**Enhancing Business Insights Through Market Basket Analysis and Sales Prediction**

**Charlene Wolthers**

Department of Data Science and Analytics, National University

ANA 630: Advanced Analytics Applications

Professor Ying Lin

August 4, 2024

**Abstract**

In the contemporary business environment, information has become a primary asset. The success of a business in a competitive landscape depends heavily on its ability to acquire, store, and utilize information effectively. Data analysis, therefore, emerges as a critical activity for acquiring new and useful insights. This research paper focuses on a dataset obtained from Kaggle, which includes sales transaction data from a UK-based e-commerce shop. The paper aims to identify products that are frequently bought together and predict the total sales for these product bundles. Techniques such as clustering, regression analysis, and data visualization are employed to achieve these objectives.

**Introduction**

The ability to analyze large datasets and extract valuable insights is a crucial competency for businesses today. As the digital revolution continues to advance, companies are inundated with data from a variety of sources. Among these, e-commerce has emerged as a significant driver of data generation. E-commerce platforms capture vast amounts of transaction data, encompassing customer preferences, purchasing behavior, and sales trends. Effectively analyzing this data can provide businesses with a competitive edge, enabling them to optimize operations, enhance customer experience, and drive revenue growth.

Market basket analysis and sales prediction are powerful analytical techniques that can be applied to e-commerce data to uncover hidden patterns and predict future sales. Market basket analysis involves examining customer transactions to identify products that are frequently purchased together. This information can be used to design effective marketing strategies, optimize inventory management, and create targeted promotions. Sales prediction, on the other hand, involves using statistical and machine learning models to forecast future sales based on historical data. Accurate sales predictions help businesses plan inventory, allocate resources, and set realistic sales targets.

**Dataset Overview**

This paper demonstrates the application of market basket analysis and sales prediction using a dataset sourced from Kaggle. The dataset includes transaction data from a UK-based e-commerce shop that has been selling gifts and homewares since 2007. The shop is based in London and caters to a global customer base, with both individual customers and small businesses making purchases.

The dataset is substantial, containing 536,350 rows and 8 columns. Each row represents a single transaction, and the columns provide various details about the transaction, including:

* **TransactionNo**: A six-digit unique number that defines each transaction. Transactions with a "C" in the code indicate cancellations.
* **Date**: The date when the transaction was generated.
* **ProductNo**: A five or six-digit unique character used to identify a specific product.
* **Product**: The name of the product.
* **Price**: The price of each product per unit in pound sterling (£).
* **Quantity**: The quantity of each product per transaction. Negative values indicate cancelled transactions.
* **CustomerNo**: A five-digit unique number that defines each customer.
* **Country**: The country where the customer resides.

**Data Cleaning and Preprocessing**

In this process, rows with missing values were removed to maintain data integrity. Negative quantities, indicative of cancelled transactions, were also eliminated. Additionally, the 'Country' and 'Date' columns were dropped as they were deemed irrelevant for the clustering and regression analyses. Lastly, the 'Price' column values were rounded to whole numbers and converted to integers. The cleaned data set contained 527,764 rows and 6 columns.

**Exploratory Data Analysis**

Exploratory Data Analysis (EDA) was undertaken to understand the distribution and relationships within the dataset comprehensively.

One of the primary visualizations used was a bar chart depicting the top 15 products by quantity sold. The bar chart revealed which products were leading in sales volume, providing clear quantitative differences between products, and showing how much more popular certain items were compared to others. For instance, products like "Popcorn Holder" and "World War 2 Gliders Asstd Designs" emerged as top sellers. Another critical visualization was a bar chart showing the prices of the top 15 products. This chart displayed a price range from £5 to £661, offering insights into the pricing strategy and product value distribution.

**Correlation Matrix Heatmap**

To understand the relationships between different variables in the dataset, a correlation matrix heatmap was utilized. This heatmap visualized the correlation coefficients between variables such as 'Quantity', 'Price', 'ProductNo', and 'TransactionNo'. The heatmap showed weak correlations among the variables, indicating that they are largely independent of each other. For instance, there was a very weak negative correlation between 'Quantity' and 'Price' (-0.01), suggesting that price changes do not significantly affect the quantity sold. The independence of 'ProductNo' and 'TransactionNo' from 'Quantity' and 'Price' underscores that product and transaction identifiers do not influence the sales volume or pricing.

**Enhanced EDA Methods and Visualizations**

To deepen the exploratory analysis, several additional methods and visualizations were used.

**Histograms**

Histograms for variables such as 'CustomerNo' and 'ProductNo' were created to understand their distributions.

**CustomerNo Distribution**

The histogram for 'CustomerNo' revealed the frequency of transactions per customer. The distribution displayed a broad range of transaction frequencies among customers. A large number of customers made relatively few purchases, as indicated by the high frequency of transactions clustered towards the lower end of the histogram. However, there were notable peaks within the histogram, representing high-frequency customers who make frequent purchases. These high-frequency customers, although fewer in number, contribute disproportionately to the overall sales volume.

**ProductNo Distribution**

Similarly, the histogram for 'ProductNo' highlighted the distribution of sales across different products, providing insights into which products had the most transactions. The distribution of 'ProductNo' displayed significant variability in the frequency of transactions for different products.

The histogram revealed that a small subset of products were sold far more frequently than others, creating distinct peaks in the distribution. These high-frequency products are the best-sellers and are critical for driving the majority of sales. Conversely, the histogram also showed a long tail of products with lower transaction frequencies.

**Box Plots**

Box plots were used to summarize the distribution of 'Quantity' and 'Price'. These visualizations showed the median, quartiles, and potential outliers.

**Quantity Distribution**

The box plot for the 'Quantity' variable offered a detailed view of the range and central tendency of quantities sold per transaction. In the 'Quantity' box plot, the lower quartile (Q1) and upper quartile (Q3) values are relatively close to the median, indicating that most transactions involve a small number of items. However, the presence of significant outliers, represented by points beyond the whiskers, indicates occasional bulk purchases. These outliers show transactions where customers purchased much larger quantities than usual, suggesting special orders, bulk buying by resellers, or promotional sales.

**Price Distribution**

The box plot for the 'Price' variable provided insights into the distribution of product prices. In the 'Price' box plot, the median price is relatively low, with the bulk of product prices clustered closely around this median. However, the plot also reveals several outliers, which are prices that lie significantly outside the normal range. These outliers indicate higher-priced items that stand apart from the general pricing trend. The presence of these high-priced outliers highlights the diversity in the product catalog, with some premium items priced much higher than the average.

**Clustering Analysis**

To identify product bundles within the dataset, a K-means clustering algorithm was employed. This algorithm helps in partitioning the data into distinct clusters, where each cluster represents a group of transactions with similar characteristics. The elbow method involves plotting the sum of squared distances from each point to its assigned cluster center and identifying the point where the rate of decrease sharply slows down, forming an "elbow" shape in the graph. This point typically suggests the most appropriate number of clusters. In this study, the elbow method indicated that the optimal number of clusters was three. After determining the number of clusters, the K-means algorithm was applied to the dataset, and the transactions were grouped into three clusters.

Cluster 0 emerged as the dominant cluster, containing the majority of the transactions. This indicated that a significant portion of overall sales was concentrated within this cluster. The high volume of transactions in Cluster 0 suggests that the products and purchasing patterns represented by this cluster are highly prevalent among the customer base.

In contrast, Cluster 1 had the fewest transactions, representing a smaller segment of the customer base. This cluster's lower transaction volume suggests that the products or purchasing behaviors it represents are less common.

Cluster 2 fell between the other two clusters in terms of transaction volume, with a moderate number of transactions. This cluster's size suggests that it encompasses a significant portion of the sales but not as much as Cluster 0. The products and purchasing patterns in Cluster 2 are moderately prevalent, indicating a substantial customer segment that is important but not dominant.

**Predictive Analysis**

In the predictive analysis phase, a linear regression model was developed with the objective of predicting total sales for each identified product bundle, or cluster. This approach leverages the relationships between independent variables (such as quantity, price, and cluster designation) and the dependent variable (total sales) to make accurate predictions.

The effectiveness of the linear regression model was evaluated using several performance metrics. The Mean Squared Error (MSE) was calculated to be 6412.45. This metric measures the average squared difference between the actual and predicted values, providing an indication of the model's prediction accuracy. A lower MSE value is desirable as it reflects fewer errors in the model’s predictions.

The Root Mean Squared Error (RMSE), which is the square root of the MSE, was found to be 80.08. RMSE is particularly useful because it expresses the error in the same units as the dependent variable, making it more interpretable. An RMSE of 80.08 suggests that, on average, the model's predictions deviate from the actual total sales by about 80 units.

In addition to these metrics, the Mean Absolute Error (MAE) was used to assess the model’s performance. The MAE, which measures the average magnitude of the errors in a set of predictions, without considering their direction, was calculated to be 17.04. This indicates that the average error in the predicted total sales is around 17 units, providing another perspective on the model's accuracy and reliability.

The R-squared (R²) value was also computed to evaluate the model's explanatory power. The R² value of 0.954 signifies that approximately 95.4% of the variance in total sales can be explained by the model. This high R² value indicates a very good fit, suggesting that the model effectively captures the underlying patterns in the data. The closer the R² value is to 1, the better the model explains the variability of the dependent variable.

These performance metrics collectively indicate that the linear regression model provides a strong predictive capability for total sales based on the input variables. The low MSE and RMSE values suggest that the model's predictions are close to the actual sales figures, while the low MAE further confirms the model's accuracy in predicting total sales with minimal average error. The high R² value underscores the model's robustness in capturing the relationship between the variables, making it a reliable tool for forecasting sales.

**Discussion**

The analyses conducted throughout this study yielded several valuable insights that can significantly inform business strategy and decision-making. A key finding from the clustering analysis is that products in Cluster 0 are the primary drivers of overall sales. This cluster contains the majority of transactions, indicating that the products and purchasing behaviors it represents are highly prevalent among the customer base. Given their substantial contribution to sales, products within Cluster 0 should be prioritized for inventory management to ensure consistent availability and prevent stockouts. Additionally, marketing efforts should focus heavily on these products to maintain and enhance their market dominance. By aligning marketing campaigns with the preferences and behaviors of customers in Cluster 0, businesses can maximize sales and customer satisfaction.

Clusters 1 and 2, despite having fewer transactions compared to Cluster 0, still present significant opportunities for growth. These clusters represent smaller segments of the customer base but highlight distinct purchasing patterns that can be leveraged through targeted marketing and promotions. By analyzing the specific characteristics and preferences of customers in Clusters 1 and 2, businesses can design customized marketing strategies to engage these segments more effectively. Promotions, discounts, and tailored product recommendations could stimulate increased purchasing activity within these clusters, thereby expanding their contribution to overall sales. Understanding the unique needs of these customer segments can help in creating a more personalized shopping experience, which is likely to foster customer loyalty and drive repeat purchases.

The correlation analysis revealed weak correlations among the variables 'Quantity', 'Price', 'ProductNo', and 'TransactionNo'. This suggests that these variables are largely independent, implying that changes in one variable do not strongly predict changes in another. For example, the weak correlation between price and quantity indicates that price adjustments may not significantly impact the number of units sold. This independence suggests that other factors might play a more crucial role in influencing sales. Potential factors could include customer demographics, such as age, gender, and income level, which often affect purchasing behavior. Seasonal trends and external events, like holidays and economic conditions, might also significantly impact sales patterns. For instance, certain products may see spikes in demand during specific seasons or events, regardless of their price or other attributes.

By incorporating customer demographic data and analyzing seasonal trends, businesses can develop a more comprehensive understanding of what drives sales. This holistic view enables more accurate forecasting and more effective strategic planning.

**Conclusion**

The use of clustering analysis in this study has effectively identified distinct product bundles and customer segments. By categorizing transactions into clusters based on purchasing behavior, businesses can gain a nuanced understanding of their sales dynamics. For instance, the identification of Cluster 0 as the dominant cluster provides a clear directive for prioritizing inventory management and marketing efforts towards the products within this cluster. This ensures that high-demand items are always in stock, thereby preventing potential lost sales due to stockouts. Additionally, targeted marketing strategies can be developed to further boost the sales of these popular products, ensuring that they continue to drive a significant portion of the overall revenue.

Similarly, regression analysis has played a crucial role in predicting total sales for each product bundle. The model's ability to explain approximately 95.4% of the variance in total sales highlights its effectiveness in capturing the underlying sales patterns. With metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) indicating a good fit, the regression model provides reliable forecasts that businesses can use to plan their inventory and resources. This predictive capability is vital for ensuring that inventory levels align with expected demand, thereby optimizing stock levels and reducing holding costs.

Moreover, the insights gained from the correlation analysis underscore the importance of considering a wide range of factors that influence sales. The weak correlations among variables like 'Quantity', 'Price', 'ProductNo', and 'TransactionNo' suggest that these factors alone do not fully explain sales patterns. This realization opens the door for future research to explore additional variables that may have a more significant impact on sales. For instance, incorporating customer demographic data, seasonal trends, and external market conditions could provide a more comprehensive understanding of what drives sales. This broader perspective can lead to more accurate predictive models and more effective business strategies.

Looking ahead, future work could involve the exploration of advanced machine learning models to further enhance predictive accuracy. Techniques such as decision trees, random forests, and neural networks have the potential to capture complex, non-linear relationships within the data that traditional regression models may miss. By leveraging these advanced models, businesses can achieve even greater precision in their sales forecasts, enabling them to stay ahead in a competitive market.

In conclusion, this research underscores the critical role of data analysis in modern business strategy. By leveraging techniques such as clustering and regression analysis, businesses can uncover hidden patterns in their data, predict future sales, and make informed decisions that optimize their operations and enhance their competitive edge.

**References**

Kaggle. (n.d.). Online Retail Dataset. Retrieved from [Kaggle website](https://www.kaggle.com).

Bruce, P., Bruce, A., & Gedeck, P. (2017). Practical Statistics for Data Scientists. O'Reilly Media.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference.