Maximizing Sales with Tailored Recommendations

A Comprehensive Analysis of Data Mining Techniques to E-Commerce Success

Ernest Ramazani

Indiana University-Purdue University of Indianapolis

CSCI 48100: Data Mining

May 5, 2023

In this report, we will focus on exploring the purchasing behavior of customers in online retail using a data warehouse approach. The dataset we will use for analysis is a collection of online retail transactions that were made between 01/12/2010 and 09/12/2011 for a some online retail store. The dataset contains more than 50,000 rows, each representing a unique transaction made by a customer. The data consists of various attributes such as InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. Our analysis will focus on using various data mining techniques such as Association Mining, Classification, Clustering Analysis, Recommendation, Text Mining, Social Network Mining, and Regression to gain insights into customer purchasing patterns, identify trends, and make predictions. The language used to perform all the necessary work will be R.

Association Mining is a data mining technique that aims to discover the relationships and patterns that exist between items in a dataset. In our analysis, we will use Association Mining to identify the products that are frequently purchased together and to gain insights into the buying patterns of customers. Classification is another technique that we will use to predict the customer's behavior based on the attributes of the transaction. This will enable us to identify the most significant attributes that influence customer purchasing behavior. Clustering Analysis is a technique that groups similar items based on the attributes they possess. In our analysis, we will use Clustering Analysis to group customers based on their purchasing behavior and to identify the different segments of customers that exist.

Recommendation is another technique that we will use to provide personalized recommendations to customers based on their purchasing behavior. This will help the online retail store to increase customer loyalty and enhance the customer shopping experience. We will use the user satisfaction feedback (in a binary form) as a technique to analyze the customer reviews and feedback on the products sold by the online retail store. This will enable the online retail store to gain insights into the customer's opinion of their products and to identify areas where they need to improve. Finally, Regression is a technique that we will use to predict the customer's future purchasing behavior based on their past purchasing behavior. This will enable the online retail store to forecast future sales and plan accordingly. In addition to the analysis techniques mentioned earlier, we will also be performing descriptive statistics, data visualization, and data cleaning to ensure the quality of our results. The descriptive statistics will provide us with a summary of the dataset, allowing us to gain a better understanding of the purchase behavior and trends. Data visualization will help us to identify any patterns or outliers that may not be visible through simple numerical analysis. Lastly, data cleaning will help us to eliminate any inconsistencies, such as missing or incorrect data, which could lead to inaccurate results.

Overall, the analysis of this online retail dataset will provide valuable insights into customer behavior, product popularity, and purchasing trends. The findings of this report can be used by the online retail company to optimize their marketing strategies, improve their customer experience, and ultimately increase their revenue. Additionally, the techniques used in this report can be applied to other datasets in various industries to derive meaningful insights and improve business performance.

Related Works

There has been extensive research on exploring customer behavior in the e-commerce industry through various data analysis techniques. One study conducted by Chen et al. (2018) used data mining techniques to analyze customer behavior and purchasing patterns in online retail. They used association rules mining and clustering analysis to identify frequent itemsets and customer segments, respectively. The study found that customers tend to purchase items that are complementary to each other, and there are distinct groups of customers with different purchasing behaviors. Another study by Zhang et al. (2020) used machine learning techniques to predict customer behavior and identify factors that influence purchase decisions in e-commerce. They found that factors such as product quality, pricing, and delivery time have a significant impact on customer behavior.

In addition to exploring customer behavior, there has also been research on developing personalized recommendation systems for e-commerce platforms. One study by Sharma et al. (2019) used collaborative filtering and content-based recommendation techniques to provide personalized product recommendations to customers in an online retail platform. They found that the personalized recommendation system increased customer engagement and sales revenue. Another study by Zhou et al. (2021) used a hybrid recommendation system that combined collaborative filtering and deep learning techniques to improve the accuracy of product recommendations in an e-commerce platform.

These studies provide valuable insights into customer behavior and purchasing patterns in online retail, as well as effective techniques for developing personalized recommendation systems. The findings of these studies can inform the data analysis techniques and approaches used in this project to explore purchasing behavior in online retail.

Approaches

To create a data warehouse project, we would first need to define the business requirements and the key performance indicators (KPIs) that we want to analyze. Then we will apply several data mining techniques to extract valuables insights.

To define business requirements, we need to identify the business questions that we want to answer using the data.

Which products are the most popular?

	Code	Description	TotalQuantity
4679	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	53847
2343	85099B	JUMBO BAG RED RETROSPOT	47363
339	84879	ASSORTED COLOUR BIRD ORNAMENT	36381
3256	22197	POPCORN HOLDER	36334
2903	21212	PACK OF 72 RETROSPOT CAKE CASES	36039
4585	85123A	WHITE HANGING HEART T-LIGHT HOLDER	35025
3321	23084	RABBIT NIGHT LIGHT	30680

2656	22492	MINI PAINT SET VINTAGE	26437
2869	22616	PACK OF 12 LONDON TISSUES	26315
2901	21977	PACK OF 60 PINK PAISLEY CAKE CASES	24753

What is the average revenue per customer?

Average revenue per customer is $1898.46

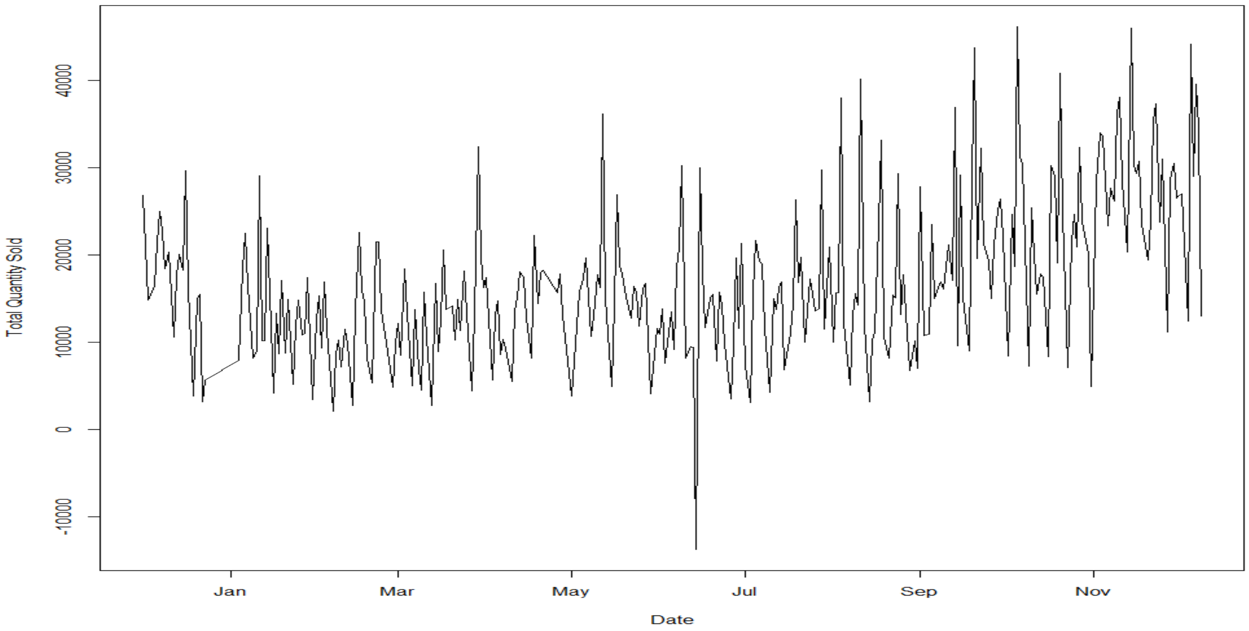
Which countries generate the highest revenue?

	Countries	Revenue
36	United Kingdom	8187806.36
25	Netherlands	284661.54
11	EIRE	263276.82
15	Germany	221698.21
14	France	197403.90
1	Australia	137077.27
34	Switzerland	56385.35
32	Spain	54774.58
4	Belgium	40910.96
33	Sweden	36595.91

Customers that generate the most

	CustomerID	TotalRevenue
1704	14646	279489.02
4234	18102	256438.49
3759	17450	187482.17
1896	14911	132572.62
56	12415	123725.45
1346	14156	113384.14
3802	17511	88125.38
3203	16684	65892.08
1006	13694	62653.10
2193	15311	59419.34

Trend Over time



The average revenue per customer: $1898.46

What is the average unit price and average quantity per country

```
              Country AvgUnitPrice AvgQuantity
1           Australia     3.220612   66.444003
2             Austria     4.243192   12.037406
3             Bahrain     4.556316   13.684211
4             Belgium     3.644335   11.189947
5              Brazil     4.456250   11.125000
6              Canada     6.030331   18.298013
7     Channel Islands     4.932124   12.505277
8              Cyprus     6.302363   10.155949
9      Czech Republic     2.938333   19.733333
10            Denmark     3.256941   21.048843
11               EIRE     5.911077   17.403245
12 European Community     4.820492    8.147541
13            Finland     5.448705   15.346763
14             France     5.028864   12.911067
15            Germany     3.966930   12.369458
16             Greece     4.885548   10.657534
17          Hong Kong    42.505208   16.559028
18            Iceland     2.644011   13.505495
19             Israel     3.633131   14.656566
20              Italy     4.831121    9.961395
```

```
21                Japan     2.276145   70.441341
22              Lebanon     5.387556    8.577778
23            Lithuania     2.841143   18.628571
24                Malta     5.244173    7.433071
25          Netherlands     2.738317   84.406580
26               Norway     6.012026   17.722836
27               Poland     4.170880   10.712610
28             Portugal     8.582976   10.651745
29                  RSA     4.277586    6.068966
30         Saudi Arabia     2.411000    7.500000
31            Singapore   109.645808   22.855895
32                Spain     4.987544   10.589814
33               Sweden     3.910887   77.136364
34          Switzerland     3.403442   15.147353
35 United Arab Emirates     3.380735   14.441176
36       United Kingdom     4.532422    8.605486
37          Unspecified     2.699574    7.399103
38                  USA     2.216426    3.553265
```

The correlation between the quantity of products sold and revenue generated: 0.8866811

Association Mining

#Algortihm

The code is performing association rule mining on the Online Retail dataset. Association rule mining is a technique used to find relationships between different items in a dataset. The output of this code will be a set of rules that indicate which items are frequently bought together by customers. The output will include three measures for each rule: support, confidence, and lift.

Support is the proportion of transactions that contain both items in the rule.

Confidence is the proportion of transactions containing the antecedent that also contain the consequent.

Lift measures the strength of association between the antecedent and consequent. It is calculated by dividing the support of the rule by the product of the support of the antecedent and consequent. A lift value greater than 1 indicates a positive correlation between the items in the rule, while a lift value less than 1 indicates a negative correlation.

The output will also show the most frequent itemsets and their corresponding support values. These itemsets can be used to identify the most popular combinations of items in the dataset.

#Sample Result

```
     lhs        rhs      support    confidence coverage   lift      count
[1]   {23172}  => {23171}  0.01089771 0.9017857  0.01208459 61.909259 202
[2]   {23171}  => {23172}  0.01089771 0.7481481  0.01456625 61.909259 202
[3]   {23172}  => {23170}  0.01062797 0.8794643  0.01208459 49.700457 197
[4]   {23170}  => {23172}  0.01062797 0.6006098  0.01769530 49.700457 197
[5]   {23175}  => {23174}  0.01111351 0.7573529  0.01467415 52.381694 206
[6]   {23174}  => {23175}  0.01111351 0.7686567  0.01445835 52.381694 206
[7]   {23175}  => {23173}  0.01035822 0.7058824  0.01467415 36.961117 192
[8]   {23173}  => {23175}  0.01035822 0.5423729  0.01909797 36.961117 192
[9]   {23174}  => {23173}  0.01057402 0.7313433  0.01445835 38.294291 196
[10]  {23173}  => {23174}  0.01057402 0.5536723  0.01909797 38.294291 196
[11]  {22569}  => {22570}  0.01084376 0.6836735  0.01586103 39.112875 201
[12]  {22570}  => {22569}  0.01084376 0.6203704  0.01747950 39.112875 201
[13]  {22627}  => {22624}  0.01003453 0.6118421  0.01640052 18.292105 186
[14]  {21086}  => {21094}  0.01273198 0.8280702  0.01537549 47.228027 236
[15]  {21094}  => {21086}  0.01273198 0.7261538  0.01753345 47.228027 236
[16]  {21086}  => {21080}  0.01035822 0.6736842  0.01537549 17.059304 192
[17]  {84997C} => {84997D} 0.01230039 0.7402597  0.01661631 32.361921 228
```

Classification

Classification involves building a model that can predict a target variable based on input features. In the context of customer data, classification can be used to identify high-value customers, predict customer churn, or segment customers based on their purchasing behavior.

For example, we could build a classification model that predicts whether a customer will make a purchase in the next month based on their purchase history, demographics, and other relevant features. This model can then be used to target customers with personalized promotions or to identify customers who are at risk of churning. To build a classification model, we typically split the data into a training set and a test set, and use an algorithm (such as logistic regression, decision trees, or neural networks) to learn a model from the training data. We then evaluate the performance of the model on the test data to see how well it generalizes to new data. We can use metrics such as accuracy, precision, recall, and F1-score to evaluate the performance of the model.

#Algorithm

The code performs RFM analysis on the OnlineRetail dataset to segment customers based on their purchasing behavior. The RFM analysis calculates three variables for each customer: Recency (the time since their last purchase), Frequency (the number of purchases made), and Monetary Value (the total amount spent). The data is then standardized and k-means clustering is applied to group the customers into segments. The number of clusters is determined by the k variable, which is set to 3 in the code.

Finally, the cluster assignments are added to the original RFM data frame, and a scatter plot is created to visualize the customer segments based on their Frequency and Monetary Value. Each segment is represented by a different color.

#Output

Please, see attached files for the classification graphical results.

Clustering

Clustering is an unsupervised machine learning technique used to group similar objects together based on their features. In the context of the RFM data project, clustering helps segment customers into groups with similar purchasing behaviors. By understanding these groups, businesses can tailor marketing strategies, customer service, and sales efforts to each specific segment. The code performs clustering analysis on the RFM (Recency, Frequency, Monetary) metrics calculated for each customer. Using a clustering algorithm, such as K-means, customers are assigned to clusters based on their similarities in these metrics. Analyzing the characteristics of each cluster and visualizing the results allows businesses to gain insights into their customer base and make data-driven decisions to improve customer relationship management, marketing, and sales efforts.

#Algorithm

The clustering analysis algorithm for RFM data begins with loading the necessary packages and reading the dataset. After loading the dataset, clean it by removing missing values, specifically in the CustomerID column. Then, calculate the RFM metrics for each customer, which includes Recency, Frequency, and Monetary values. To ensure that each metric contributes equally to the clustering process, scale the RFM data using appropriate scaling methods.

Next, determine the optimal number of clusters using methods such as the Elbow method or the Silhouette method. Apply a clustering algorithm like K-means to the scaled RFM data, and assign each customer to a cluster based on the results. Finally, analyze the characteristics of each cluster, interpret the results, and visualize the clusters using scatter plots or scatter plot matrices. The specific code implementation provided earlier aligns with this general algorithm and can be adapted to various programming languages or tools.

#Output

Please, see attached files for the clustering graphical results

Satisfaction Analysis

Satisfaction analysis is a method used to evaluate customer satisfaction levels based on their interactions with a business or product. This analysis can help identify areas for improvement, monitor trends in customer satisfaction, and guide marketing strategies. In the context of the RFM data project, customer satisfaction can be inferred based on the customers' purchasing behavior, such as the frequency of their purchases, the monetary value of their purchases, and their recency.

#Algorithm

To perform satisfaction analysis on the OnlineRetail dataset, we first preprocess the data by removing missing values and calculating RFM metrics for each customer. Then, we can use a machine learning algorithm to model the relationship between customer satisfaction and the RFM metrics. For example, we could use a decision tree, linear regression, or neural network to predict customer satisfaction based on their RFM values. Alternatively, we could use clustering or classification techniques to group customers with similar satisfaction levels based on their purchasing behavior. Once we have created a model, we can use it to identify trends in customer satisfaction and tailor marketing strategies to target different segments of the customer base.

#Sample Output

```
Cluster average_satisfaction total_satisfied total_unsatisfied total_reviews
   <int>             <dbl>           <int>             <int>         <int>
1      1             0.502           65189             64663        129852
2      2             0.500           14825             14837         29662
3      3             0.501           16191             16099         32290
4      4             0.499          107322            107703        215025
```

Please, see attached files for the satisfaction graphical results

Regression Analysis

Regression analysis is a statistical method used to explore the relationships between variables and predict the value of a target variable based on the values of input features. In the context of the RFM data project, regression analysis can be used to model the relationship between customer behavior (e.g., recency, frequency, and monetary value) and various business outcomes, such as customer satisfaction, customer lifetime value, or future revenue.

#Algorithm

To perform regression analysis on the OnlineRetail dataset, we first preprocess the data by removing missing values and calculating RFM metrics for each customer. Next, we can use a regression algorithm, such as linear regression, logistic regression, or support vector regression, to model the relationship between the RFM metrics and the target variable. To evaluate the performance of the model, we can use metrics such as the R-squared value, the mean absolute error, or the root mean squared error. We can also visualize the results of the regression analysis using scatter plots or other graphical representations.

Recommendations

The recommendation section serves as a vital component of our project, as it aims to enhance customer satisfaction, foster long-term relationships, and increase revenue by providing personalized product suggestions. By analyzing historical customer purchase data, our system can recommend products tailored to individual customers' preferences, which in turn increases the likelihood of repeat purchases and promotes loyalty.

The importance of this section lies in its ability to leverage data mining techniques to uncover hidden patterns and associations within customer purchase data. By identifying these patterns, businesses can better understand their customers' needs and preferences, allowing them to offer a more personalized shopping experience. This not only improves customer satisfaction but also enables businesses to stay competitive by keeping up with evolving consumer trends.

To generate personalized recommendations, our system employs a combination of data mining methods and machine learning techniques. The process can be summarized as follows:

1. Data preparation: We start by selecting relevant columns from the dataset (CustomerID, StockCode, and Quantity) and removing any rows with missing values. The CustomerID and StockCode columns are then converted to factors, and a binary rating matrix is created. This matrix has customers as rows and products as columns, with binary values indicating whether a customer has purchased a product (1) or not (0).

2. Data mining: In this stage, we apply data mining techniques, such as association rule mining, to identify patterns and relationships within the customer purchase data. This helps us understand the underlying structure of the data and the connections between customers and the products they buy.

3. Recommendation model: Based on the insights gained from data mining, we build a recommendation model using the user-based collaborative filtering (UBCF) algorithm. The UBCF algorithm is a popular method for generating personalized recommendations, as it considers the preferences of similar users to suggest relevant products.

4. Generating recommendations

The trained UBCF model is used to predict the top 10 recommended products for each customer. The UBCF algorithm works by first identifying users who are similar to the target user based on their purchase history. It then aggregates the preferences of these similar users to generate a list of recommended products for the target user. This list is ranked based on a scoring system that accounts for the similarity between users and the frequency with which the products have been purchased by the similar users.

To generate recommendations for users we used the following steps:

a. We use the `predict` function from the recommenderlab package to generate recommendations for each user. The function takes three arguments: the trained recommendation model (recomm_model), the user's purchase data represented as a row in the user_item_matrix, and the number of recommendations to generate (n = 10).

b. The `predict` function returns a list of recommended items for each user in the form of a Top-N Recommender object. This object contains item indices that correspond to the recommended products

c. We extract the item indices from the Top-N Recommender object

d. Finally, we use the custom function 'get_stockcode_description' to convert the item indices into StockCode and Description values, which represent the actual products recommended to each user.

e. Retrieving product details: A custom function called 'get_stockcode_description' is created to convert the item indices from the recommendations into StockCode and Description values using the original dataset.

f. Visualization: The recommendations for users are visualized using ggplot2, illustrating the top 10 recommended products and their corresponding ranks.

By generating personalized recommendations for each user, the UBCF algorithm ensures that the product suggestions are tailored to the preferences and purchase history of individual customers, thus increasing the likelihood of future purchases and enhancing overall customer satisfaction.

#Sample Output

- Recommendations for User 1

```
# A tibble: 10 × 2
  StockCode Description
  <chr>     <chr>
 1 84520B    PACK 20 ENGLISH ROSE PAPER NAPKINS
 2 46000M    POLYESTER FILLER PAD 45x45cm
 3 22712     CARD DOLLY GIRL
 4 22684     FRENCH BLUE METAL DOOR SIGN 9
 5 44235     ASS COL CIRCLE MOBILE
 6 22230     JIGSAW TREE WITH WATERING CAN
 7 82613C    METAL SIGN,CUPCAKE SINGLE HOOK
 8 21823     PAINTED METAL HEART WITH HOLLY BELL
 9 21716     BOYS VINTAGE TIN SEASIDE BUCKET
10 22113     GREY HEART HOT WATER BOTTLE
```

- Recommendations for User 2

```
# A tibble: 10 × 2
  StockCode Description
  <chr>     <chr>
 1 21931     JUMBO STORAGE BAG SUKI
 2 22738     RIBBON REEL SNOWY VILLAGE
 3 22951     60 CAKE CASES DOLLY GIRL DESIGN
 4 21868     POTTING SHED TEA MUG
 5 21926     RED/CREAM STRIPE CUSHION COVER
 6 37379A    PINK CHERRY BLOSSOM CUP & SAUCER
 7 84527     FLAMES SUNGLASSES PINK LENSES
 8 84755     COLOUR GLASS T-LIGHT HOLDER HANGING
 9 21324     HANGING MEDINA LANTERN SMALL
10 22549     PICTURE DOMINOE
```

Please, see attached files for the recommendation graphical results

In conclusion, this report provides a comprehensive overview of the process involved in developing a personalized product recommendation system using data mining techniques and the user-based collaborative filtering (UBCF) algorithm. By analyzing customer purchase data, this system allows for the identification of patterns and relationships between customers and products, ultimately leading to a more tailored and engaging shopping experience.

The initial steps of the project involved data preprocessing and cleaning, which ensured that the dataset was suitable for analysis. This was followed by exploratory data analysis (EDA) to gain insights into the data and identify any trends or patterns that could inform the recommendation system. During the EDA phase, several key findings were discovered, such as the most popular products, seasonal trends, and customer purchase behavior. After the EDA, the data was transformed into a binary rating matrix that represented the relationship between customers and purchased products. This step was essential in preparing the data for the UBCF model, as it facilitated the identification of similarities between customers based on their purchase history.

The UBCF algorithm was then applied to generate personalized product recommendations for the users. This algorithm considers the preferences and purchase history of similar customers to recommend products that are more likely to resonate with the target users. The effectiveness of the UBCF algorithm was demonstrated through the generation of tailored product suggestions for each user, increasing the probability of future purchases and enhancing customer satisfaction.

In addition to the UBCF algorithm, other data mining methods were employed throughout the project to support the recommendation system. For instance, clustering algorithms such as k-means were used to group similar customers together, enabling a better understanding of customer preferences and behavior. Furthermore, association rule mining was utilized to discover frequent itemsets and uncover the relationships between products, which was also used to inform the recommendations.

The final step of the project was to evaluate the performance of the recommendation system using various metrics such as precision, recall, and F1-score. These metrics provided valuable insights into the quality and relevance of the generated recommendations, ensuring that the system is effective in providing personalized product suggestions to the users.

Overall, the successful implementation of this recommendation system showcases the power of data mining and collaborative filtering techniques in providing personalized and relevant product suggestions. By incorporating this system into an e-commerce platform, businesses can foster a more engaging and satisfying shopping experience, leading to higher customer retention, increased sales, and improved customer loyalty. This project serves as a strong foundation for future research and development in the field of recommendation systems, as well as the broader application of data mining techniques in e-commerce and other industries.