associated with fraudulent activities, the company can take preventive measures to mitigate future fraud risks and safeguard their revenue. Addressing fraud-related issues and optimizing delivery processes in these regions can contribute to reducing the significant loss of approximately 3.9 million, thus enhancing the overall financial performance of the company.

The category of products that is experiencing the most frequent late deliveries is Cleats, followed by Men's Footwear. As indicated in the data, some orders are flagged with a risk of late delivery. By comparing these orders with the actual late delivered products, it becomes evident that the Cleats category has the highest occurrences of late deliveries, with Men's Footwear being the second highest. Ensuring timely delivery is crucial for customer satisfaction in the supply chain industry. When orders are not delivered on time, it can lead to dissatisfied customers, potentially impacting the company's reputation and customer loyalty. Therefore, focusing on improving the delivery processes for the Cleats and Men's Footwear categories can be instrumental in reducing late deliveries and enhancing overall customer satisfaction. Supply chain companies may need to identify and address the underlying reasons behind these delays, such as logistical challenges, inventory management, or operational inefficiencies, to improve their delivery performance in these product categories.

A graph of different colored bars

Description automatically generated with medium confidence

Figure : Top 10 products with most late deliveries

A graph of a graph

Description automatically generated with medium confidence

Figure : Different types of shipping methods

As expected the most number of late deliveries for all regions occured with standard class shipping,with same day shipping being the one with least number of late deliveries.Both the first class and second class shipping have almost equal number of late deliveries.

# **Modeling:**

**Future sales prediction using order item quantity:**

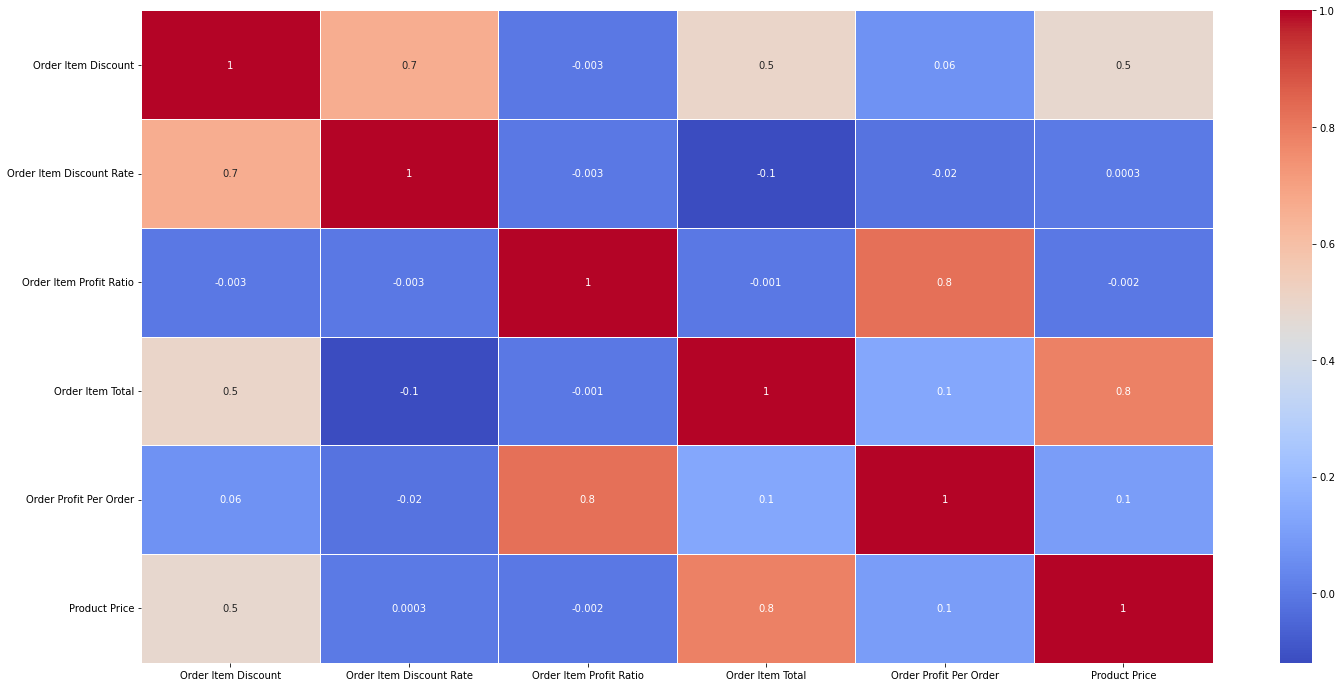
****

Figure : Heatmap

The provided heatmap shows the correlation coefficients between various features in the dataset, with a focus on the significant multicollinear values. Multicollinearity occurs when two or more variables are highly correlated, which can lead to inflated coefficient estimates and affect the accuracy of statistical models. From the heatmap, we observe some significant multicollinear relationships. For instance, there are strong positive correlation of 0.8 between the "Order profit per order" and "Order item profit ratio" variables, “Order item total” and “Product price” variables, indicating that the two are closely related.

Subsequently, three different machine learning models (Decision tree, Random Forest and XGBoost) and two statistical models (Lasso regression and Ridge regression) were employed to make predictions, and the MAE and RMSE for each model were calculated as follows:

**Number of independent variables:** 22 variables

**Response Variable:** “Order Item Quantity"

Table : MAE and RMSE values for each models

|  |  |  |
| --- | --- | --- |
| **Model** | **MAE** | **RMSE** |
| Lasso Regression | 0.3712 | 0.5507 |
| Ridge Regression | 0.3687 | 0.5493 |
| Random forest Regression | 0.0005 | 0.0086 |
| XGBoost Regression | 0.0037 | 0.0137 |
| Decision Tree | 0.00013 | 0.0117 |

In this comprehensive analysis aimed at predicting the "Order Item Quantity," the Ridge and Lasso Regression models were employed instead of Linear Regression due to the presence of multicollinearity among the independent variables, as indicated by the heat map. The independent variables used in the study included 'Type', 'Category Name', 'Customer City', 'Customer Country', 'Customer Segment', 'Customer State', 'Department Name', 'Market', 'Order City', 'Order Country', 'Order Item Discount', 'Order Item Discount Rate', 'Order Item Profit Ratio', 'Order Item Total', 'Order Profit Per Order', 'Order Region', 'Order State', 'Order Status', 'Product Name', 'Product Price', 'Shipping Mode', and 'late\_delivery'.

The Ridge and Lasso Regression models were evaluated based on two performance metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The results indicated that both models exhibited comparable predictive abilities, with the Ridge Regression model achieving an MAE of approximately 0.3687 and an RMSE of around 0.5493. Similarly, the Lasso Regression model produced an MAE of about 0.3712 and an RMSE of approximately 0.5507. These results suggest that both regularization techniques effectively addressed the multicollinearity issue, resulting in models with similar predictive performance.

Furthermore, the analysis also included non-linear regression models such as Random Forest Regression and XGBoost Regression. These models achieved notably lower MAE values of 0.00049 and 0.0036, and RMSE values of 0.0084 and 0.0131, respectively. The Random Forest and XGBoost models' superior performance can be attributed to their ability to handle non-linear relationships among the independent and response variables.

Surprisingly, the Decision Tree model demonstrated exceptional performance, yielding an impressively low MAE of 0.00024 and a relatively low RMSE of 0.015. The Decision Tree model's effectiveness in capturing underlying patterns and interactions among the independent variables showcases its potential as a robust predictor for the "Order Item Quantity."

To ensure the models' reliability and generalization capabilities, it is crucial to validate them on unseen data and conduct a feature importance analysis to identify the most influential independent variables in predicting the "Order Item Quantity." In conclusion, the Ridge and Lasso Regression models were employed to mitigate the multicollinearity issue observed in the heat map. The Decision Tree, Random Forest Regression, and XGBoost Regression models also displayed promising performance, outperforming the Ridge and Lasso Regression models.

**Customer segmentation:**

* K-Means clustering is an effective technique for unsupervised machine learning, particularly useful for customer segmentation. This method divides a dataset into distinct clusters, each representing a group of similar data points. The algorithm iteratively assigns data points to the nearest cluster center, updating the centers based on these assignments until they stabilize. This minimizes the squared distances between data points and their cluster centers.
* In the realm of customer segmentation, K-Means proves invaluable. It allows businesses to categorize customers into groups based on purchasing behavior, preferences, or relevant features, enabling tailored marketing strategies and improved customer service.
* To determine the optimal cluster count for K-Means, the "elbow method" is often employed. This entails plotting the sum of squared distances (inertia) for varying cluster numbers and observing where the plot's decline slows, resembling an "elbow." This point signifies the optimal cluster count.
* In this scenario, the elbow method indicates the optimal cluster count as 3. Consequently, a K-Means model is fitted using this value. The outcome is three customer segments, each with distinct characteristics:

A graph with blue dots

Description automatically generated

Figure : Elbow Plot

A screenshot of a computer screen

Description automatically generated

Figure : Customer Segmentations

Cluster 1: This segment tends to have more discount rates for orders. But the quantity of ordering is less compared to the other two groups. So, this may be mostly the individual customers.

Cluster 2: This segment tends to purchase more items at once. This group is having the second highest order item totals. But order profit per order is minimum for this group. Hence this group may be the people who are making bulk purchasing at discounts.

Cluster 3: This segment is similar to the second segment group. They also do bulk purchasing. Order item total is high for this group. The highest profit can be obtained by targeting this segment customers.

**Product recommendation system:**

In this context, we harness the power of collaborative filtering to develop an effective product recommendation system. Collaborative filtering relies on user-item interactions and feedback to understand the preferences and interests of individual users. By analyzing historical data, such as user ratings, reviews, and purchase behavior, the system identifies similarities between users with similar tastes and preferences. Leveraging this collective knowledge, the collaborative filtering algorithm then generates personalized recommendations for each user, presenting them with products that align with their unique interests. This approach not only enhances the user experience by tailoring recommendations to individual preferences but also benefits businesses by increasing customer satisfaction, engagement, and ultimately driving revenue through improved sales and customer retention.

Data Preprocessing:

* The product data underwent preprocessing, including text cleaning and normalization. The processed text, labeled 'Processed\_Product', was obtained by removing stop words, special characters, and converting the text to lowercase. A dictionary was used to map the original product names to their cleaned counterparts.

TF-IDF Vectorization:

* The importance of words in product descriptions was quantified using the TF-IDF vectorization technique. The TF-IDF vectorizer transformed the processed product text into numerical vectors, enabling the calculation of similarity between products based on their descriptions.

Cosine Similarity Calculation:

* The TF-IDF vectors allowed to compute the cosine similarity between products. The cosine similarity matrix measured the cosine distance between each pair of products, indicating the level of similarity between them.

Recommendation Algorithm:

* The collaborative filtering approach involved identifying similar products for a given input product. Using the processed text, the system found the index of the product in the data frame and retrieved its similarity scores with other products. The system returned the top 10 similar products based on their cosine similarity scores, excluding the input product itself.

A screenshot of a computer

Description automatically generated

Figure : Product Recommendation using the system

As an illustration, when the product name "Clicgear 8.0 Shoe Brush" was provided as input, the system efficiently identified and presented relevant product recommendations along with their respective similarity scores. The top ten recommended products were listed, each exhibiting a varying level of similarity with the input product. Notably, the "Clicgear Rovic Cooler Bag" received the highest similarity score of 0.2930, indicating a significant resemblance in product attributes and likely appealing to users with an interest in golfing accessories. Following closely were the "Ogio Race Golf Shoes" and the "Nike Men's Free 5.0+ Running Shoe," both exhibiting similarity scores of 0.2089 and 0.2082, respectively.

# **Performance evaluation:**

**Future sales prediction using order item quantity:**

*Selected best model: Random Forest Regressor*

A white screen with a black border

Description automatically generated

Figure : VIP plot for the selected best model 9Random Forest Regressor)

The VIP plot is a visualization technique used to interpret the importance of different features in a predictive model. In this case, the plot was created using a set of features, each assigned an importance score based on their contribution to the model's predictive performance. Analyzing the provided importance scores, it becomes evident that the two most influential features in the model are "Order Item Discount Rate" and "Product Price," with importance scores of 0.4438 and 0.5369, respectively. These two features have a significant impact on the model's ability to make accurate predictions. Additional features, such as "Department Name" and "Product Name," also demonstrate non-negligible importance with scores of 0.0068 and 0.0101, respectively. While their influence is not as pronounced as the top two features, they still contribute meaningfully to the model's predictive power.

Conversely, several other features, including "Type," "Customer City," "Customer Country," "Customer Segment," "Customer State," "Market," "Order City," "Order Country," "Order Item Total," "Order Profit Per Order," "Order Region," "Order State," "Order Status," "Shipping Mode," and "late\_delivery," exhibit relatively lower importance scores ranging from approximately 1e-07 to 1e-04. These features are likely to have minimal impact on the model's predictions or may be less relevant in this specific context.

# **Insights and conclusion:**

In this analysis, regression models were employed to predict order item quantity, with the best-performing models being Random Forest and Decision Trees. These models exhibited low Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values, indicating their efficacy in estimating order quantities accurately. Additionally, customer segmentation was achieved through KMeans clustering, identifying an optimal cluster count of 3. This segmentation approach enables businesses to tailor strategies to distinct customer groups based on purchasing behaviors and preferences.

Furthermore, to enhance customer experience and engagement, a collaborative filtering-based product recommendation system was implemented. By leveraging the preferences and behaviors of similar customers, this system provides personalized product suggestions, contributing to increased sales and customer satisfaction.