AMSTERDAM
UNIVERSITY OF
APPLIED
SCIENCES

08/11/2024

Business Statistics

Masoud Rezapour Najafi
Ferdinand Iguodala
Ruta Gebremedhin
Joep Welsch

Methodology

This report focuses on analyzing real business data to identify actionable insights that could help optimize sales, improve user engagement, and increase purchase completion rates. The methodology consists of two sub-assignments, each targeting distinct business questions and research objectives. The approach is based on a combination of data segmentation, clustering, RFM analysis, regression models, and recommendation systems, applied to both product and customer data. Below is a detailed breakdown of the methodology followed for this project:

1. **Preprocessing**

We used business data, Tao Yin, and Retail Rocket, which included items, customer attributes, and Session and User Engagement. Data cleaning and preparation were done to address missing values and outliers before analysis (We also analyzed outliers).

2. Item Segmentation and Analysis

Items were clustered based on characteristics like recency, frequency, price, and profitability. Descriptive statistics and RFM (Recency, Frequency, Monetary) analysis identified high-performing item segments. A regression model was used to evaluate how these factors influence profitability.

3. Customer Segmentation and Analysis

Customers were segmented using clustering techniques based on behavior and engagement metrics. We identified the most valuable segments in terms of revenue potential and targeted specific clusters for promotions, retention, or cross-selling.

4. Session and User Engagement Analysis

We examined how session characteristics (e.g., activity frequency and session duration) influenced purchase completion rates. User engagement was analyzed to predict purchase likelihood and future session returns.

5. Recommendation System

The recommendation system was optimized to suggest high-demand and cross-sell items to increase average order value and user engagement.

6. Data Visualization and Model Evaluation

Visualizations (e.g., scatter plots, bar charts) were used to interpret results, and model performance was evaluated using metrics like AUC to ensure the robustness of our insights.

This methodology allowed us to apply statistical theory to real-world business data, generating insights to enhance sales strategies and customer retention.

1.1 Business Question

How can we increase sales by identifying and targeting distinct item and customer segments from our business data?

To increase sales, the business can leverage distinct item and customer segments identified through data analysis. For items, clustering revealed that factors like recency, frequency, and price influence profitability, with certain segments showing higher engagement and sales potential. By targeting these high-performing item segments with tailored marketing and inventory strategies, the business can optimize sales. On the customer side, segments like the "Premium, High-Margin Customers" represent the most valuable group, ideal for targeting with new premium products, while the "Occasional, PriceSensitive Shoppers" cluster offers opportunities for retention and cross-selling efforts. Additionally, the "Low Average Order Value" segment could benefit from targeted promotions to boost engagement and spending, enhancing overall sales growth.

Research Question 1: What distinct item segments can be identified from the business data?

1.2 We focused on item segmentation by first dividing the data into 8 separate dataframes based on the main groupings. From these, we selected two groups, A and B, and performed clustering and analysis on each group separately to address the three sub-questions outlined below. Ruta did for B group and Ferdinand did for group A.

Sub-Question-1: What are the key characteristics (such as recency, frequency, profitability, and price) of items within each cluster? (Ferdinand & Ruta)

- This subquestion helps to identify patterns and distinguish between clusters, allowing us to understand how the items in each cluster differ across these features. We used descriptive statistics and visualization techniques to analyze the central tendencies (mean), variability (spread), and distribution of these features within each cluster

Sub-question-2: How does the behavior of item segments differ in terms of RFM? (Masoud & Joep)

- We analyze the behavior of items in each cluster based on their Recency, Frequency, and Monetary scores, along with the overall RFM_Score. This analysis helps us understand how different item segments perform in terms of customer engagement and profitability, allowing for better-targeted marketing and inventory strategies. By examining these RFM scores, we can identify which items are most likely to drive sales and which may require additional attention or promotions.

Sub-Question-3: Which item features strongly influence profitability? (Ruta & Ferdinand)

- This analysis examines how **Recency**, **Frequency**, and **Price** influence **profitability** using a multiple linear regression model. The model evaluates the relationship between these variables and the **profit of the item**

Sub-question-4: What distinct customers segments can be identified from the business data?

- The business data reveals several distinct customer segments, each with unique characteristics and potential value. The "Premium, High-Margin Customers" cluster emerges as the most valuable in terms of revenue potential, driven by high margins and monetary value per 1000 units. This group is also the ideal target for new premium product offerings. In contrast, the "Low Average Order Value" segment could benefit from more targeted marketing or promotions to boost engagement and spending. Additionally, the "Occasional, Price-Sensitive Shoppers" cluster, with its high recency and frequency, presents an opportunity for retention and crossselling efforts to further increase customer lifetime value.

Sub-Question-5 Which customer cluster represents the most valuable segment for the business in terms of revenue potential? (Masoud)

- Based on the graph, the "Premium, High-Margin Customers" cluster has the highest values for metrics like "margin per 1000 units" and "monetary value per 1000 units". This suggests that this cluster represents the most valuable segment in terms of revenue potential for the business.

Sub-Question-6: Which customer cluster would be the most suitable target for a new premium product offering? (Masoud)

- The "Premium, High-Margin Customers" cluster would likely be the most suitable target for a new "premium product offering", given their higher "margin per 1000 units" and "monetary value per 1000 units" metrics.

Sub-Question-7: Are there any customer segments that could benefit from more targeted marketing or promotions to increase their engagement? (Masoud)

1.3 The "Low Average Order Value" cluster has relatively "low values" across multiple metrics like "recency", "frequency", and "margin". This suggests that this segment could potentially benefit from more "targeted marketing" or "promotions" to increase their engagement and spending.

Sub-Question-8: Are there any customer segments that the business should focus on retaining or crossselling to? (Ferdinand)

The "Occasional, Price-Sensitive Shoppers" cluster has relatively high "recency" and "frequency" values compared to some other clusters. This indicates that they are engaged customers, so the business may want to focus on "retaining" and "cross-selling" to this segment.

2.1 Business Question

How can session characteristics and user engagement drive higher purchase completion rates to boost revenue?

To drive higher purchase completion rates and boost revenue, businesses should focus on enhancing user engagement through personalized recommendations, analyzing session characteristics that encourage interaction, and optimizing the browsing experience to convert interest into sales. By effectively leveraging these insights, companies can create strategies that not only improve the purchasing process but also foster long-term customer relationships

2.2 Research Question:

How can we session characteristics, user segmentation, and engagement levels influence purchase completion rates and optimize personalized recommendations?

- User segmentation based on engagement and session characteristics, combined with predictive modeling of browsing behavior, can significantly enhance purchase completion rates and optimize personalized recommendations by tailoring suggestions to each segment's unique behaviors and preferences, thus driving both immediate purchases and sustained engagement.

Sub_question-1: how can we segment users based on their engagement and behavior metrics? (Ruta)

- I segmented users based on engagement and behavior metrics by creating new features that capture essential aspects of user interactions. Using these engagement features, I clustered the users and conducted statistical analysis on each cluster to identify distinct behavioral patterns. This analysis, conducted in the Python notebook, provided insights that directly answer my research question. By examining how user engagement and behavioral metrics distinguish user segments, I highlighted unique characteristics and tendencies within each cluster.

Sub-question-2: How accurately can we predict the likelihood of a purchase based on browsing behavior? (Ruta)

- I focused on creating features that capture user engagement, specifically num_items_viewed and view_count. These metrics were chosen because they reflect the user's level of interest and interaction with items during browsing. Using these features, we built a predictive model aimed at estimating the likelihood of a purchase, providing insights into how browsing patterns are connected to buying decisions. I then evaluated the model's accuracy to assess the effectiveness of these engagement indicators in predicting purchase outcomes. This approach allowed us to understand the extent to which browsing behavior can signal purchase intent, helping to inform strategies that could target users with higher engagement

Sub-question-3: How do session characteristics influence the purchase completion rate within user segments? (Joep)

- Session characteristics such as activity frequency, event duration, and time since last purchase are the main drivers of purchase completion rates, particularly in clusters with higher engagement. Total viewing time often indicates indecision and can negatively impact purchase likelihood, while time since last event generally does not play a significant role across most clusters.

Sub-Question-4: Influence of User Engagement on Return for Future Sessions? (Ferdinand):

- Higher engagement during sessions, particularly through key actions and longer session lengths, significantly increases the likelihood of users returning for future sessions. This indicates that engaging users effectively can foster loyalty and repeat visits.

Sub-question-5: Impact of User Engagement on Purchase Behavior (Ferdinand):

- User engagement metrics, such as session length and focus on key actions, are critical drivers of purchase behavior. Specifically, increased engagement correlates with higher purchase rates, suggesting that the more users interact with the platform and its features, the more likely they are to make a purchase. However, excessive engagement per event may indicate friction, negatively affecting the purchase likelihood.

Sub-question-6: How does the recommendation system's focus on popular and cross-sell items enhance user engagement and revenue potential? (Masoud)

- The recommendation system is designed to boost user engagement and sales by promoting highdemand items, such as those frequently viewed or purchased. For example, users who buy a popular item like item 200793 are also recommended related products that are frequently bought together. This focus on popular products increases the likelihood of additional purchases. Additionally, the system's cross-selling capabilities suggest complementary items in real-time, enhancing the average order value. By aligning recommendations with users' interests and behaviors, the system supports both increased engagement and revenue growth.

Joep

This project has been both challenging and valuable for me. It marked the first time I had to balance two demanding projects alongside competitive sports, which I approached with seriousness and a strong desire to learn as much as possible.

Initially, formulating the research questions was tough, as we needed to approach the problem from a business perspective. This was interesting but challenging, especially in connecting the questions to a relevant dataset. In retrospect, I think I could have done better here, given my experience with business projects and research question formulation. Starting the project went smoothly, although I initially needed time to find my footing. However, as I progressed, my understanding deepened because I was actively engaging with the data and seeing the elements in action.

Sub assignment 2 was completed individually, and this presented a new challenge as it was my first Python project. Initially, I struggled with where to start, but as I moved forward, I learned the importance of understanding the "why" behind each step rather than accepting results at face value. For example, I mistakenly formulated a metric I thought represented the conversion rate, but it turned out not to be accurate.

Looking ahead, I'll focus on developing a deeper understanding of underlying principles and concepts to ensure I'm aligning project metrics accurately and making well-informed decisions.

Ruta

This project presented unique challenges and taught me a great deal. Balancing two major projects at once, along with the demands of competitive sports, was a first for me and pushed me to manage my time and energy with greater care. I was committed to getting the most out of the experience, but there were aspects I could have approached differently, particularly in terms of communication and collaboration.

Starting with the research questions, I found it difficult to establish a clear, business-oriented focus that would connect smoothly with the available data. This was an eye-opening part of the process, as it required looking beyond typical project angles to identify questions that aligned with business goals. In hindsight, I could have been more proactive in refining these questions, especially since I have experience with business-related projects. Taking more time here might have led to a stronger foundation for the rest of the project.

Once the project was underway, I felt like I made steady progress, but it took some time to fully understand the scope. My confidence grew as I worked with the data, and I saw how each element fit together, which made the project more engaging and meaningful. However, while I was focused on my individual responsibilities, I could have improved how I communicated with the team. Misunderstandings slowed us down at times, and a clearer exchange of ideas would have likely benefited the overall project outcome.

In the Tao Yin segment, I collaborated with Ferdinand, which was a good opportunity to dive deeper into the project while learning from a teammate. Despite feeling that I could contribute meaningfully, I noticed that we could have coordinated our efforts better. This is an area I'm committed to working on in the future, as strong communication can help projects run more smoothly and efficiently.

Later, I worked on a separate sub-assignment, which I completed individually. This was my first handson project using Python, and initially, I felt lost trying to determine the best starting point. Through trial

and error, I realized that it's crucial to understand the reasoning behind each step in coding. I initially set up a metric that I thought represented the conversion rate, but later found it didn't actually align with our goals, which highlighted how essential it is to understand metrics thoroughly.

For future projects, I plan to develop a stronger foundational knowledge of core concepts to ensure my metrics are accurate and decisions are well-informed. Additionally, I'll focus on improving communication and collaboration, as effective teamwork and clear exchanges of ideas can have a big impact on the project's success.

Ferdinand

Reflection on the Project

Embarking on this business statistics project was a transformative experience for me. Working with real datasets gave me a taste of the challenges and triumphs that come with analyzing actual business scenarios. I felt a sense of excitement as I connected theoretical concepts to practical applications, making the data come alive in a way that purely academic exercises couldn't.

However, balancing this project with my academic commitments was tough. The pressure of tight deadlines sometimes felt overwhelming, but it pushed me to refine my time management skills. I learned the value of being organized and the importance of prioritizing tasks to ensure that I stayed on track without compromising the quality of my work.

Looking back, I realize that a bit more upfront planning would have made a difference. There were moments when I felt a little lost, unsure of the best way to approach certain aspects of the analysis. Despite this, the experience taught me to embrace the learning process, to be patient with myself, and to seek help when needed.

What stands out most is how much I grew through this project. I not only improved my analytical skills but also gained insights into how data can inform and influence business decisions. Overall, this project was a valuable journey that reinforced my enthusiasm for data analysis and its role in real-world contexts.

Masoud

Working on this business statistics project was a truly valuable experience. It was great to work with real datasets, as it allowed us to apply theoretical knowledge to actual business scenarios. The challenge of analyzing and interpreting real-world data was both exciting and educational, providing us with deeper insights into the complexities of business analytics.

Having learned statistical theory throughout the course, we were able to approach the dataset and analysis from a more analytical and informed perspective.

However, balancing this project with a busy academic schedule was a bit overwhelming at times. Given the intense block schedule, we had to manage our time effectively to complete the project. While it was a lot of work, it pushed us to improve our time management and collaboration skills.

In hindsight, perhaps we could have planned a bit better to manage the workload more effectively. A bit more preparation in advance would have helped us streamline the process, but overall, the experience was incredibly rewarding. We learned a great deal, not only about business statistics but also about handling the pressures of tight deadlines and multitasking. This project has certainly enhanced our analytical skills and reinforced the importance of realworld data in decision-making.