Master in Management + Analytics

Predictive Customer Lifetime value modeling: Improving customer engagement and business performance

Autoría: Delgado, Matías M.
Fecha de defensa de la tesis: 2023

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

PREDICTIVE CUSTOMER LIFETIME VALUE MODELING: IMPROVING CUSTOMER ENGAGEMENT AND BUSINESS PERFORMANCE

THESIS

Matias M. Delgado

May 2023

Tutor: Lionel Barbagallo

Acknowledgments

I am deeply grateful for the unwavering support and guidance I have received from my family and friends, who have taught me the most important things in life. The invaluable encouragement and mentorship from my leaders at Despegar, Fernando Viola and Deborah Durlacher, have pushed me to reach beyond my comfort zone. I am indebted to Alvaro Osores for paving my way into the world of data. The remarkable leadership of Sam Giligich, Peter Garcia, and Collin Roth at Cookunity has provided me with an incredible platform to grow both personally and professionally. My work experience has been truly enjoyable thanks to the talent and dedication of my teammates. Lastly, I am immensely grateful to Nerea Villata, whose companionship has brought joy and beauty to my life.

Abstract

CookUnity, a meal subscription service, has witnessed substantial annual revenue growth over the past three years. However, this growth has primarily been driven by the acquisition of new users to expand the customer base, rather than an evident increase in customers' spending levels. If it weren't for the raised subscription prices, the company's customer lifetime value (CLV) would have remained the same as it was three years ago. Consequently, the company's leadership recognizes the need to adopt a holistic approach to unlock an enhancement in CLV.

The objective of this thesis is to develop a comprehensive understanding of CLV, its implications, and how companies leverage it to inform strategic decisions. Throughout the course of this study, our central focus is to deliver a fully functional and efficient machine learning solution to CookUnity. This solution will possess exceptional predictive capabilities, enabling accurate forecasting of each customer's future CLV. By equipping CookUnity with this powerful tool, our aim is to empower the company to strategically leverage CLV for sustained growth.

To achieve this objective, we analyze various methodologies and approaches to CLV analysis, evaluating their applicability and effectiveness within the context of CookUnity. We thoroughly explore available data sources that can serve as predictors of CLV, ensuring the incorporation of the most relevant and meaningful variables in our model. Additionally, we assess different research methodologies to identify the top-performing approach and examine its implications for implementation at CookUnity.

By implementing data-driven strategies based on our predictive CLV model, CookUnity will be able to optimize order levels and maximize the lifetime value of its customer base. The outcome of this thesis will be a robust ML solution with remarkable prediction accuracy and practical usability within the company. Furthermore, the insights gained from our research will contribute to a broader understanding of CLV in the subscription-based business context, stimulating further exploration and advancement in this field of study.

Resumen

CookUnity, un servicio de suscripción de comidas, ha experimentado un crecimiento sustancial en sus ingresos anuales en los últimos tres años. Sin embargo, este crecimiento se ha impulsado principalmente mediante la adquisición de más usuarios para ampliar la base de clientes, sin una mejora notable en los niveles de gasto de los clientes. Si no fuera por el aumento en los precios de suscripción, la empresa seguiría generando el mismo valor de vida del cliente (CLV, por sus siglas en inglés) que hace tres años. Por lo tanto, la máxima prioridad del liderazgo de la empresa es adoptar un enfoque integral para aumentar el CLV.

El objetivo de esta tesis es obtener una comprensión integral del CLV, sus implicaciones y cómo las empresas lo aprovechan para tomar decisiones estratégicas. A lo largo del desarrollo de este estudio, nuestro objetivo principal es entregar una solución de aprendizaje automático totalmente funcional y efectiva a CookUnity. Esta solución poseerá capacidades predictivas excepcionales, lo que le permitirá pronosticar con precisión el CLV futuro de cada cliente. Al proporcionar a CookUnity una herramienta tan poderosa, buscamos capacitar a la empresa para aprovechar el CLV como un activo estratégico para su crecimiento.

Para lograr este objetivo, analizamos varias metodologías y enfoques de análisis de CLV, considerando su aplicabilidad y efectividad en el contexto de CookUnity. Exploramos exhaustivamente las fuentes de datos disponibles que pueden servir como predictores de CLV, asegurándonos de incorporar las variables más relevantes y significativas en nuestro modelo. Además, evaluamos diferentes metodologías de investigación para identificar el enfoque más eficiente y examinar las implicaciones de su implementación para CookUnity.

Con la implementación de estrategias basadas en datos impulsadas por nuestro modelo predictivo de CLV, CookUnity podrá optimizar los niveles de pedidos y maximizar el valor de vida del cliente de su base de clientes. El resultado de esta tesis será una solución de aprendizaje automático sólida con una notable precisión en las predicciones y una utilidad práctica dentro de la empresa. Además, los conocimientos adquiridos en nuestra investigación contribuirán a una comprensión más amplia del CLV en el contexto de los negocios basados en suscripciones, estimulando una mayor exploración y avance en este campo de estudio.

Content Index

Figures Index

Tables Index

Equations Index

1. Introduction

1.1. CookUnity (CU): Transforming the Dining Experience

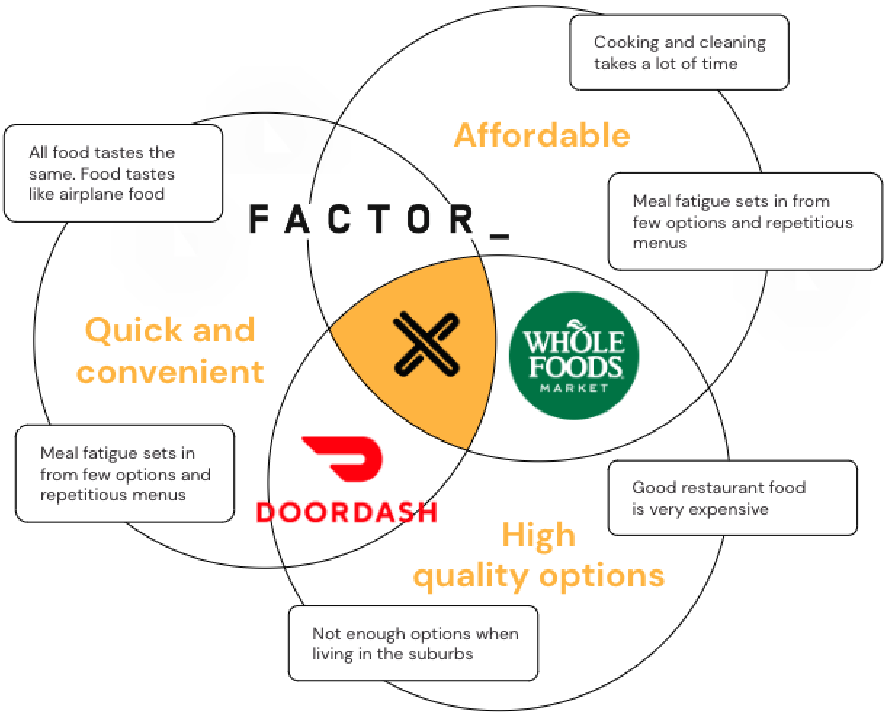
CookUnity (https://www.cookunity.com) is an innovative culinary platform that has gained significant popularity in the United States. It serves as a marketplace where independent chefs can showcase and sell their signature dishes directly to discerning consumers. This company has successfully transformed the traditional food delivery model by prioritizing exceptional, nourishing, and sustainable meals.

The platform offered by CookUnity allows customers to embark on a unique culinary journey, enjoying chef-prepared meals conveniently delivered to their doorsteps. Through a recurring weekly subscription, customers can access a diverse range of meticulously crafted dishes made with locally sourced ingredients. CookUnity's focus on health-consciousness and flavor ensures an exceptional dining experience.

One of CookUnity's strengths lies in its partnerships with a diverse array of chefs and culinary experts. By bringing together talented food creators, including vegan and vegetarian chefs as well as renowned culinary trailblazers, CookUnity expands the accessibility of their exceptional creations. This approach revolutionizes the way people appreciate and savor food, while promoting a sustainable and inclusive food culture.

CookUnity's platform has ushered in a new era of culinary possibilities, captivating the palates of discerning consumers. By connecting passionate chefs with eager customers, CookUnity not only transforms the dining experience but also champions sustainability and inclusivity in the culinary world.

Figure 1: CookUnity Value Proposal Benchmark


Cooking and cleaning takes a lot of time
Affordable
All food tastes the same. Food tastes like airplane food
FACTOR_
Meal fatigue sets in from few options and repetitious menus
Quick and convenient
WHOLE FOODS MARKET
Meal fatigue sets in from few options and repetitious menus
DOORDASH
Good restaurant food is very expensive
High quality options
Not enough options when living in the suburbs

1.1.1. Business Model: Empowering Customers with Choice

At the core of CookUnity's business model is a subscription service that offers customers the freedom to curate their own culinary experience. With an extensive selection of over 400 meal options (based on the user's zip code) that continues to grow, customers are presented with an abundance of choices each week.

Upon subscribing to CookUnity, customers are guided to personalize their meal preferences, indicating their culinary tastes and dietary restrictions. They also have the flexibility to select a meal plan that suits their needs, ranging from 4 to 16 meals per week[1]. This personalized approach ensures that customers receive meals that align with their specific preferences and dietary requirements.

CookUnity unveils its enticing menus for the upcoming week two weeks in advance. Customers are presented with a diverse range of options and have the freedom to make their own selections. They can handpick the specific meals they desire, customizing their culinary experience to suit their preferences. Alternatively, they have the option to skip a particular week or pause their meal deliveries for an extended period.

For customers seeking a seamless experience, CookUnity offers an auto-pilot feature. In the absence of specific actions taken by the customer, the company's algorithms match each customer with the most suitable meals based on their preferences and automatically place the order on their behalf. This convenient option ensures that customers never miss out on the delight of a delectable meal, even when they are pressed for time.

CookUnity's meals are delivered fresh, rather than frozen, and come in packaging that is compostable, recyclable, or reusable. Each package is thoughtfully labeled with expiration dates, heating instructions, and comprehensive nutrition information, empowering customers to make informed choices about their meals.

Through its innovative and customer-centric business model, CookUnity revolutionizes the dining experience by empowering individuals to embark on a culinary journey tailored to their tastes and preferences. With a steadfast commitment to freshness, sustainability, and convenience, CookUnity sets the stage for an exceptional dining experience.

1.1.2. Growth Drivers: Subscriptions, Order Rate, and Average Order Value

CookUnity embarked on its subscription-based business journey in the dynamic market of New York in the second half of 2019. Since then, the company has experienced promising growth in its annual recurring gross revenue (Gross ARR). This steady upward trajectory propelled CookUnity to successfully secure series A and B fundraising rounds, with the most recent one completed by April 2021.

Looking towards the future, CookUnity faces a new challenge presented by its investor group. The company is now tasked with maintaining its growth momentum without significantly increasing its marketing expenditure. In other words, CookUnity must find ways to maximize the efficiency of its marketing spend to facilitate the journey towards reaching breakeven.

Considering CookUnity's core business model is a weekly subscription service, the company generates revenue on a weekly basis. This revenue can be calculated as the product of three key

[1] Snapshots of CookUnity’s web platform can be found in Appendix A.

factors: the number of customers subscribed (Customers), the ratio of customers who place an order (Order Rate), and the average value of each order (AOV).

$$Revenue\ = Customers\ *\ Order\ Rate * \ AOV \tag{1}$$

Where:

- Customers: Represents the number of customers subscribed to the business.
- Order Rate: Refers to the ratio of customers who place an order.
- AOV: Represents the average order value (revenue).

By using this formula, we can illustrate the revenue generation process. For example, if CookUnity has 50,000 active customers in a given week, with 50% of them placing an order (25,000 ordering customers), and an average order value of $100 USD, the revenue for that particular week would amount to

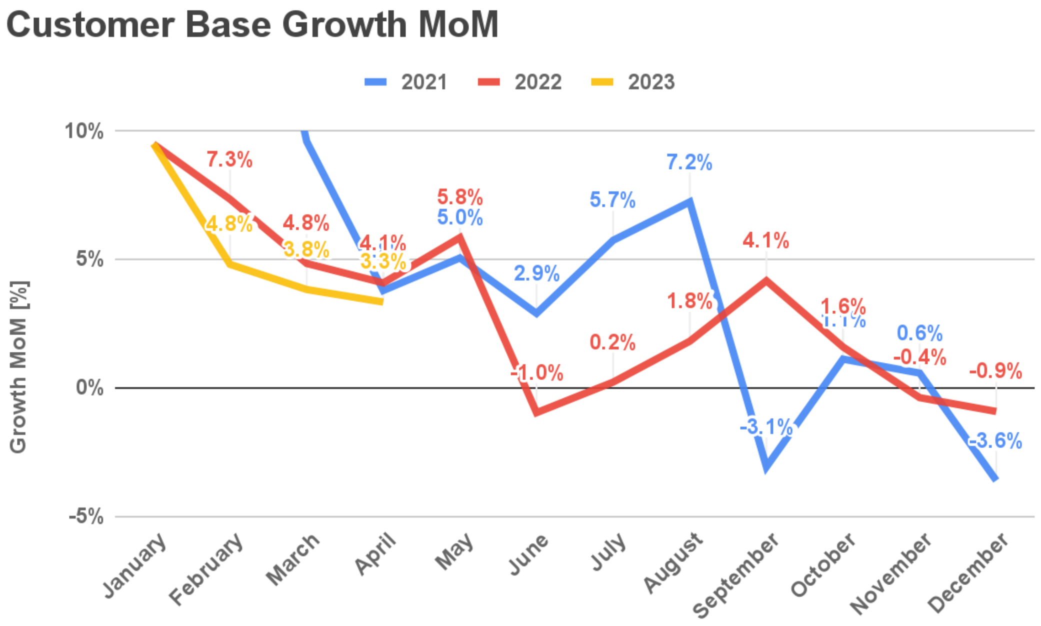
$$Revenue\ = 50{,}000\ customers * 0.5\ \frac{order}{customer} *\ 100\ \frac{usd}{order}\ =\ 2.500.000\ usd \tag{2}$$

This revenue formula allows us to analyze CookUnity's growth through three distinct levers: the number of subscribed customers, the order rate, and the average order value. By carefully evaluating and strategically enhancing these growth drivers, Cookunity can seize the opportunity to propel its business forward, demonstrating the company's ability to adapt, evolve, and thrive in the competitive subscription-based market landscape.

1.1.2.1. Subscribed Customers: Seasonality and Retention Insights

The growth of CookUnity's customer base shows a clear pattern influenced by seasonality, as depicted in Figure 2. From January to December over the past three years, there are noticeable fluctuations in the monthly customer growth rates.

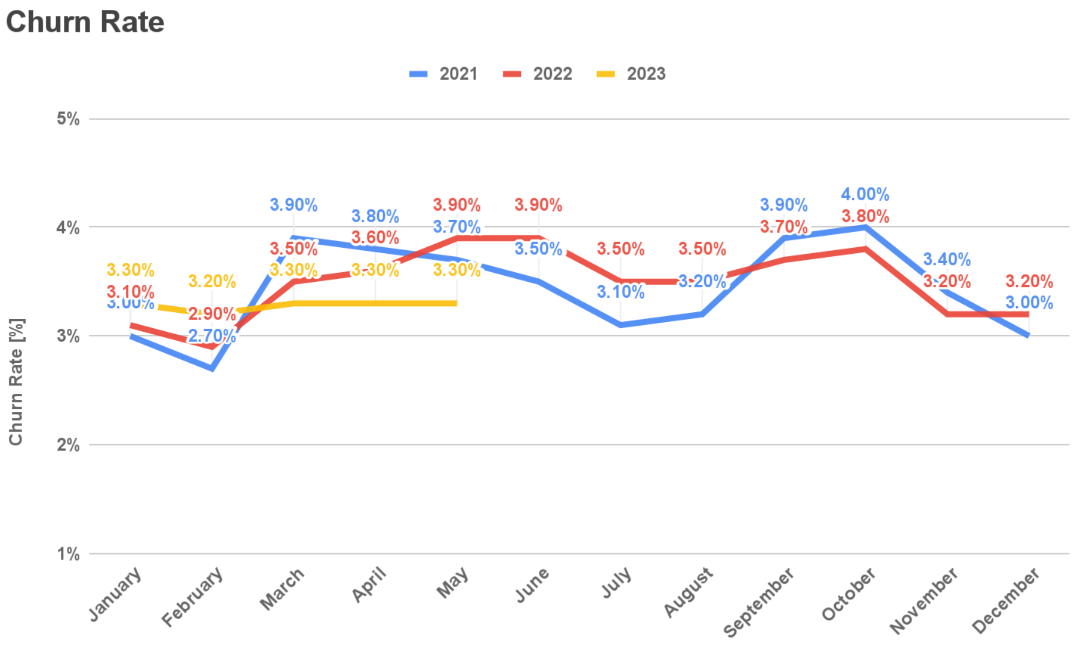
Figure 2: Subscribed Customer Base Growth MoM



The fluctuations in the customer base can be primarily attributed to two factors: customer acquisition levels (the number of customers subscribing) and churn levels (the number of customers unsubscribing).

Analyzing the churn rate provides valuable insights into customer retention. Remarkably, Figure 3 showcases the churn rate over the last three years, indicating that CookUnity's performance in retaining users has remained relatively consistent.

Figure 3: Churn Rate Performance - Excluding New Users

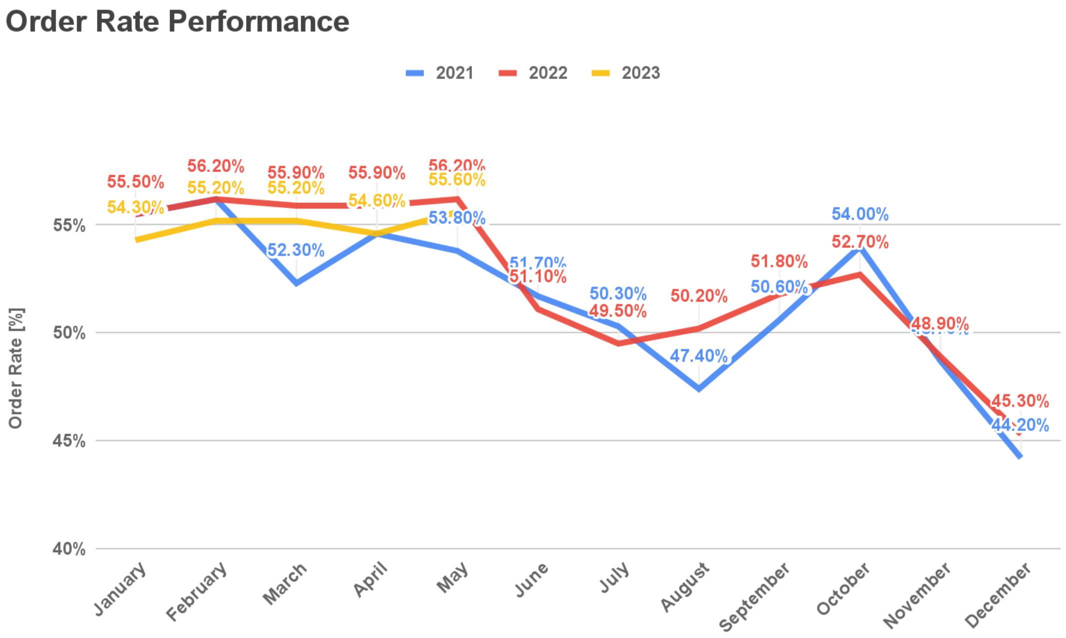


These findings suggest that while CookUnity experiences fluctuations in its customer base due to seasonality, the company has not made significant strides in improving its user retention rates. It's important to note that, due to CookUnity's flexible subscription options, a customer not churning does not necessarily imply placing orders. Customers have the option to skip or pause their subscriptions, which contributes to a higher number of customers remaining subscribed to the platform.

1.1.2.2. Order Rates: User Ordering Behavior

Analyzing the monthly order rates over the past three years reveals a cyclic pattern influenced by seasonality. High seasons coincide with new year health resolutions and the convenience of staying at home during winter, while low seasons occur during summer breaks and end-of-year religious holidays. As shown in Figure 4, CookUnity's ability to drive customers to place orders has not shown significant advancements over the past three years.

Figure 4: Order Rate Performance

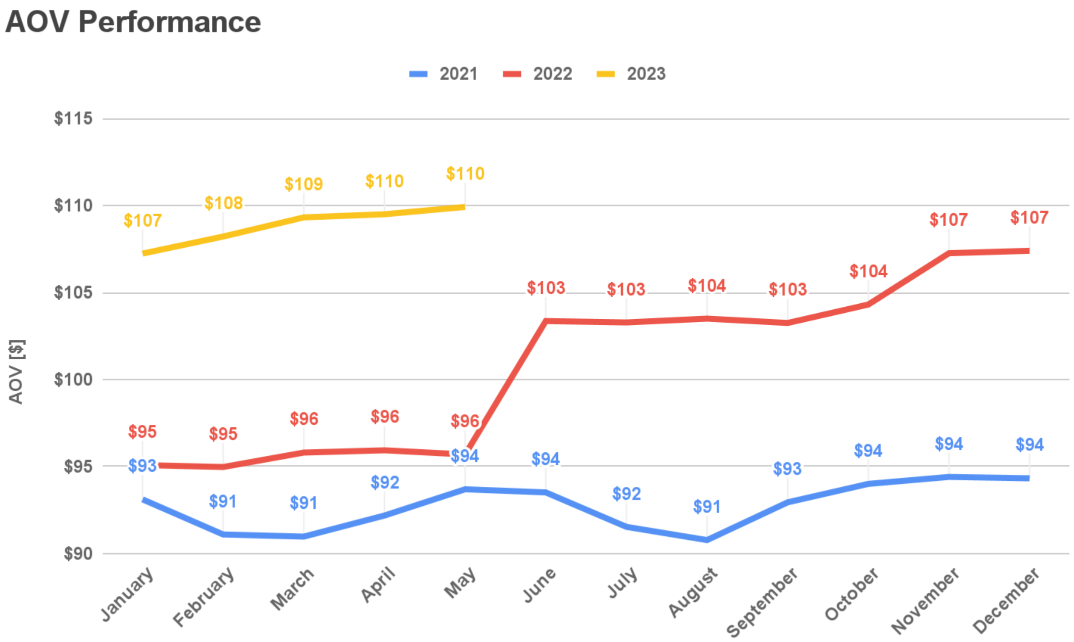


Understanding user behavior and motivations behind order placement is pivotal for CookUnity to optimize its order rates and stimulate growth. By identifying potential barriers or incentives that influence customers' decision-making, CookUnity can devise targeted strategies to enhance the frequency and consistency of customer orders. A data-driven approach, combined with effective marketing and communication tactics, can empower CookUnity to maximize its order rates, driving revenue growth and fostering stronger customer engagement.

1.1.2.3. Average Order Value (AOV): Stability and Price Adjustments

In contrast to the previously discussed metrics, the average order value (AOV) exhibits a relatively consistent pattern throughout the year without being influenced by seasonality, as depicted in Figure 5. Over the past three years, strategic adjustments were made to CookUnity's subscription plans in response to US inflation rates, resulting in an upward trend in AOV.

Figure 5: AOV [Gross] Performance



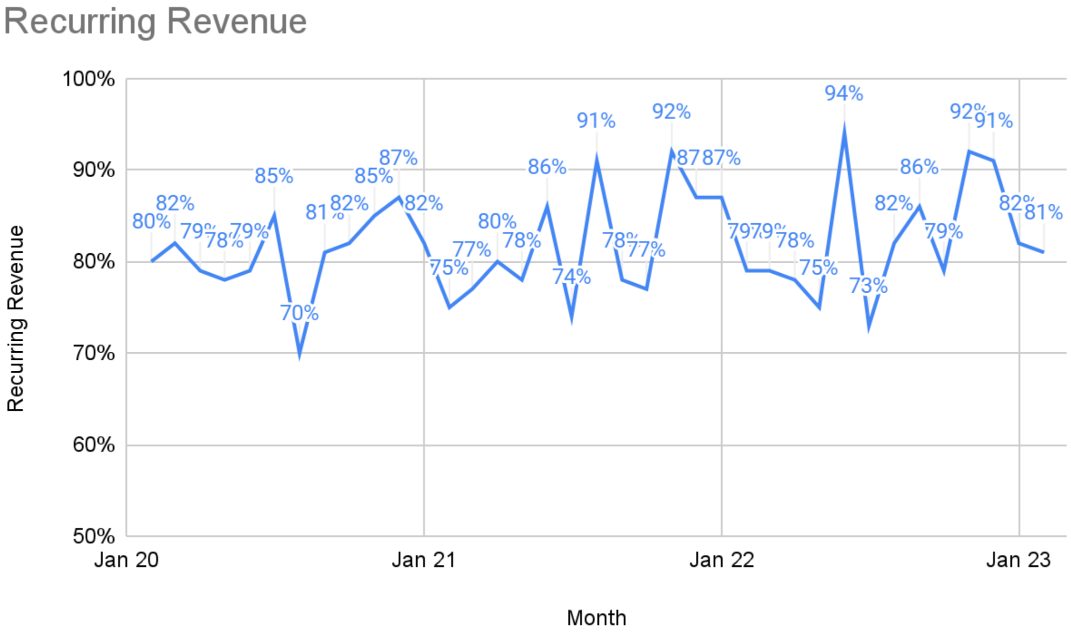
Significant increases in AOV were implemented in June 2022 for all users, in October 2022 targeting new subscribers, and in February 2023 for all users. These price adjustments have successfully raised the AOV and contributed to the company's overall revenue growth.

However, CookUnity should not rely solely on these price adjustments to drive revenue growth. It is crucial for the company to explore additional avenues and strategies to expand its income streams and foster sustainable revenue growth. By diversifying its approaches and focusing on value creation, CookUnity can ensure long-term financial success.

1.1.3. Sustaining Growth: Acquisition Dependency

Figure 6 shows the percentage of the monthly revenue being generated from previous months acquisition or in other words which share of the total revenue would have the company made if it hadn't acquired users on each given month. This analysis of CookUnity's Gross ARR reveals a consistent trend: a significant portion of monthly revenue, from 20% to 30%, is lost from one month to the next. This indicates the challenge of retaining ordering customers over time.

Figure 6: Recurring Revenue



Remarkably, this loss of ordering users from month to month has remained consistent over the years, aligning with the earlier analysis of the growth drivers. The stability of these metrics indicates that CookUnity has been compensating for the loss of ordering customers by continuously acquiring new customers. However, the board members have expressed concerns about the long-term sustainability and growth trajectory of the company due to this acquisition dependency.

To address these concerns, CookUnity's leadership is urged to prioritize the improvement of customer ordering retention. This can be achieved through strategies that enhance the overall customer experience, personalized recommendations, targeted promotions, and improved communication channels. By engaging customers on a deeper level and fostering stronger relationships, CookUnity can encourage longer ordering periods and generate more revenue over the customer's subscription period. This total amount of revenue a customer is expected to generate over the course of his relationship with the business is defined as Customer Lifetime Value.

Optimizing marketing spend return on investment (ROI) is also crucial in improving customer retention and Customer Lifetime Value (CLV). By leveraging data analytics and customer insights, CookUnity can identify effective marketing channels, refine targeting strategies, and allocate resources to campaigns with the highest ROI. This data-driven approach will maximize the value derived from each customer and reduce dependency on constant customer acquisition.

Addressing CLV not only mitigates acquisition dependency but also makes CookUnity a more attractive investment opportunity. By demonstrating a strong focus on sustainable revenue growth and customer ordering retention, CookUnity can instill confidence in potential investors and pave the way for successful fundraising opportunities.

1.2. Understanding Customer Lifetime Value (CLV)

Customer Lifetime Value (CLV) is a crucial metric that quantifies the total revenue a customer is expected to generate for a business throughout their entire relationship with the company. This metric holds particular importance for subscription-based businesses, like CookUnity, as it allows them to evaluate the long-term value and profitability of each customer. By estimating CLV, companies can make informed decisions regarding customer acquisition, retention strategies, and overall business growth.

While there are various methods to calculate CLV, we will explore them further in the literature review section. However, one direct calculation approach involves multiplying two essential factors: Average Revenue per User (ARPU) and the Contribution Margin (CM%).

$$Historic\ CLV\ = ARPU\ * CM\% \tag{3}$$

ARPU, which stands for Average Revenue per User, represents the average amount of revenue generated by a customer during their lifespan with the business. It can be calculated by multiplying three key variables: the expected length of the customer relationship (lifespan), the ordering cadence during that lifespan (order rate), and the average order value (AOV).

$$ARPU\ = Lifespan\ *\ Order\ Rate *\ AOV \tag{4}$$

For instance, consider a subscription-based business where customers have an average lifespan of 24 weeks, order on average 70% of those weeks, and have an average order value of $50. In this scenario, the ARPU would amount to $840. This means that, on average, each customer is expected to generate $840 in revenue over the course of their lifetime with the company.

The Contribution Margin (CM%) represents the proportion of revenue that remains after deducting the cost of goods sold (COGS). In the case of CookUnity, the company calculates the overall revenue for each month and then subtracts the total cost of goods sold (COGS) to obtain the Contribution Margin (CM). However, at this stage of CookUnity's maturity, the company does not calculate CM% at the product level. Instead, a constant CM% is assumed for the company as a whole. This means that strategic decisions based on metrics such as Average Revenue Per User (ARPU) or Customer Lifetime Value (CLV) have a consistent impact on the company's profitability.

CLV can provide valuable insights to CookUnity's decision-makers and investors. By assessing the projected revenue from each customer over their lifetime, the company can prioritize initiatives aimed at maximizing customer value, improving retention rates, and optimizing marketing and operational strategies.

In the upcoming section, we will delve into how CookUnity can leverage CLV to enhance customer retention, increase ordering levels, and ultimately drive long-term growth for the company.

1.2.1. Leveraging CLV for Long-Term Growth

CookUnity can utilize CLV as a strategic tool to enhance customer retention, increase ordering levels, and drive long-term growth. By understanding the long-term value of each customer, the company can tailor its initiatives and marketing efforts to maximize customer satisfaction and loyalty.

Analyzing customer behavior and preferences enables CookUnity to personalize the customer experience, offer tailored promotions, and address pain points, encouraging customers to continue ordering for longer durations. By understanding the ordering cadence and preferences of different customer segments, CookUnity can develop strategies to increase order frequency and consistency. This can involve flexible subscription plans, new menu options, and incentives to encourage customers to order more frequently.

Furthermore, CLV can guide CookUnity in optimizing its marketing and acquisition strategies for long-term growth. By understanding the lifetime value of customers, the company can allocate marketing resources effectively, targeting segments with high CLV potential. This includes identifying acquisition channels that attract customers with higher CLV, refining messaging and targeting strategies, and optimizing the cost-to-value ratio of customer acquisition. Nurturing existing customers through personalized marketing campaigns and loyalty programs can also drive customer advocacy and word-of-mouth referrals, contributing to growth.

By integrating CLV into its decision-making processes, CookUnity can align its initiatives and resources towards maximizing customer lifetime value. This customer-centric approach enhances satisfaction, loyalty, and fosters sustainable long-term growth for the company.

1.3. Objective of The Study: Unlocking CLV as Pathway to Compounded Growth

In order to achieve sustainable long-term growth, companies must go beyond individual metrics such as order rate or churn rate. CookUnity, despite experiencing significant growth in annual recurring gross revenue (ARR) driven by customer acquisition and price adjustments, has faced challenges in improving the ordering behavior of its users. Simply focusing on optimizing one metric at a time will not suffice to meet the requirements for a successful series C fundraising. To overcome this hurdle, CookUnity must embrace a holistic approach by leveraging Customer Lifetime Value (CLV). By developing a predictive CLV model at the customer level, the company can gain deeper insights into each customer's potential future value and make data-driven decisions to enhance customer retention, increase ordering levels, and ultimately drive sustainable growth.

1.3.1. Main objective: “Develop a CLV Predictive Model”

The main objective of this study is to develop a robust CLV predictive model at the customer level for CookUnity. By leveraging historical data and customer characteristics, the model will forecast the potential value that each customer can bring, enabling the company to enhance customer retention, increase ordering levels, and drive sustainable growth. Embracing CLV as the guiding metric will optimize resource allocation and strengthen CookUnity's position in the highly competitive market.

1.3.2. Secondary Objective: “Understanding Consumer Ordering Drivers”

The CLV predictive model will provide valuable insights into what drives consumer spending within CookUnity's customer base. By accurately predicting CLV, the company can identify high-value customers and tailor personalized strategies to increase their engagement. This understanding of consumer behavior will inform customer relationship management (CRM) campaigns, guide product development initiatives, and uncover new features to further drive ordering behavior and maximize revenue generation.

Utilizing the CLV predictive model, CookUnity will be able to identify customers who exhibit lower ordering engagement. By pinpointing these customers in advance, the company can implement targeted retention initiatives and re-engagement campaigns. This proactive approach will enable CookUnity to prolong customer relationships, and maximize CLV.

1.4. Structure of the Study

The thesis will be structured as follows:

- **Literature Review**: This section will provide an overview of how subscription businesses work and their main performance metrics. Will tackle CLV and its significance in subscription businesses, different approaches to calculate CLV, and the strategic decisions based on CLV. Additionally, it will review various modeling approaches for predictive CLV and reference previous work in this area.
- **Methodology**: This section will outline the methodology adopted in the study. It will explain the target variable, available data sources, modeling approach, and model validation techniques.
- **Development**: This section will dive into the practical aspects of the study. It will cover exploratory data analysis (EDA) to understand customer behavior, feature engineering techniques, cross-validation, evaluation of different models (such as Random Forest, XG Boost, Light GBM, Neural Networks), and their performance assessment.
- **Discussion on Results**: This section will analyze and discuss the findings of the models developed. It will explore the model performance across customer segments, identify key performing segments, highlight challenges and limitations, and propose strategies for targeting and engaging specific customer segments.
- **Conclusion and Recommendations**: This section will summarize the study, emphasizing the main findings and implications. It will provide recommendations for further research and implementation based on the insights gained from the predictive CLV model.

By following this structured approach, the thesis aims to provide a comprehensive understanding of CLV, its application in CookUnity's context, the development of a predictive CLV model, and the analysis of its impact on customer retention, ordering levels, and long-term growth.

2. Literature review

Before delving into our own solution, it is essential to lay a strong foundation by gaining a comprehensive understanding of key concepts and previous work in the field. This section serves as an exploration of the literature surrounding subscription businesses and Customer Lifetime Value (CLV). We will delve into the significance of CLV in the context of subscription-based models, examining performance metrics and the strategic decisions that can be made based on CLV insights. Additionally, we will explore different approaches to calculate CLV, including historic CLV, cohort CLV, and predictive CLV. Furthermore, we will review various modeling approaches used for predicting CLV at the customer level. By delving into the existing body of knowledge and research, we will establish a solid framework for our own solution and gain valuable insights to inform our methodology and findings.

2.1. Subscription Business

Subscription businesses have disrupted traditional models by shifting the emphasis from selling individual products or services to creating ongoing, long-term relationships with customers. Rather than focusing solely on the transactional aspects of a sale, these businesses prioritize building strong, lasting connections with their subscribers. This approach allows them to establish a reliable, recurring revenue stream and, in turn, foster greater customer loyalty and engagement (Warrillow, 2015).

Subscription businesses have changed traditional business models in several ways, among which we can emphasize:

- **Recurring Revenue**: Subscription businesses create predictable revenue streams by charging customers on a recurring basis. This allows companies to plan for growth and invest in new products or services (Tzuo & Weisert, 2018).
- **Customer Loyalty**: Subscription businesses prioritize customer satisfaction and retention, which leads to long-term relationships with customers. This is different from traditional business models, which often focus on one-time sales (Tzuo & Weisert, 2018).
- **Data-Driven Decisions**: Subscription businesses collect data on their customers' behavior and preferences, which can be used to improve their products and services. This data can also be used to make data-driven decisions about marketing, pricing, and other business strategies (Baxter, 2015).
- **Flexibility**: Subscription businesses can offer customers flexible plans that can be adjusted based on their needs. This allows customers to customize their experience and only pay for what they need (Ries, 2011).
- **Upselling and Cross-Selling**: Subscription businesses can use data to identify opportunities to upsell or cross-sell products or services to customers. This can increase revenue and provide additional value to customers (Tzuo & Weisert, 2018).

Overall, subscription businesses have revolutionized traditional business models by shifting their focus from one-time transactions to building enduring relationships with customers. By prioritizing recurring revenue, customer loyalty, data-driven decision-making, flexibility, and upselling/cross-selling opportunities, these businesses have redefined the way companies interact with their customers.

2.1.1. Performance Metrics

Measuring the success of a subscription business requires a comprehensive understanding of key performance metrics. Notable authors such as Tzuo & Weisert (2018) have emphasized the significance of these metrics in assessing and optimizing business performance. The following performance metrics are considered essential in evaluating the success and sustainability of a subscription business:

- **Annual Recurring Revenue (ARR)**: This metric quantifies the total revenue generated annually by the subscription business, providing insights into its growth or decline over time.
- **Customer Acquisition Cost (CAC)**: CAC encompasses the expenses associated with acquiring new customers, including marketing and sales costs. Monitoring CAC enables businesses to assess the efficiency of customer acquisition efforts and identify potential areas for improvement.
- **Churn Rate:** This is the rate at which customers cancel their subscriptions. A high churn rate can indicate that a business is failing to provide value to its customers, and therefore they decide to cancel their subscription, while a low churn rate indicates that customers are satisfied and likely to continue their subscriptions.
- **Gross Margin:** This is the difference between the revenue generated by the subscription business and the cost of goods sold (COGS). A high gross margin indicates that a business is generating revenue efficiently and is likely to be profitable.
- **Average Revenue Per User (ARPU):** This is the average revenue generated per customer. A higher ARPU demonstrates substantial revenue generation from each customer, contributing to overall profitability.
- **Customer Lifetime Value (CLV):** This is the total revenue a customer is expected to generate over the lifetime of their subscription. A high CLV indicates that a business is successfully retaining customers and generating recurring revenue over a long period of time.

By diligently tracking and analyzing these performance metrics, subscription businesses gain valuable insights into their operational strengths and weaknesses. This data-driven approach enables informed decision-making, empowers performance improvement, and facilitates sustainable business growth.

2.2. Customer Lifetime Value

2.2.1. Strategic Decisions based on CLV

In order to make strategic decisions that drive customer acquisition and retention, businesses rely on the insights provided by Customer Lifetime Value (CLV). CLV serves as a powerful metric that enables teams to focus their energies on the areas that yield the greatest returns and enhance the overall customer experience.

Retention, as a direct measure of customer experience, plays a crucial role in determining whether customers will continue their engagement with a business. By analyzing customer behavior and identifying key moments of truth throughout the customer journey, businesses can proactively address potential issues and provide the necessary resources and support to ensure customer success. This approach involves mapping out the customer journey, from onboarding to renewal, and understanding the critical milestones that influence customer satisfaction and loyalty (Heath & Heath, 2017).

To achieve customer loyalty, it is essential to understand and cater to the unique needs and preferences of different customer segments. By identifying high-value and low-value customers, businesses can personalize experiences and create targeted strategies that maximize engagement and build loyalty. High-value customers, who contribute significantly to revenue, require tailored approaches to foster their continued engagement. Conversely, low-value customers may benefit from initiatives aimed at increasing their purchasing frequency or expanding their product adoption. By studying customer behavior and engagement patterns, businesses gain insights into how to effectively engage each customer segment and create experiences that drive loyalty and increase retention.

By calculating CLV, businesses gain a comprehensive understanding of the long-term value that customers bring to the organization. This knowledge guides strategic decision-making, such as allocating resources for customer acquisition and retention efforts. A business might choose to invest more in marketing campaigns targeting customers with a higher CLV, recognizing the potential for greater returns on investment. Simultaneously, efforts to reduce churn and enhance customer satisfaction can be prioritized to increase the lifetime value of existing customers (Mehta et al., 2016).

By identifying the factors that contribute to customer value and analyzing engagement patterns, companies can implement targeted strategies to enhance the overall ordering experience. This could involve refining the onboarding process, personalizing recommendations, or implementing retention initiatives aimed at increasing customer loyalty and maximizing CLV (Fleming, 2016).

By leveraging CLV as a guiding metric, businesses like CookUnity can make informed strategic decisions that optimize resource allocation, enhance customer experiences, and drive sustainable growth in the competitive market landscape.

2.2.2. Approaches to calculate CLV

In order to accurately assess the long-term value of customers, various approaches have been developed to calculate Customer Lifetime Value (CLV). These approaches take into account different factors and considerations to provide businesses with insights into the revenue potential and profitability associated with their customer base. By understanding the different approaches to calculating CLV, businesses can make informed decisions regarding customer acquisition, retention strategies, and overall business growth. This section explores the key methodologies used to estimate CLV and their implications for strategic decision-making.

2.2.2.1. Historic CLV

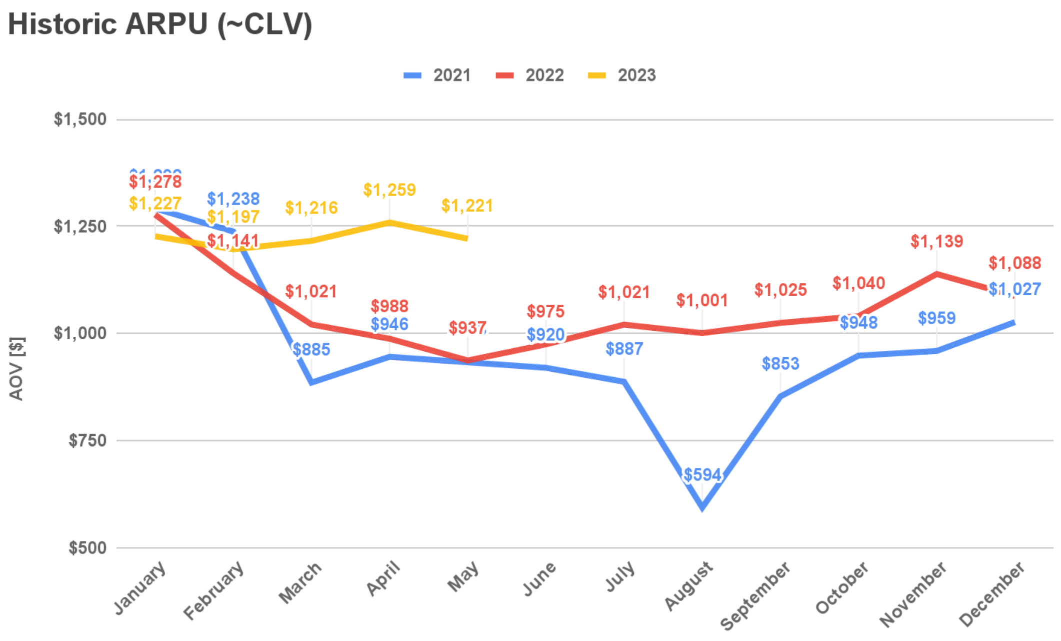
Historic CLV, as the name suggests, relies on analyzing customers' past behavior to estimate their future value. This approach involves calculating the average revenue generated per customer over a specific time period and multiplying it by the average customer lifespan. By considering historical data, businesses can gain insights into the revenue potential of their customer base (Equation 3).

For example, let's consider the evaluation of the Average Revenue Per User (ARPU) for January 2023. To calculate this metric, we replace the relevant variables in Equation 4. The average customers' lifespan is determined to be 24.3 weeks, the order rate for that month is 54.3%, and the Average Order Value (AOV) is calculated using the net revenue, which amounts to $93.07. By substituting these values into the equation, we obtain an ARPU of $1,227 for January 2023.

From a financial perspective, historic CLV is often used as a benchmark by investor groups to assess the value of subscription businesses. Its simplicity and straightforward calculation make it a popular choice for evaluating business performance (Gupta, 2014) (Gallo, 2014).

CookUnity also tracks CLV using this method to monitor its financial performance (Figure 7), as it provides a general overview of the customer value over time. However, it is important to note that historic CLV may have limitations when it comes to accurately representing the true value of customers. In the following section, we will explore another approach called Cohort CLV, which provides a more granular and representative analysis of customer lifetime value.

Figure 7: CookUnity's Historic ARPU



2.2.2.2. Cohort CLV

Cohort CLV is an approach that focuses on analyzing the lifetime value of groups of customers who share similar characteristics or exhibit similar behaviors. By examining cohorts, which are groups of customers who joined the business during a specific period (e.g., a particular month), businesses can gain deeper insights into customer behavior patterns and make targeted marketing and retention strategies.

To calculate Cohort CLV, we can follow the formula:

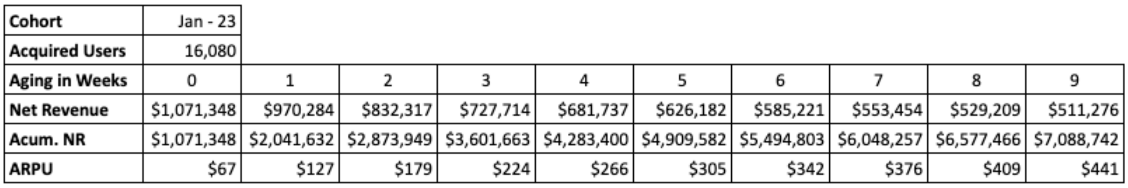
$$Cohort\ CLV\ = (\sum_{i=1}^{n}\ Revenue\ at\ Wi\ )/\ Acquired\ Users \tag{5}$$

Where:

- ***Wi*** is the week of aging the customer is at.
  - For example if a customer subscribed at week 22 (iso calendar) by week 25 he will be three weeks of aging.
- **Acquired Users** is the total number of subscribers acquired in a specific period.

As exemplified in Table 1, in January 2023, CookUnity acquired a total of 16,080 new users. By week 9, these acquired users had collectively generated $7.1 million in net revenue. Calculating the average net revenue per acquired user, we find that it amounts to $441 per user.

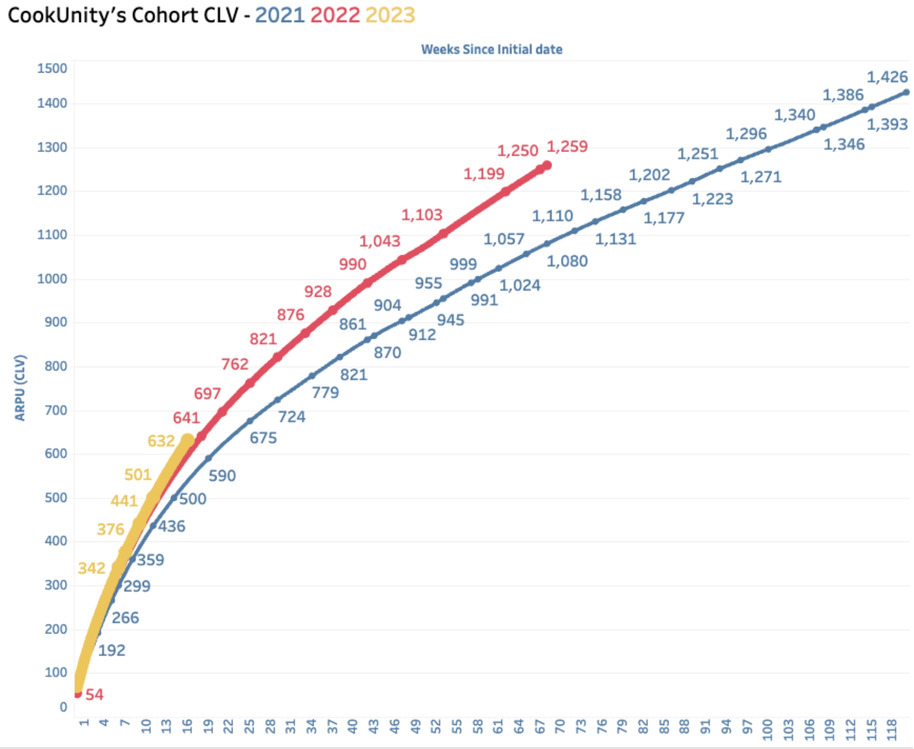
Table 1: CookUnity's Jan-23 Cohort CLV



Cohort	Jan - 23									
Acquired Users	16,080									
Aging in Weeks	0	1	2	3	4	5	6	7	8	9
Net Revenue	$1,071,348	$970,284	$832,317	$727,714	$681,737	$626,182	$585,221	$553,454	$529,209	$511,276
Acum. NR	$1,071,348	$2,041,632	$2,873,949	$3,601,663	$4,283,400	$4,909,582	$5,494,803	$6,048,257	$6,577,466	$7,088,742
ARPU	$67	$127	$179	$224	$266	$305	$342	$376	$409	$441

By applying this calculation to CookUnity's customer data, we can observe interesting trends (see Figure 8). For instance, when analyzing customers acquired in January '21 (blue line), we see that 118 weeks after being acquired their average revenue per user (ARPU) has reached $1.426 and this value maintains an upward trend. Similarly, customers acquired in January '22 (red line) have achieved an ARPU of $1.259 by their week 70, indicating that their lifetime value is expected to surpass that of the previous January cohort when reaching the same maturity.

Figure 8: CookUnity's Cohort CLV



These findings highlight the limitations of the linear calculation performed with the historic CLV method. Cohort CLV provides a more nuanced understanding of customers' actual lifetime value and enables businesses to make more accurate predictions about future revenue potential.

2.2.2.3. Predictive CLV

Predictive CLV is an advanced approach that leverages statistical modeling techniques to forecast a customer's future value based on a comprehensive set of factors. This method takes into account various variables, including historical purchase behavior, demographic information, and customer engagement with the business. By utilizing predictive CLV, businesses can obtain more accurate estimations of a customer's lifetime value, as it incorporates a wider range of data points and can adapt to changes in customer behavior over time.

Unlike historic CLV, which provides a retrospective analysis of customer value, predictive CLV offers insights at the individual customer level**.** This granularity enables businesses to develop tailored marketing and retention strategies that cater to the unique needs and preferences of each customer. By understanding the expected future value of customers, CookUnity can provide personalized solutions and experiences, enhancing customer satisfaction and driving long-term loyalty.

While CookUnity currently lacks a customer-level predictive CLV solution, implementing this approach will empower the company to unlock valuable insights and optimize its strategies accordingly. In the upcoming section, we will explore previous studies and examine different methodologies employed to address predictive CLV. This assessment will guide us in determining the most effective path for applying the predictive CLV approach within CookUnity's operations.

2.3. Predictive CLV - Modeling Approaches

Customer Lifetime Value (CLV) modeling has been a subject of extensive research in marketing and machine learning, aiming to address the fundamental question of identifying valuable customers and allocating resources effectively. Early statistical models of CLV, known as 'Buy 'Til You Die' (BTYD) models, focused on parametric distributions of customer lifetime and purchase frequency, primarily utilizing customer recency and frequency as inputs (RFM). Limited computational resources and data availability constrained these early solutions to simple parametric statistical models.

A comprehensive review done by Gupta et al. (2006) identified six primary types of modeling approaches, including the widely used recency, frequency, and monetary value models (RFM), which relate to the three stages of the customer-company relationship: acquisition, retention, and expansion. While RFM and related models have proven effective, evidence suggests that machine learning prediction models outperform them by incorporating a broader range of variables. In churn prediction competitions, random forest models (Breiman, 2001) have demonstrated strong performance. Further historical insights into CLV prediction can be found in (Kumar & Wiener, 2007) (Dwyer, 1997) (Rosset et al., 2003).

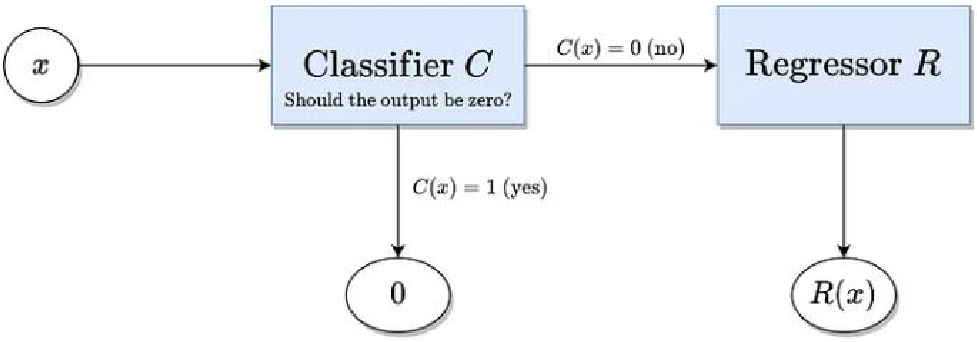
The abundance of customer data available to modern e-commerce companies has posed challenges in incorporating all relevant factors within the RFM and BTYD frameworks. This has led to the emergence of computer science framework approaches, such as optimization (Crowder et al., 2007), support vector machines (SVM) (Zhen-Yu & Peng, 2012), quantile regression (Benoit & Van den Poel, 2012), and portfolio analysis (Čermák, 2015).

In the context of predictive CLV, it is important to consider the potential presence of a zero-inflated problem[2] as the predicted value is the customer's future revenue, which can potentially take the value of zero. When dealing with a zero-inflated problem, it becomes crucial to employ modeling techniques that can effectively account for and handle the excess zeros. Failing to address this issue can lead to biased or inaccurate predictions and interpretations.

[2] A zero-inflated problem refers to a statistical phenomenon where the data contains an excessive number of zero values that cannot be explained by a typical distribution.

A novel approach proposed by Vanderveld et al. (2016) aimed to incorporate user engagement features and swiftly detect changes in customer value in an e-commerce business. They utilized a two-stage random forest model for different user behavior segments, employing a comprehensive feature set. Their solution to the zero-inflated problem was tackled by the two-stage modeling where the first stage predicted purchase propensity, while the second stage predicted the dollar value for users identified as potential purchasers in the first stage as shown in figure 9.

Figure 9: Two-stage model proposed by Vanderveld et al. (2016)



Although the two-stage approach provides insights into the various factors influencing CLV, **it** introduces complexity by maintaining two models. Chamberlain et al. (2017) explored a deep neural network (DNN) solution and found that, with a sufficient number of hidden units, it achieved comparable performance to the random forest approach proposed by Vanderveld et al. (2016).

These various modeling approaches highlight the evolution of CLV prediction methods and the ongoing exploration of more sophisticated techniques to capture the intricacies of customer behavior and value. In a further section, we will leverage methodologies employed in previous studies to evaluate the most suitable approach for applying predictive CLV within the context of CookUnity.

2.3.1. Target Variable

The selection of an appropriate target variable is crucial in developing a successful CLV prediction model. One common issue faced in CLV modeling is the infeasibility of predicting longer-term CLV due to the length of historical data required to construct training labels, and the waiting time needed to evaluate the model's predictive performance. Specifically, to model a target variable such as CLV in a 3-year time frame, one would require not only 3 years of representative history but also would need to wait an additional 3 years to test the predictive capacity of the model.

To address this challenge, many researchers have chosen to predict CLV using a shorter time horizon. For instance, Vanderveld et al. (2016) and Chamberlain et al. (2017) both choose to predict CLV with a 1-year prediction horizon. This allows for a more practical implementation of the model while still providing meaningful insights into customer behavior and future profitability.

However, the choice of the prediction horizon must be carefully considered, as it can impact the model's accuracy and usefulness in a particular business context. Companies with high customer turnover or operating in rapidly changing industries may benefit from a shorter prediction horizon. On the other hand, companies with a stable customer base and long-term customer relationships may require a longer time horizon to capture the full extent of customer value.

In addition to the prediction horizon, other factors must also be considered when setting the target variable, such as the company's business goals, customer behavior, and data availability. A well-defined target variable, that accurately captures the underlying business problem, is a critical component in developing an effective CLV prediction model.

2.3.2. Predictive Features

As mentioned previously, CLV modeling involves the use of various predictive features to accurately forecast future customer behavior and identify valuable customer segments. Chamberlain et al. (2017) employed 132 handcrafted features highlighting several important categories of predictive features that are commonly used in CLV models.

Purchasing Behavior:

One of the fundamental feature sets used in CLV models is user purchasing behavior. Purchase history metrics such as frequency, recency, and monetary value of purchases are key variables in determining CLV. Features within this segment include the number of days since the most recent purchase, time intervals between purchases, and the number of unredeemed vouchers a user currently holds.

Engagement - Web/App Session Logs:

Engagement features provide insights into user interactions and their level of engagement with the product. Tracking front-end event data from mobile or web-app interactions, as well as interactions with communication channels like emails, SMS, and in-app notifications, helps detect changes in customer value earlier than relying solely on purchase history data.

User Experience:

User experience is another crucial feature set in CLV models. It encompasses two main aspects. Firstly, the product offering and the level of availability to each customer can influence their future purchasing behavior. Secondly, customer service plays a vital role in shaping the user experience. Tracking metrics such as the number of refunds and customer service tickets, customer service phone and email wait times, and survey responses provide valuable insights into customer satisfaction and future behavior. For businesses dealing with physical merchandise, average shipping times can also be an important user experience feature to consider.

Customer Demographics:

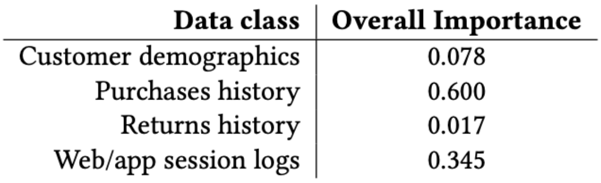
Basic demographic variables such as gender, age, location (including city size and distance between home and city center), and circumstances related to a user's original subscription and first purchase (including cohort years) are commonly utilized in CLV models. These demographic features help in understanding customer characteristics and their impact on future behavior.

2.3.2.1. Feature Importance

It is worth noting that the importance and weights of predictive features can vary across different customer segments. Vanderveld et al. (2016) evaluated their model on each purchase cohort, knowing that model performance and feature importance will vary for each group of users.

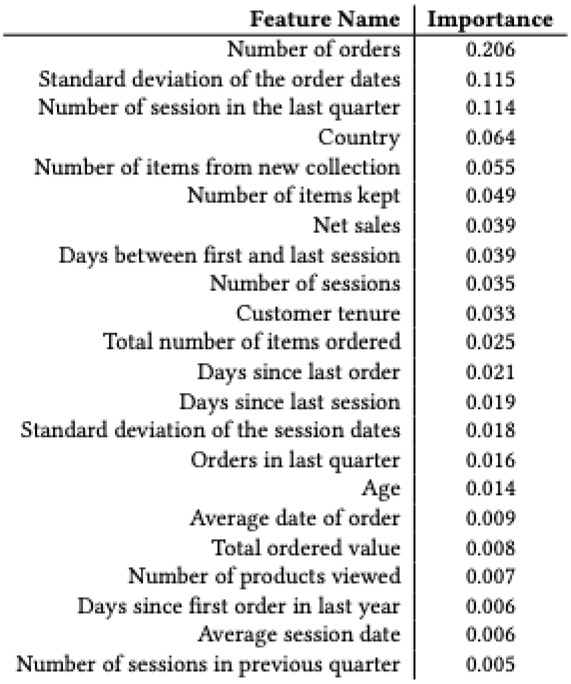
In their study, Chamberlain et al. (2017) were able to determine feature weights based on their model. By analyzing the weights assigned to each feature, they identified the key drivers of CLV within their business context. These findings enabled them to focus their marketing and retention strategies on the most influential factors.

Table 2: Chamberlain et al. (2017) Feature importance by data class.



Data class	Overall Importance
Customer demographics	0.078
Purchases history	0.600
Returns history	0.017
Web/app session logs	0.345

Table 3: Chamberlain et al. (2017) Individual feature importance.



Feature Name	Importance
Number of orders	0.206
Standard deviation of the order dates	0.115
Number of session in the last quarter	0.114
Country	0.064
Number of items from new collection	0.055
Number of items kept	0.049
Net sales	0.039
Days between first and last session	0.039
Number of sessions	0.035
Customer tenure	0.033
Total number of items ordered	0.025
Days since last order	0.021
Days since last session	0.019
Standard deviation of the session dates	0.018
Orders in last quarter	0.016
Age	0.014
Average date of order	0.009
Total ordered value	0.008
Number of products viewed	0.007
Days since first order in last year	0.006
Average session date	0.006
Number of sessions in previous quarter	0.005

2.3.3. Cross-Validation Techniques

Though the studied resources did not deep dive into the validation strategies used, considering that CookUnity sales are significantly affected by seasonality, it is crucial to employ a time-awareness modeling approach. Time-aware modeling is a technique utilized in machine learning to account for the time dimension when constructing models that make predictions based on time-series data (Shmueli & Lichtendahl, 2016).

When training data includes a temporal aspect, such as customer transactions recorded over time, it is important to consider the chronological order of the data during model evaluation (Hyndman & Athanasopoulos, 2018). This mimics real-world scenarios where the model is trained on historical data and tested on future data. Time series cross-validation is specifically designed for datasets with a temporal aspect, where the order of data points matters. In time series cross-validation, it is crucial to maintain the temporal ordering of the data to avoid introducing data leakage. The validation set should always come after the training set in terms of time.

Contrasting this with a simpler form of cross-validation, where data is randomly partitioned into training and validation folds (Figure 10), this approach acknowledges that time-series data often exhibit patterns and seasonality that change over time. Models trained on data from one period may not perform well on data from another period. By adopting a time-aware approach (Figure 11), the model can learn to adjust its parameters to the given seasonality and changes in patterns over time, leading to more accurate predictions for future periods.

Figure 10: Cross-Validation Example

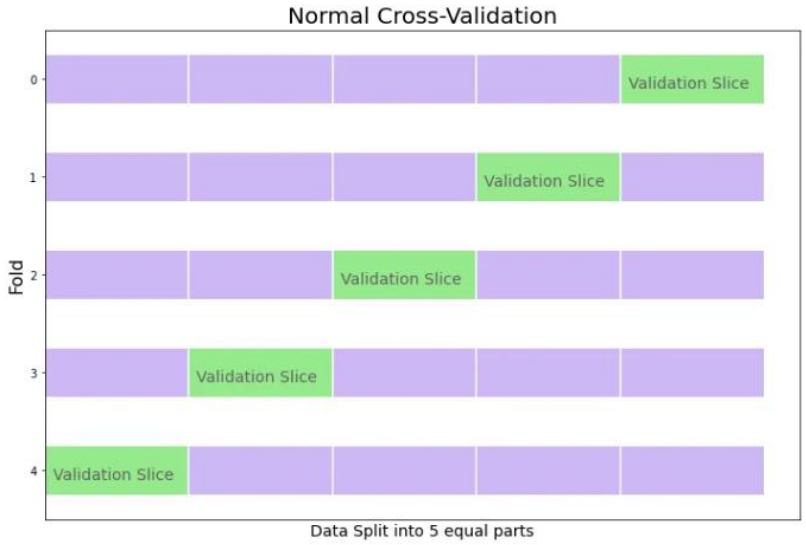
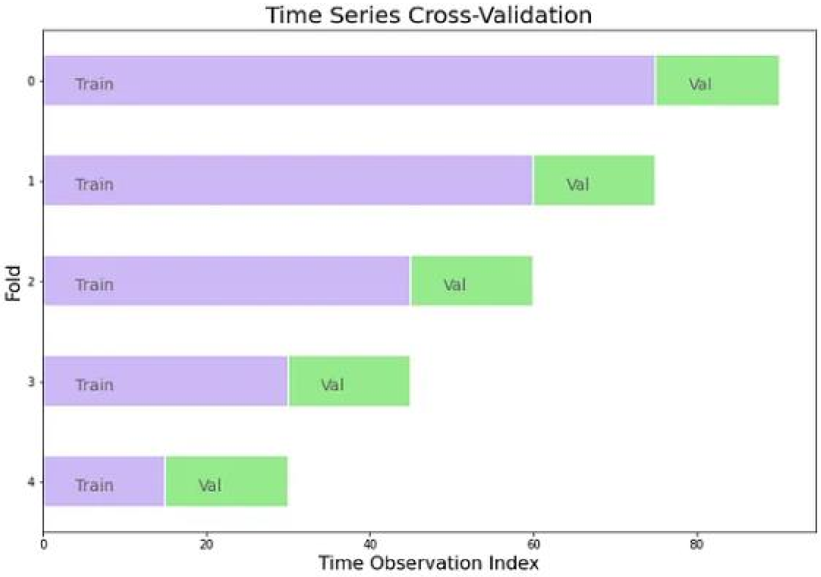


Figure 11: Time Series Cross-Validation Example



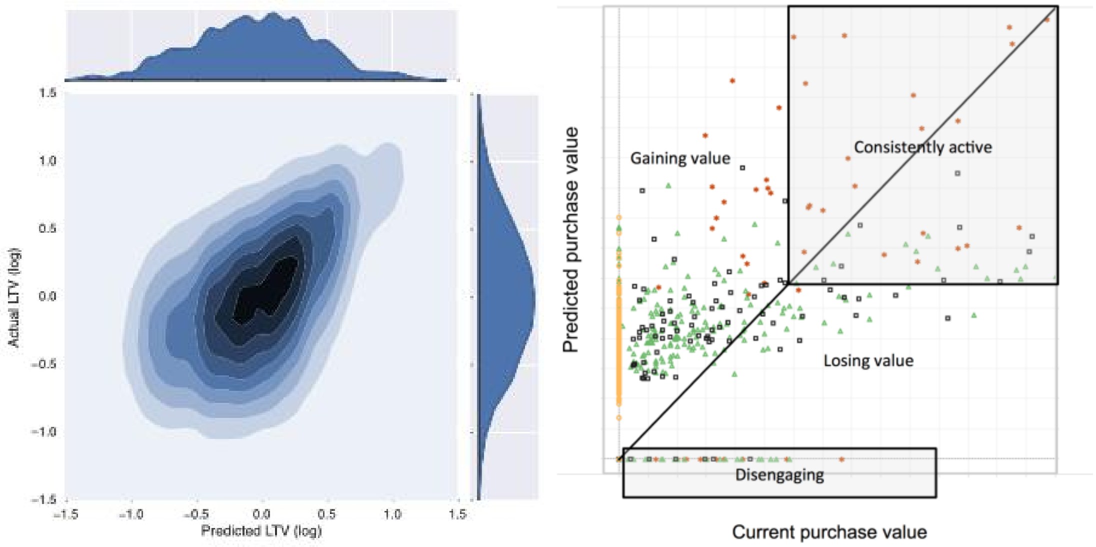
The significance of employing this time-aware modeling and time series cross-validation is that it provides valuable insights into the model's ability to capture seasonality and adapt to changing patterns. By evaluating performance on unseen data, we can confidently assess the model's generalizability and make informed decisions based on its predictions.

In the subsequent sections, we will delve into the specific implementation of our time-aware modeling approach and describe the steps taken to execute the time series cross-validation, ensuring the reliability and effectiveness of our CLV prediction models.

2.3.4. Model Evaluation and Performance Metric

Regarding model evaluation, as the target variable represents an amount, the CLV prediction problem can be treated as a regression problem. Both Vanderveld et al. (2016) and Chamberlain et al. (2017) evaluated the performance of their models by measuring the Root Mean Squared Error (RMSE) and comparing the actual versus predicted distributions (see Figure 12). Additionally, other metrics such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) can also be employed to assess the model's accuracy. These metrics provide insights into the magnitude of prediction errors and the proportion of error relative to the actual values. Furthermore, the coefficient of determination (R-squared) can be used to evaluate the model's ability to explain the variance in the target variable, offering a measure of overall performance.

Figure 12: Vanderveld et al. (2016) and Chamberlain et al. (2017) CLV Predicted vs Actual



3. Methodology

Building upon the literature review presented earlier, this section outlines the methodology employed to achieve the main objective of this work: predicting customer lifetime value (CLV) at the individual level. The methodology description encompasses critical aspects, including the definition of the target variable, identification of available data sources, model selection, and rigorous evaluation.

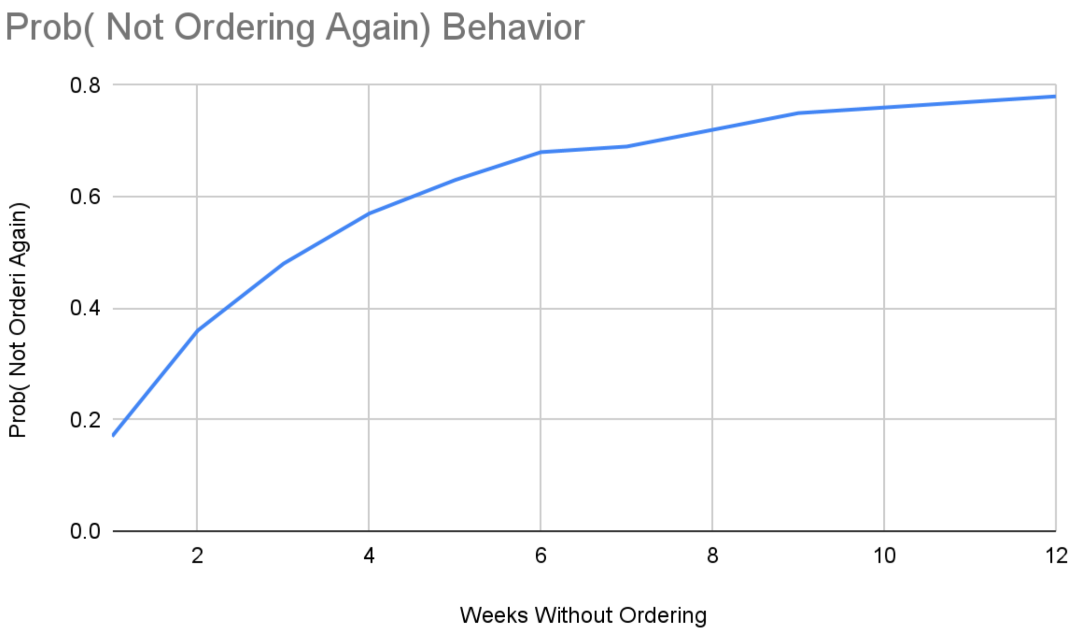
Through this methodology, our objective is to develop a robust and reliable CLV prediction model that empowers CookUnity to make data-driven decisions, optimize customer relationships, and enhance long-term profitability.

3.1. Target Variable

To define the target variable for CookUnity's CLV model, several considerations need to be taken into account. As the company is developing its initial CLV model, it is essential to prioritize agility in learning and iteration. Therefore, a shorter time window prediction is deemed more suitable.

One approach to determining the probability of a customer not ordering again within a specific timeframe is by analyzing the ordering cadence. Observing customer behavior, it is observed that the probability of not ordering again goes above 60% after approximately five weeks of inactivity, as indicated in the figure below. Identifying customers who are generating low CLV in the subsequent five weeks becomes crucial for maximizing customer ordering retention and overall CLV.

Figure 13: Probability of Not Ordering Over N Weeks Without Ordering



As mentioned in section 1.2, CookUnity currently works with an overall contribution margin percentage (CM%) for all orders, therefore, it is equally feasible to work with either CLV or average revenue per user (ARPU) from a mathematical perspective.

Based on the factors mentioned above, we have decided that the target variable for our CLV model at CookUnity will be set as the Net Revenue in the next 28 days for each customer. This time window was chosen as it provides a balance between providing a meaningful prediction and allowing for agility in learning and iteration. By predicting the net revenue, we can better understand which customers are more likely to generate low CLV in the near future, allowing CookUnity to take appropriate actions to retain them. Additionally, since the company currently works with an overall CM% for all orders, using Net Revenue in the next 28 days as the target variable is equivalent to using CLV for mathematical purposes.

3.2. Available Data Sources

Building accurate predictive features for the target variable necessitates considering a wide range of available data sources that can offer valuable insights into customer behavior and characteristics. In this section, we will discuss the various cleaned and modeled datasets provided by CookUnity, that will be leveraged to create the feature set for our CLV prediction model.

CookUnity has a wealth of data at its disposal (see Appendix B), encompassing various aspects of customer interactions and transactions. Leveraging these multiple data sources, we can derive meaningful predictors that enhance the accuracy and effectiveness of our CLV model. The data sources utilized include:

Purchasing Behavior: Models that encompass information about customers' purchasing habits and experience.

- **fact_events:** Contains information on customer ordering behavior for each week (If ordered, skipped, pause or unsubscribed).
- **fact_orders:** Provides comprehensive order information (creation date, shipping, revenue).
- **fact_order_line_items:** Contains details about the items ordered (id of the items, and quantities).
- **fact_recommendations:** Captures information on recommended orders (which items were recommended).
- **fact_available_products:** Contains information on products being offered in the menu each day.

Customer Satisfaction:

- **fact_reviews:** Captures customer reviews placed on the ordered items.
- **fact_nps_survey_response:** Captures information on NPS survey responses.
- **fact_menu_satisfaction_surveys:** Contains data on menu satisfaction responses.
- **bronze_product_added_to_favorites:** Contains tracking of products added to favorites.
- **fact_customer_snapshots_referrals:** Contains information on customers referrals.

Customer Service:

- **fact_ops_logistics_orders:** Contains information on order shipments.
- **fact_ops_failed_orders:** Contains information on operational issues.
- **fact_credits:** Contains information on compensation credits given.
- **fact_refunds:** Provides details on refunded orders**.**

Customer Profile:

- **dim_customers:** Provides details about customer dimensions.
- **dim_customer_acquisition_sources:** Contains information on customer marketing acquisition sources.
- **fact_subscriptions:** Provides insights into customer subscriptions.
- **silver_dim_user_preferences:** Contains information on customer meal preferences settings.

Engagement:

- **fact_menu_browse:** Contains tracking of customers browsing the menu.
- **fact_customer_snapshots_sessions:** Contains tracking of customers sessions starts.
- **fact_communications:** Contains information on communications being sent to customers.
- **fact_promotions_activations:** Contains data on promotions generated at customer level.
- **fact_emails:** Contains data on email conversion.

One of the key challenges in this project is to integrate and consolidate data from these various sources into a unified dataset that can be used in our CLV model. This integration process will be accomplished through dbt[3] by combining the described datasets into a single comprehensive table, referred to as One Big Table (OBT).

In the subsequent sections, we will delve into the specifics of each data source and discuss their respective contributions to the feature engineering process. Additionally, we will explore the data preprocessing techniques employed to ensure data quality, consistency, and relevance in our CLV prediction model.

3.3. Time Series Cross-Validation

Ensuring accurate and robust predictions within CookUnity's subscription-based business, which experiences strong seasonality in customer behavior, necessitates the implementation of a comprehensive cross-validation technique. In this section, we will outline the time series cross-validation strategy that will be employed for evaluation.

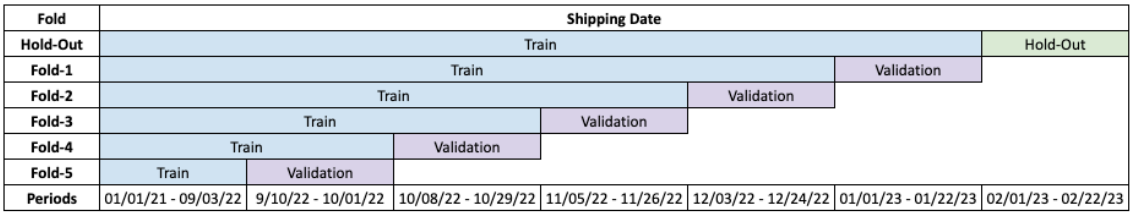
Adopting a time-aware modeling approach involves dividing our backtests into three distinct periods: training, validation, and hold-out. During the training period, the model learns from historical data up until a specific point in time, allowing it to capture temporal dynamics. The validation period is dedicated to fine-tuning the model's hyperparameters using out-of-sample data, while the hold-out period assesses the model's performance on unseen data, providing a reliable measure of its predictive capabilities (Stoffer & Shumway, 2006).

Given that our primary target variable is revenue, and orders are invoiced upon preparation and shipment, we designate the shipping date as the time dimension for our cross-validation framework. This ensures that the evaluation is conducted based on the actual timeline of order fulfillment. Also, considering that our target variable has a time-frame prediction target of 28 days, the length of each validation period was set to the same length.

[3] dbt (getdbt.com) is an open-source command line tool that helps analysts and engineers transform data in their warehouse more effectively.

In line with reviewed Figure 11, Table 3 showcases how our data set cross-validation folds settings looks like.

Table 4: Time Series Cross-Validation Folds



<table>
<tr><th>Fold</th><th colspan="7">Shipping Date</th></tr>
<tr><td>Hold-Out</td><td colspan="6">Train</td><td>Hold-Out</td></tr>
<tr><td>Fold-1</td><td colspan="5">Train</td><td>Validation</td><td></td></tr>
<tr><td>Fold-2</td><td colspan="4">Train</td><td>Validation</td><td></td><td></td></tr>
<tr><td>Fold-3</td><td colspan="3">Train</td><td>Validation</td><td></td><td></td><td></td></tr>
<tr><td>Fold-4</td><td colspan="2">Train</td><td>Validation</td><td></td><td></td><td></td><td></td></tr>
<tr><td>Fold-5</td><td>Train</td><td>Validation</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Periods</td><td>01/01/21 - 09/03/22</td><td>9/10/22 - 10/01/22</td><td>10/08/22 - 10/29/22</td><td>11/05/22 - 11/26/22</td><td>12/03/22 - 12/24/22</td><td>01/01/23 - 01/22/23</td><td>02/01/23 - 02/22/23</td></tr>
</table>

3.4. Modeling Approaches to Evaluate

In order to effectively leverage the extensive amount of available data for CLV prediction, we need to go beyond parametric models and explore machine learning solutions. Our objective is to determine which of the studied solutions performs better for our CLV prediction problem and aligns with our business requirements.

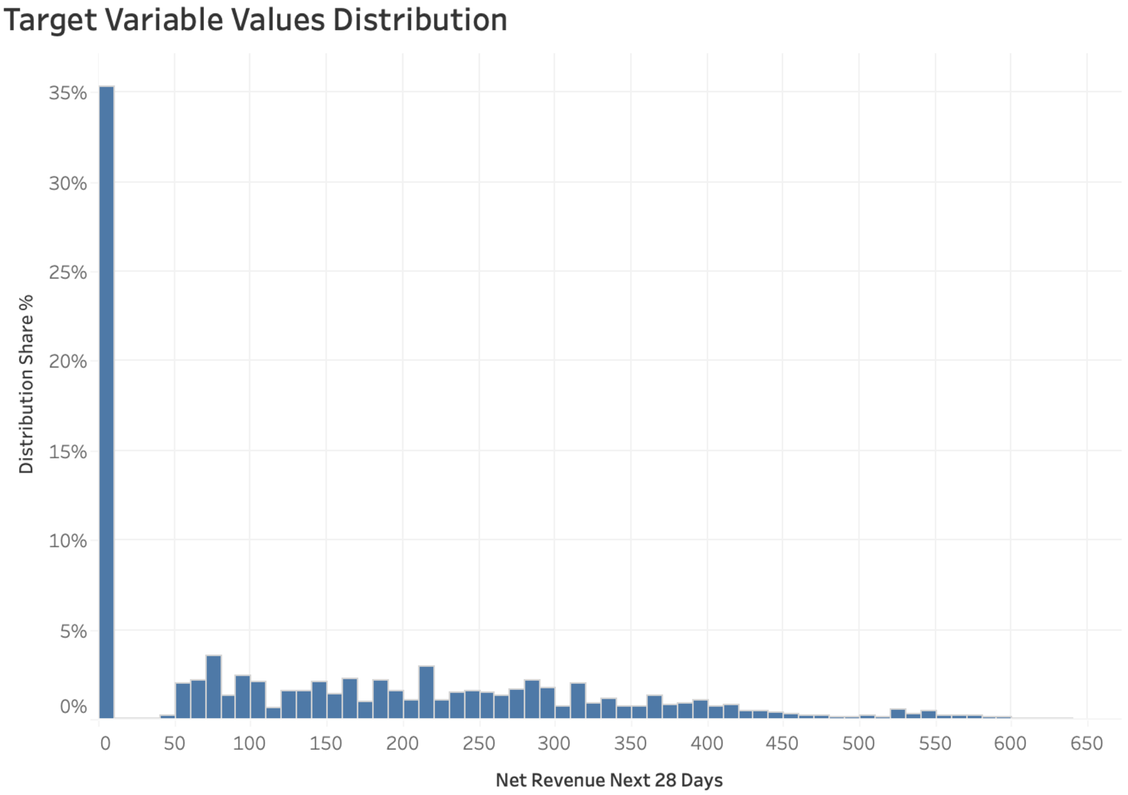
First, we will delve into gradient boosting machines (GBM), a technique similar in concept to the random forests methodology employed by Vanderveld et al. (2016). GBM fits individual decision trees to random re-samples of the input data, with each tree trained on a bootstrap sample of the dataset's rows and N arbitrarily chosen columns, where N is a configurable parameter. However, GBMs differ from random forests in a significant aspect: each successive tree is fitted to the residual errors from all the previous trees combined. This approach is advantageous as it focuses each iteration on the examples that are most challenging to predict, ultimately improving the accuracy of the model.

Furthermore, we will also explore neural network solutions, similar to the approach proposed by Chamberlain et al. (2017). By considering neural networks, we aim to compare their performance against GBMs and determine which approach better aligns with our prediction problem. Neural networks offer a different modeling perspective and possess the capability to capture complex patterns and relationships within the data.

It is important to note that both of these reviewed methodologies deal with predicting CLV in the context of retailing e-commerce. Given that most customers do not place orders, they tackle a zero-inflated problem by building a two-stage solution: one stage predicts whether a customer is going to purchase, and the second stage predicts the monetary value of those predicted to make a purchase in the first stage.

Although our problem is similar, it is not exactly the same as CU operates as a subscription-based business. This provides a better scenario compared to the one faced by Chamberlain et al. (2017). However, since CookUnity is a subscription model with a monthly order rate of 70%, most users do place an order, and those orders have a fixed value determined by the subscription plan (these can be assessed from the peaks along the tail of the distribution).

Figure 14: Net Revenue Next 28 Days Distribution



Based on these factors, we aim to build a single regression solution without the need for an intermediate classification (order/not order). Our goal is to select the modeling approach that best suits our prediction problem and exhibits superior performance in predicting CLV.

3.5. Model Validation

The performance of the evaluated models will be measured using Root Mean Squared Error (RMSE), a widely recognized metric in the literature for regression problems like CLV prediction. RMSE is chosen due to its ease of interpretation and its ability to provide meaningful insights into the magnitude of prediction errors. By taking the square root of the average squared differences between the actual and predicted values, RMSE gives a measure of the typical size of errors made by the model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (yi - \hat{y})2}{N}} \tag{6}$$

Where:

- n is the number of samples or data points.
- y represents the actual or observed values of the target variable.
- ŷ represents the predicted values of the target variable.

Moreover, RMSE is particularly useful in the context of CLV prediction as it penalizes larger errors more heavily. Since the target variable represents an amount, it is crucial to accurately estimate and minimize the deviations between the predicted and actual values. RMSE emphasizes larger errors, making it a suitable choice for assessing the performance of CLV prediction models.

While RMSE will serve as the primary evaluation metric, we acknowledge the importance of considering other performance metrics as well. Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) will also be employed to provide additional insights into the models' accuracy and the proportion of error relative to the actual values. Furthermore, the coefficient of determination (R-squared) will be used to evaluate the model's ability to explain the variance in the target variable, offering a measure of overall performance.

By considering multiple evaluation metrics, we aim to gain a comprehensive understanding of the models' predictive capabilities and their suitability for accurately estimating customer lifetime value. This holistic evaluation approach will enable us to make informed decisions and effectively communicate the performance of the models to stakeholders.

4. Development

In this section, we present the development of our CLV prediction model, encompassing a comprehensive exploration of the key stages involved in constructing a robust and effective solution. We commence by providing a detailed description of the data utilized in our analysis and its relevance. Subsequently, we delve into the process of feature engineering, carefully selecting and transforming variables that exhibit the highest potential for predicting CLV accurately. Furthermore, we undertake the model selection process, comparing the machine learning algorithms reviewed in our literature analysis to identify the most suitable approach that yields the highest precision in CLV predictions. Additionally, we discuss the evaluation metrics employed to assess the performance of our models, thereby enhancing our understanding of their applicability.

4.1. Exploratory Data Analysis (EDA)

In accordance with the previously outlined methodology, our analysis leverages an extensive range of data sources, presenting a vast pool of information for constructing a comprehensive set of predictors for our CLV model. It is noteworthy that while CookUnity has not previously engaged in CLV prediction or ARPU estimation, the company has conducted studies to identify user segments and behaviors that exhibit distinct CLV patterns.

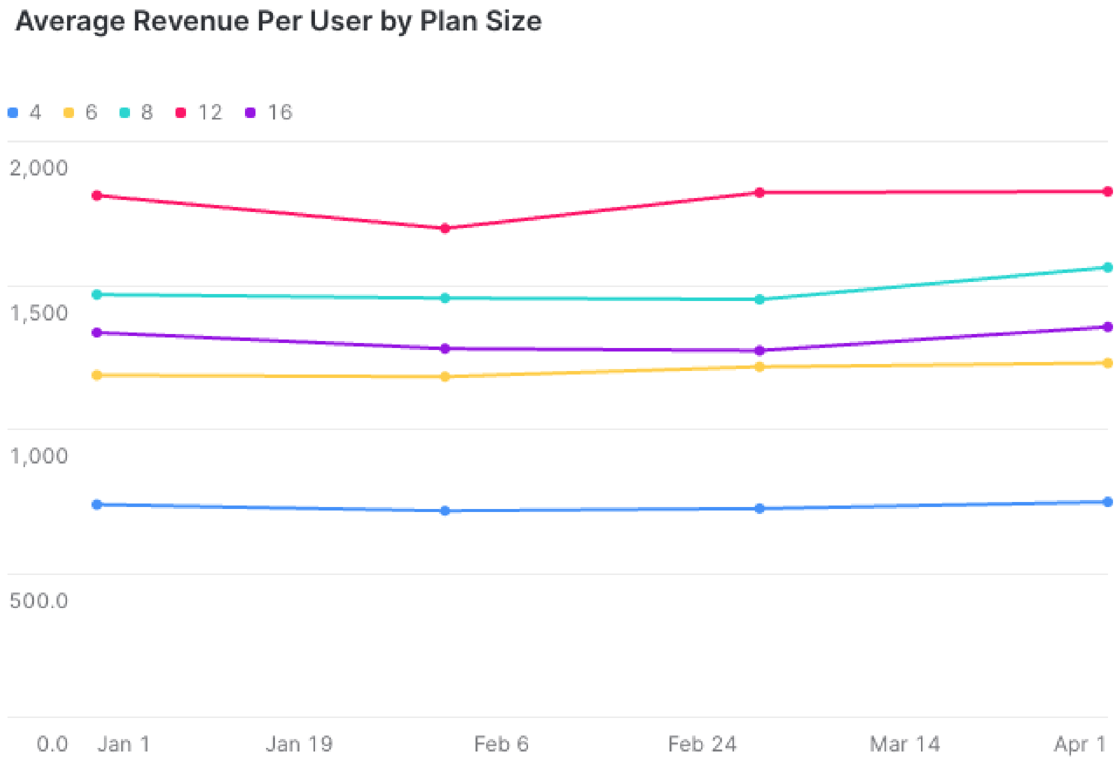
In this section, we embark on a meticulous exploration of these significant customer behaviors, which warrant inclusion as predictive features. This exploration serves the purpose of facilitating the creation of a robust and representative training feature set. It is important to acknowledge that providing an exhaustive enumeration of all the features developed for this project would exceed the scope of this section. However, a comprehensive listing of these features can be found in Appendix C, offering an overview of the incorporated predictors.

4.1.1. Subscription Plan

Upon subscribing to CookUnity (CU), customers are presented with a range of subscription plans from which to choose, (see Appendix A). The selection of a particular plan is driven by individual needs and preferences. Given that our target variable is revenue-based, it naturally follows that the monetary value of each order is determined by the specific subscription plan a customer has enrolled in. Consequently, customers who have subscribed to higher-tier plans are expected to incur greater expenses.

However, a preliminary assumption that customers with higher subscription plans exhibit higher Customer Lifetime Value (CLV) does not hold universally true. Figure 15 illustrates the evolution of Average Revenue per User (ARPU) across different subscription plans, revealing that customers on the sixteen-meal plan perform worse than those on the twelve-meal plan. This unexpected observation can be attributed to the fact that some customers who receive sixteen meals per week may not consume all of them, resulting in a lower likelihood of placing subsequent orders. In other words, although the order value associated with the sixteen-meal plan is higher, customers on this plan tend to place fewer orders compared to those on lower-tier plans, resulting in a lower ARPU.

Figure 15: ARPU Performance Evolution by Plan Size (Q1 - 2023)



In addition to the impact of subscription plans on CLV, it is crucial to consider the dynamic nature of customers' plan selections. Customers have the flexibility to switch their plans from week to week. Hence, it is not only important to ascertain the most recent plan chosen by a customer, but also to examine the frequency and directionality of plan changes. Understanding whether customers are increasing or decreasing the number of meals in their plans provides valuable insights into their engagement level with the service.

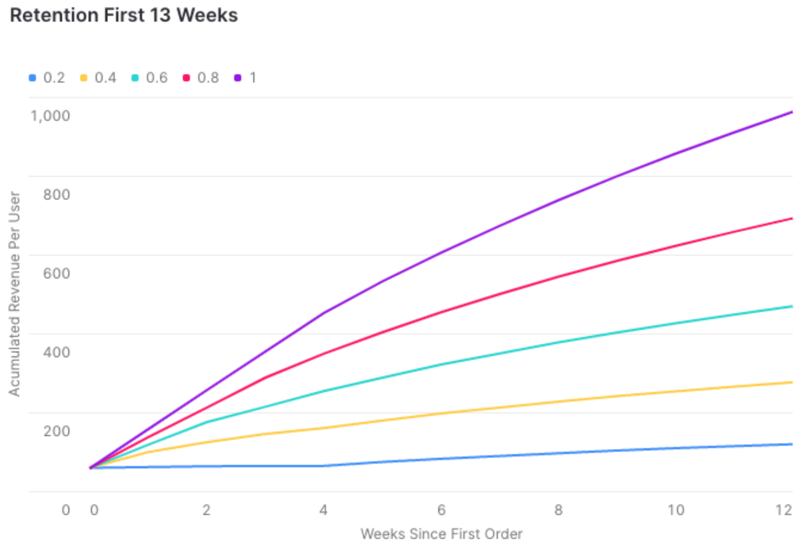
4.1.2. Building Habit

As highlighted during the definition of our target variable, establishing a sense of habit among CookUnity users is of paramount importance. It has been observed that once a customer initiates an order, the likelihood of reordering significantly increases, fostering the development of an ordering cadence.

We can gain further insights into Customer Lifetime Value (CLV) by examining it from a cohort perspective, as suggested by the alternative methodology reviewed in the literature. By considering all users who remain subscribed for a minimum of five weeks, we can segment them based on the frequency of their orders within the first five weeks (referred to as "Order Rate"). Figure 16 showcases the revenue growth trajectory of these cohorts, revealing that customers who exhibit a higher Order Rate (Order Rate >= 0.8) continue to generate increasing revenue at a steeper rate compared to those with lower order frequency.

This analysis emphasizes the significance of fostering a consistent ordering habit among customers. The sustained ordering behavior not only maintains customers in their established ordering cadence but also contributes to generating a higher CLV over time.

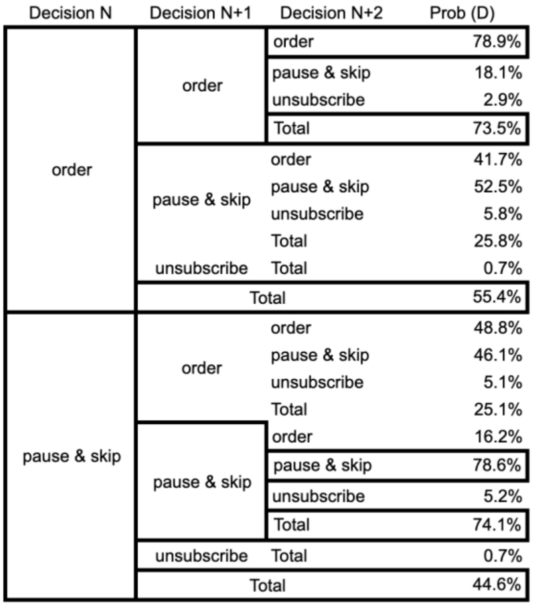
Figure 16: ARPU Growth by Order Segment at W5



A complementary perspective on this behavior can be obtained by examining the probability of repeated ordering, as presented in Table 3. On average, in any given week, there is a 55.4% probability of a customer placing an order. Among those who place an order, the probability of reordering the following week is 73.5%, and this likelihood further rises to 78.9% for subsequent orders. Similarly, in any given week, there is a 44.6% probability of a customer skipping the order, leading to a 74.1% probability of skipping again the following week, which further increases to 78.6%. These conversion probabilities indicate that once a customer falls into the habit of either ordering or skipping, the likelihood of changing that behavior becomes relatively low.

Hence, it becomes imperative for the company to actively encourage customers to develop an ordering habit. By facilitating and nurturing this habit formation, CookUnity can mitigate the chances of customers deviating from the established behavior of consistent ordering or pausing, ultimately fostering long-term customer engagement and maximizing CLV.

Table 5: Next Decision Probability



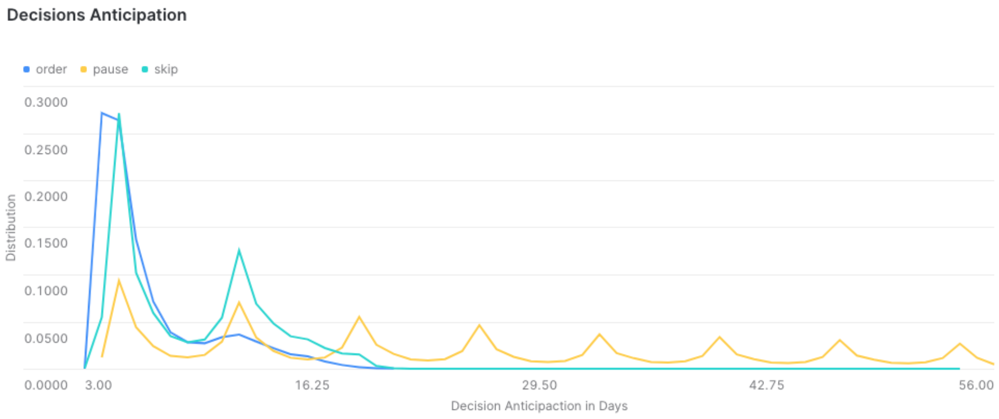
Decision N	Decision N+1	Decision N+2	Prob (D)
order	order	order	78.9%
		pause & skip	18.1%
		unsubscribe	2.9%
		Total	73.5%
	pause & skip	order	41.7%
		pause & skip	52.5%
		unsubscribe	5.8%
		Total	25.8%
	unsubscribe	Total	0.7%
	Total		55.4%
pause & skip	order	order	48.8%
		pause & skip	46.1%
		unsubscribe	5.1%
		Total	25.1%
	pause & skip	order	16.2%
		pause & skip	78.6%
		unsubscribe	5.2%
		Total	74.1%
	unsubscribe	Total	0.7%
	Total		44.6%

4.1.3. Decision anticipation

It is important to recognize that despite being a subscription-based service, CookUnity allows customers the flexibility to skip or pause their subscriptions for a certain period. However, this decision must be made four days prior to the scheduled delivery. For instance, if the delivery day is on Friday, customers are required to place their order or skip it by Monday to avoid an autopilot order being generated automatically.

The ability to measure the number of days customers anticipate their ordering decisions holds significant value within the CookUnity business model. Figure 17 illustrates the distinct distributions of decision anticipation in days for different customer behaviors.

Figure 17: Decision Anticipation Distribution by Days



Notably, a large proportion of orders are placed within a narrow timeframe, typically four to six days ahead of the delivery date. In contrast, when customers choose to skip an order, they tend to do so either before the autopilot order is generated or during the week preceding the delivery. On the other hand, when customers pause their subscription, subsequent orders are paused several weeks in advance, displaying a relatively even distribution.

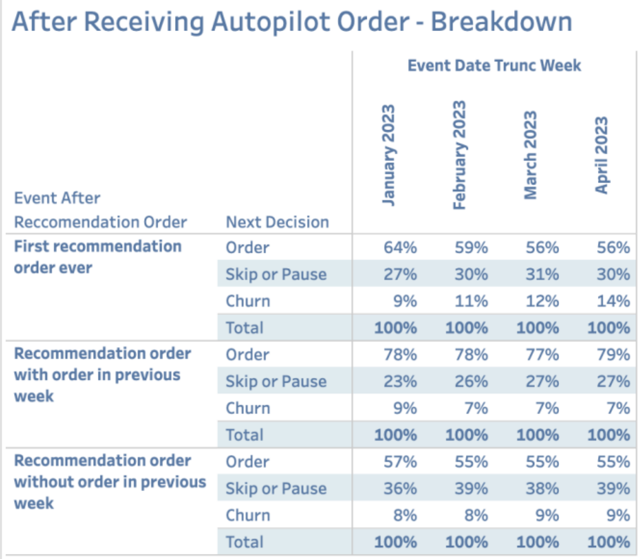
Considering the average anticipation period for each customer in determining whether to place an order or not becomes crucial in understanding their behavioral patterns and preferences. By leveraging this information, CookUnity can enhance its understanding of customer decision-making processes and tailor its strategies to optimize customer engagement and satisfaction.

4.1.4. Autopilot Recommendation

The autopilot order feature plays a significant role in CookUnity's subscription business model. When customers neither create an order nor choose to skip or pause their subscription, CookUnity's recommendation algorithm generates an order on their behalf. While some customers appreciate this feature and continue to utilize it after receiving an autopilot order, others perceive it as an unwanted service, leading to disengagement from the subscription.

Table 4 demonstrates the behavioral patterns associated with the autopilot order feature. When customers receive their first-ever recommendation, the probability of placing an order again is approximately 60%, while the likelihood of unsubscribing reaches 15%. Similar patterns emerge when customers have been skipping or pausing their orders, and suddenly receive a recommendation. The ordering probability stands at 50%, while the churn rate remains high at 10%. In contrast, customers who have been regularly placing orders and receive a recommendation exhibit a positive response. They have an 80% chance of ordering again, and the churn rate decreases to 7%.

Table 6: Next Decision Probability After Receiving Autopilot Order



After Receiving Autopilot Order - Breakdown

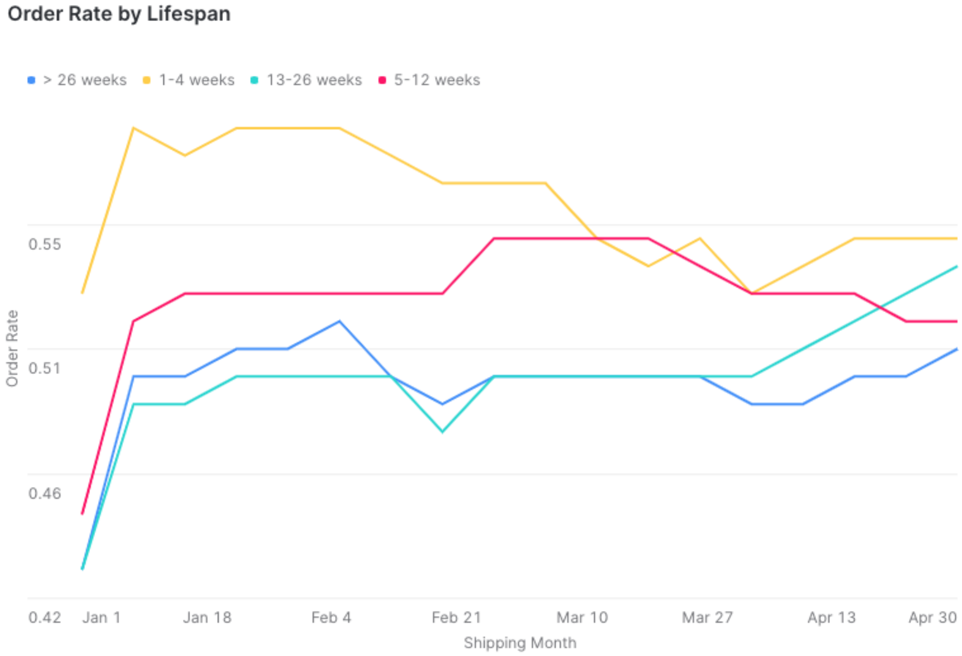
		Event Date Trunc Week			
Event After Reccomendation Order	Next Decision	January 2023	February 2023	March 2023	April 2023
First recommendation order ever	Order	64%	59%	56%	56%
	Skip or Pause	27%	30%	31%	30%
	Churn	9%	11%	12%	14%
	Total	**100%**	**100%**	**100%**	**100%**
Recommendation order with order in previous week	Order	78%	78%	77%	79%
	Skip or Pause	23%	26%	27%	27%
	Churn	9%	7%	7%	7%
	Total	**100%**	**100%**	**100%**	**100%**
Recommendation order without order in previous week	Order	57%	55%	55%	55%
	Skip or Pause	36%	39%	38%	39%
	Churn	8%	8%	9%	9%
	Total	**100%**	**100%**	**100%**	**100%**

Given the varying user interactions with the recommendation feature offered by CookUnity, understanding and incorporating this aspect within our feature set becomes essential.

4.1.5. Customer Lifespan

In line with typical subscription business dynamics, CookUnity observes distinct patterns in user behavior based on their lifespan or duration of subscription. New users, who initially join the service as trials, tend to exhibit different engagement levels compared to more established users. Figure 18 illustrates the ordering levels of users categorized by their lifespan.

Figure 18: Order Rate by Lifespan (Q1 - 2023)



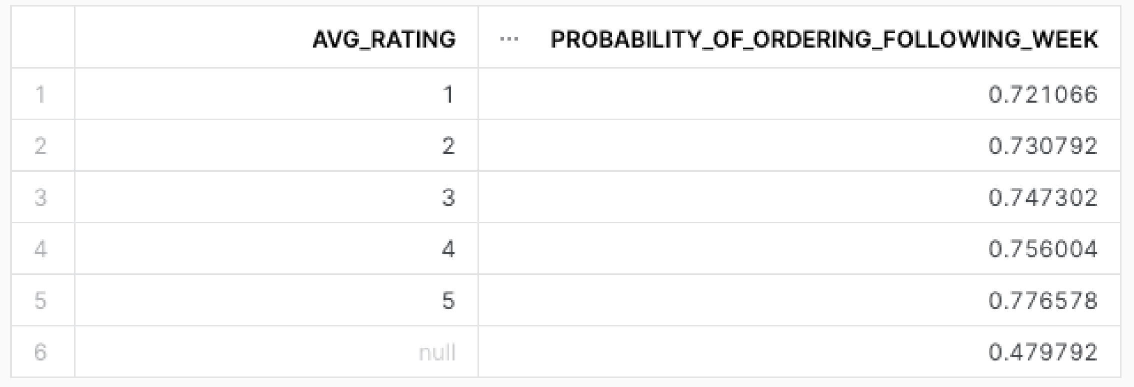
It is evident that as users progress and become more mature in their subscription journey, their ordering frequency tends to decrease. This trend can be attributed to several factors. Initially, when users subscribe, many of them take advantage of early discounts and promotions, leading to a higher volume of orders. However, as their trial period ends and they transition into regular subscribers, their ordering behavior becomes more varied. Some users continue to order on a consistent basis, while others exhibit lower frequency in placing orders.

4.1.6. Customer Satisfaction

Customer satisfaction is a critical aspect of CookUnity's service, and while there are various approaches to assessing satisfaction, we will focus on a few key indicators. These include measuring delivery issues, problems with meals received, and overall satisfaction with the service. Although this discussion will not cover all possible assessment methods, these examples serve as illustrative measures.

Table 5 presents insights into the relationship between customer satisfaction ratings and subsequent ordering behavior. It is evident that higher average ratings correspond to a higher probability of customers placing orders again in the following week. Conversely, a survey without a response significantly reduces the likelihood of a repeat order, even more so than a low rating.

Table 7: Probability of Ordering by Avg. Rating



	AVG_RATING	PROBABILITY_OF_ORDERING_FOLLOWING_WEEK
1	1	0.721066
2	2	0.730792
3	3	0.747302
4	4	0.756004
5	5	0.776578
6	null	0.479792

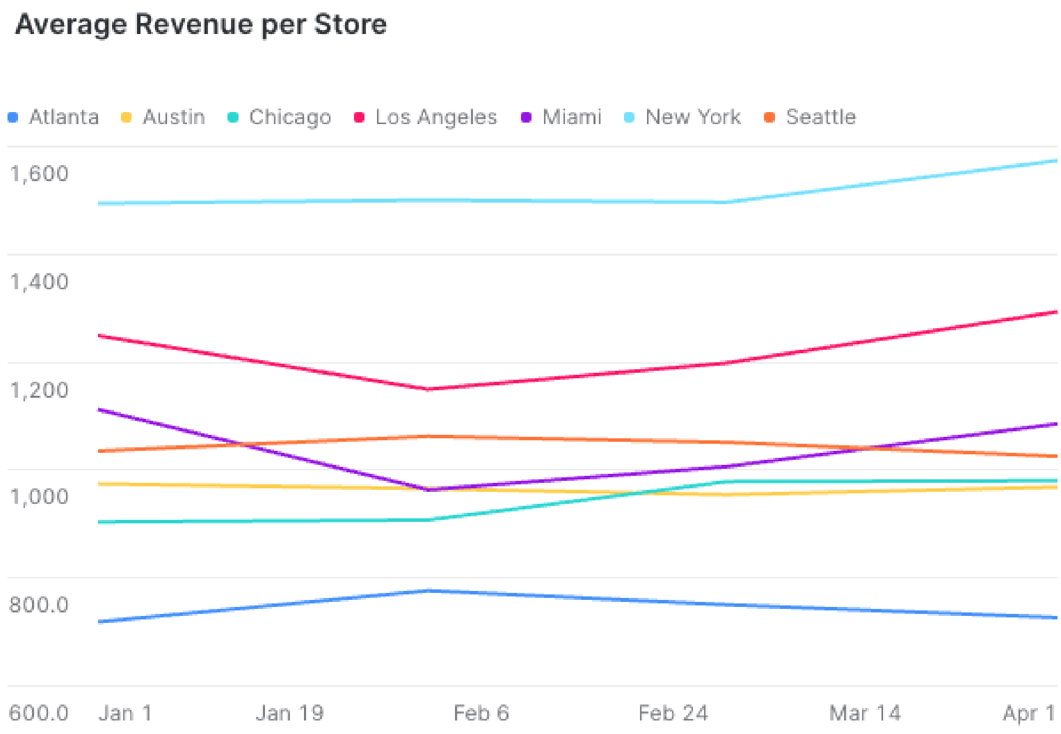
4.1.7. Demographics

The demographic dimension plays a significant role in understanding customer behavior and performance within CookUnity. However, it is important to acknowledge that there is currently a gap in available demographic information such as age, gender, ethnicity, income and others.

Nevertheless, location, as a key demographic factor, provides valuable insights. Each CookUnity kitchen operates in different markets with varying product availability and competition. Additionally, customers in each city have distinct acquisition capabilities, further emphasizing the significance of location.

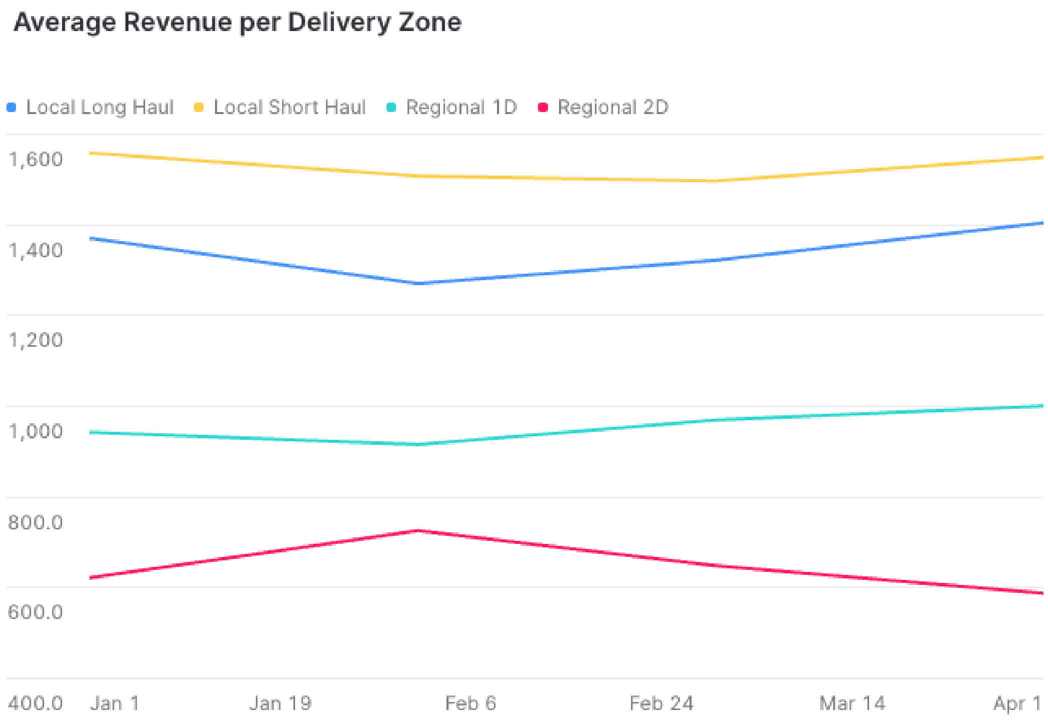
Figure 19 illustrates the relationship between location (represented by the store) and average revenue per user (ARPU). Notably, customers in New York exhibit the highest performance in terms of ARPU, maintaining a consistent lead over customers in Los Angeles. The remaining stores, including AU, CHI, MIA, and SEA, fall within the middle range, while ATL lags behind as the lowest-performing store.

Figure 19: Customers' ARPU by Store (Q1 - 2023)



Furthermore, it is worth noting that when customers subscribe to CookUnity, they are assigned a delivery region based on their shipping address. This delivery setting is closely associated with operational challenges, particularly for regions requiring longer-distance shipping. As depicted in Figure 20, regions with longer distances have lower ARPU due to the logistical complexities involved.

Figure 20: Customers' ARPU by Delivery Zone (Q1 - 2023)



4.1.8. Other essential groups

As previously discussed, the available dataset contains a wealth of information that could be further explored. However, in the interest of maintaining the agility of the analysis process and focusing on the core objective of modeling CLV, we will conclude this exploratory data analysis by highlighting other essential groups of features within the feature set.

- Promotion Offerings: It includes discounts, special deals, and incentives aimed at driving customer engagement and increasing CLV.
- Menu Availability and Meals Preferences: These features capture information related to the availability of different menu options and customers' preferences for specific meals.
- Compensation Credit and Refunds: This group of features involves the tracking of compensation credits and refunds issued to customers.
- Engagement with Communications (Emails): Tracking how customers interact with promotional emails, newsletters, and other communication materials.
- Engagement with Platform, Sessions, and Menu Browsing: These features can shed light on customers' level of engagement and their propensity to make repeat purchases.
- Seasonality Patterns: This group of features examines how customers' ordering habits and preferences change throughout the year, allowing for the incorporation of seasonality factors into the CLV model.

While we have not delved into the details of these feature groups within this section, their inclusion in the feature set acknowledges their importance and potential impact on CLV prediction. A comprehensive listing of all the developed features can be found in Appendix C, providing a detailed overview for further exploration and reference.

4.2. Feature Engineering

The constructed data model necessitates additional engineering to render certain variables amenable to our machine learning models, while simultaneously enhancing prediction performance. Key feature engineering tasks encompass missing values imputation, conversion of categorical variables into numeric representations, and various numeric transformations (Zheng, 2017). This section delves into the implementation of these techniques, elucidating their impact on the dataset and subsequent CLV predictions.

4.2.1. Missing Values Imputation

Missing values imputation is a common preprocessing step in machine learning because many algorithms, including random forest, cannot handle missing values directly. These algorithms rely on numerical computations and mathematical operations, which cannot be performed when missing values are present.

Missing values can lead to incomplete datasets, which can result in biased or inaccurate models. By imputing missing values, you retain the available data and potentially enhance the performance of the algorithm. But if missing values are not handled appropriately, it can introduce bias into the analysis. For example, if certain records have missing values for a specific feature, omitting those records from the analysis can result in biased predictions or inaccurate representations of the underlying patterns.

There are various techniques for missing values imputation, such as mean imputation, median imputation, mode imputation, or more advanced methods like multiple imputation or regression-based imputation (Enders, 2010).

For our specific use case, we have to take in account that all the data sources we reviewed in the EDA are modeled with data validation processes and therefore any missing values we may encounter in the data are meant to be there. Therefore, having a null value in our variables is something to be considered a value on itself and not to be replaced with another such as a median imputation, otherwise we will be creating biased predictions. That's why the numerical missing values imputation strategy for our model is to impute rows of missing values with an arbitrary (default: 9999). This is effective for tree-based models, as they can learn a split between the arbitrary value (9999) and the rest of the data (which will not overlap this value).

4.2.2. Categorical Variables Transformation

Within the training features we have several categorical ones which need to be transformed to numerical for our models to handle them. According to the content of those variables different transformation methods could be applied.

Ordinal encoding and one-hot encoding are both techniques used in machine learning to represent categorical data as numerical values. However, they differ in how they encode the categorical data and the types of data they are best suited for.

Ordinal encoding is a technique that assigns a numerical value to each category in a categorical feature based on the order or rank of the categories. Ordinal encoding is useful when there is an inherent order or ranking between the categories, such as with clothing sizes (S, M, L, XL), education level (High School, Bachelor's, Master's, PhD), or socioeconomic status (Low, Middle, High). This technique is applied to:

- **Delivery Type**: This attribute contains four categories which can be ordered by proximity to the kitchen (shipment origin).
- **Lifespan Segment**: This attribute is a bin grouping of the customer lifespan, therefore it can also be ordered by customer maturity.
- **Previous Shipping Date Event & Last Shipping Date Event**: These attributes correspond to the last and previous ordering decision the customer has made, therefore it can also be ordered as (i) order (ii) skip (iii) pause (iv) churn.

One-hot encoding, on the other hand, is a technique that represents each category as a binary feature. For example, if we have a categorical feature called "Color" with three categories (Red, Green, Blue), we can represent each category as a binary feature by creating three new features: "Is Red", "Is Green", and "Is Blue". One-hot encoding is useful when there is no inherent order or ranking between the categories and when there are many categories. This technique is applied to:

- **Shipping WeekDay**: Day of week customers choose to receive their orders.
- **Acquisition Channel**: Marketing channel where customers subscribed from.
- **Store Name**: Store from which customers receive their meals.

These techniques are the only ones we are going to apply as for now to our categorical variables.

4.3. Models Evaluated

Out of the literature review we found two main similar solutions in this area, one is from Vanderveld et al. (2016) who implemented a tree-based solution and the other was from Chamberlain et al. (2017) who motivated by Vanderveld et al. (2016) solution tackled a similar problem with DNN. This last achieved similar performance, therefore we will try which of both achieves a better on our training data.

As mentioned in the methodology for tree-based decision models we will be implementing a random forest and two gradient boosting machine solutions (XGBoost and LightGBM), and for the DNN solution we will be working with a neural network regressor. All of these models will be working with the same data preprocessing reviewed in previous sections and be calibrated with the mentioned cross-validation strategy.

4.3.1. Random Forest

Random Forest is an ensemble learning algorithm that combines the predictions of multiple decision trees to make accurate and robust predictions. Each decision tree in the Random Forest is trained independently on a random subset of the training data and features, reducing the risk of overfitting. The final prediction is made by aggregating the predictions of all the trees, either through majority voting for classification or averaging for regression. Random Forest is known for its simplicity, ability to handle missing values, and providing feature importance measures. It is widely used due to its robustness, good performance, and suitability for various domains. However, the interpretability of Random Forest may be lower compared to a single decision tree due to its ensemble nature (Breiman, 2001).

4.3.2. XG Boost

XGBoost, short for Extreme Gradient Boosting, is a powerful machine learning algorithm that excels in tackling a wide range of predictive modeling tasks. It belongs to the ensemble learning family and is based on the concept of gradient boosting, where weak learners (typically decision trees) are iteratively combined to create a strong predictive model. XGBoost incorporates several innovative techniques to enhance its performance, including regularization, parallel processing, and a unique optimization objective that combines both a loss function and a complexity penalty. These features make XGBoost highly effective in handling large datasets with high dimensionality, while also providing excellent accuracy and robustness (Chen & Guestrin, 2016).

4.3.3. Light GBM

Light GBM is a high-performance gradient boosting framework that is designed to deliver fast and accurate results in various machine learning tasks. Similar to other gradient boosting algorithms like XG Boost, Light GBM combines weak learners, typically decision trees, to create a strong predictive model. However, Light GBM introduces some unique optimizations that make it particularly efficient and effective. These optimizations, along with other techniques such as regularized learning and early stopping, enable Light GBM to handle large-scale datasets with high dimensionality efficiently while delivering competitive accuracy. Overall, Light GBM is a powerful and widely-used tool in the data scientist's toolbox, offering a balance between speed, memory efficiency, and predictive performance (Ke & Meng, 2017).

4.3.4. Neural Networks

A neural network is a type of machine learning model inspired by the structure and functioning of the human brain. Neural networks consist of interconnected layers of artificial neurons that process and transform input data to produce meaningful output predictions. With their ability to capture complex patterns and relationships in data, neural networks excel in tasks such as classification, regression, and even more advanced tasks like image recognition and natural language processing. The model's architecture, including the number of layers and neurons, can be customized to fit the specific problem at hand. Training a neural network involves iteratively adjusting its weights and biases based on the input data and desired outputs, a process known as backpropagation. Regularization techniques, such as dropout and weight decay, are commonly employed to prevent overfitting. Additionally, advanced architectures like convolutional neural networks (CNNs) for image data or recurrent neural networks (RNNs) for sequential data have been developed to tackle specific types of problems. Neural networks are widely used in data science due to their flexibility, adaptability, and ability to handle large and complex datasets. However, they also require careful hyperparameter tuning, significant computational resources, and a sufficient amount of labeled training data for optimal performance.

Keras is a high-level library for building neural networks using the Tensorflow framework for deep learning models. Keras provides flexibility for rapidly incorporating state-of-the-art deep learning models. Keras also supports sparse data, which can be particularly important for text-heavy data or categorical data with many levels.

4.4. Models Performance

In this section we will present the performance obtained after having run and calibrated the mentioned solutions (Random Forest, XGBoost, Light GBM and Keras Neural Network). The outcomes of these experiments shed light on the accuracy, robustness, and overall effectiveness of each solution in capturing the dynamic nature of CLV.

We focus on three key aspects to evaluate the performance of the top-performing model. Firstly, we assess its performance over time, considering its ability to adapt to changing customer behavior patterns. Secondly, we examine the model's capability to differentiate between ordering and non-ordering customers, aiding CookUnity in targeting high-value customers. Lastly, we analyze the overall predicted CLV compared to the actual CLV, providing insights into the model's accuracy and reliability.

By comparing and analyzing the predictive performance of these models, we aim to provide insights into the most suitable machine learning approaches for predicting CLV in various business contexts, ultimately enabling CookUnity to make informed decisions and optimize customer-centric strategies.

4.4.1. Top performing model

After carefully calibrating and evaluating the performance of all four proposed models through a random search process, we have obtained the best-performing results in terms of root mean square error (RMSE) for each model (refer to Table 6).

Table 8: Evaluated Models Performance

	RMSE		
Model	Fold 1	All Folds	Hold-Out
Random Forest	117.7611	111.2385	115.5486
XGBoost	112.4171	106.5759	111.2617
Light GBM	111.7625	106.5600	111.0845
DNN	115.8791	110.9814	115.8349

Our findings reveal that the boosting machine models, XGBoost and Light GBM, outperformed the traditional Random Forest model in terms of predictive accuracy. This outcome can be attributed to the inherent strengths of boosting algorithms, such as their ability to handle complex feature sets effectively and capture intricate patterns in the data. The boosting models' ensemble approach, which combines multiple weak learners to form a strong predictor, enables them to achieve superior performance.

On the other hand, the deep neural network (DNN) solution achieved comparable performance to the Random Forest model, as suggested by Chamberlain et al. (2017). However, it is important to note that the DNN may not have fully exploited its potential due to the specific characteristics of our feature set, which consists of a large number of predictors. Boosting models, like XGBoost and Light GBM, are known for their effectiveness in handling high-dimensional feature spaces, which may explain their superior performance in our case.

Although both XGBoost and Light GBM demonstrated similar performance, we have chosen to proceed with Light GBM as our preferred model. This decision is based on its slightly better performance in the hold-out evaluation. Nonetheless, it is worth noting that either of these models would be equally suitable for our predictive CLV solution, given their comparable performance and robustness.

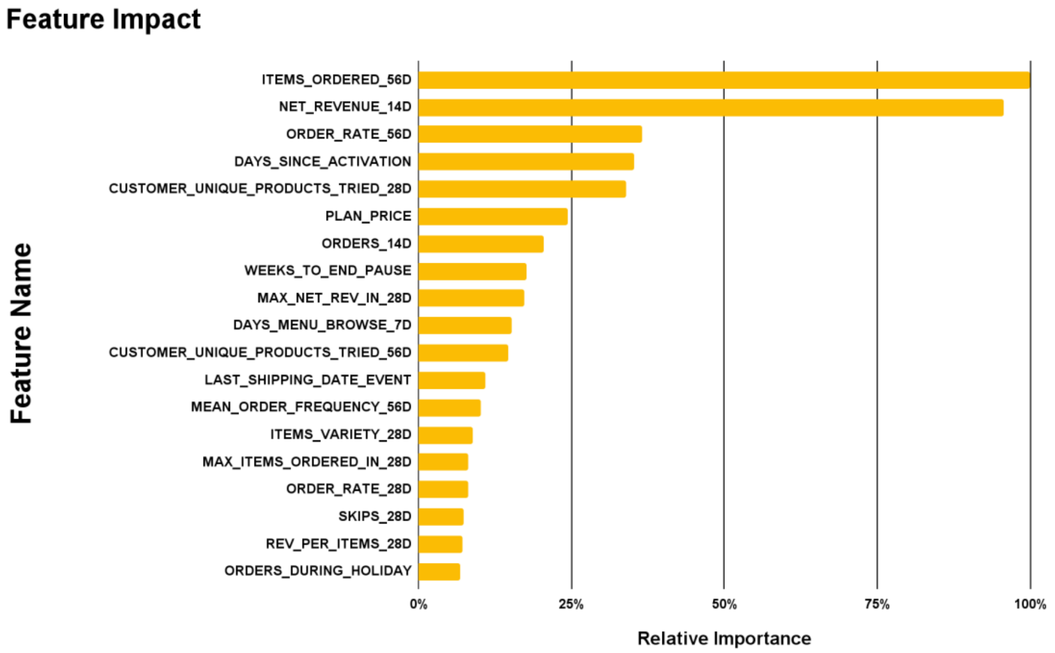
By selecting the best-performing model and considering the unique characteristics of our dataset, we can proceed with confidence in utilizing this model to predict customer lifetime value effectively.

4.4.2. Feature Importance

Before delving into the model's output and its predictions, it is essential to examine the feature importance as identified by the selected model. This analysis serves two purposes: validating the conclusions drawn during the exploratory analysis and ensuring that the predictions are not solely reliant on a single dominant feature, as this could indicate potential data leakage.

The top three features that demonstrate the greatest predictive capacity are frequency, recency, and monetary value, commonly referred to as RFM variables. These findings align with the original models reviewed in the literature, which were primarily based on the RFM framework. Additionally, the feature "DAYS_SINCE_ACTIVATION," representing the time difference between the customer's activation and the date of their first full-price order, is expected to play a crucial role in capturing customer maturity.

Figure 21: Top Feature Impact by Relative Importance



Furthermore, the subscription plan-related features, such as "PLAN_PRICE" and "REV_PER_ITEMS_28D," are also influential factors in the model's predictions. These features provide insights into the customer's chosen subscription plan and therefore their expenditure levels. The importance of meal variety is evident through features like "UNIQUE_PRODUCTS_TRIED" and "ITEMS_VARIETY," which aim to capture the diversity of meals ordered by the customer.

Moreover, in line with the concept of building habit, the features "MAX_NET_REVENUE_IN_28D" and "MAX_ITEMS_ORDERED_IN_28D" demonstrate significant relevance. These features reflect the customer's highest net revenue and maximum number of items ordered within a 28-day period, reinforcing the notion that consistent engagement in ordering over consecutive weeks contributes to long-term customer retention. Lastly, "ORDERS_DURING_HOLIDAY" emerges as another impactful feature, indicating how many times a customer placed orders during special low season weeks.

Overall, the feature importance ranking aligns with the expectations derived from the exploratory analysis, further validating the relevance of these factors in predicting customer lifetime value. Furthermore, the absence of any apparent data leakage suggests that the model's performance is not solely reliant on a single feature, but rather takes into account a comprehensive set of influential predictors.

4.4.3. Results Evaluation

Within the evaluation of the top-performing model, we prioritize three crucial aspects to assess its performance thoroughly.

Firstly, we examine its performance over time, aiming to ensure its adaptability to changing seasonal patterns and trends. Validating if by utilizing time-series cross-validation, we effectively mitigate the risk of overfitting to specific seasonality, enhancing the model's robustness

Secondly, we evaluate the model's efficacy in distinguishing between ordering (CLV > $0) and non-ordering (CLV=$0) customers. Validating if by addressing the zero-inflated issue employing a logarithmic transformation instead of a complex multilayer model was successful.

Lastly, we conduct a comprehensive analysis of the model's predicted CLV compared to the actual CLV. This analysis serves to quantify the model's accuracy and reliability in forecasting the revenue potential of individual customers, providing valuable insights for business decision-making.

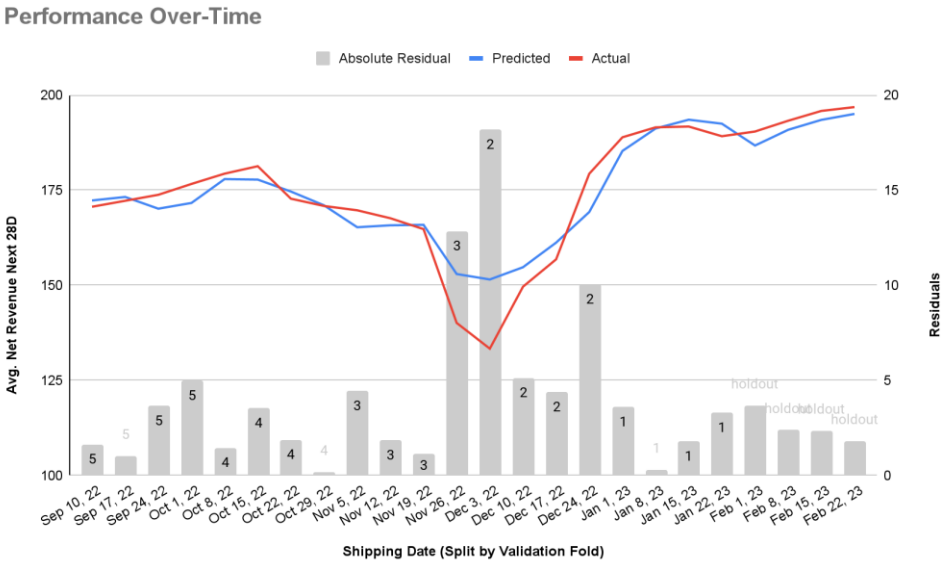
By thoroughly examining these aspects, we gain a comprehensive understanding of the top-performing model's performance, ensuring its efficacy in accurately predicting CLV and enabling CookUnity to optimize customer-centric strategies effectively.

4.4.3.1. Performance Over-Time

When evaluating the model's performance, it is crucial to ensure that it does not overfit to a specific season during training but maintains a consistent average performance across different seasons, considering the inherent ordering seasonality experienced by the company. While seasonality features have been identified as significant predictors, it is essential to assess whether the model can adjust its predictions accurately across different seasons.

Having in mind that once in production the model will be scoring the expected net revenue for the next 28 days for all active users, to gain insights into the model's performance over time, we examine Figure 22, which contrasts the average predicted revenue (represented by the blue line on the left axis) in a given week with the actual average results (represented by the red line). Additionally, the bar chart on the right axis illustrates the residuals (identified with the validation fold number they belong to), representing the differences between the predicted and actual values.

Figure 22: Model Performance Over-Time



From the analysis, we observe that the average revenue per user (ARPU) during Q3 and Q4 is notably lower compared to Q1. The model successfully captures this pattern and reflects it in its predictions. However, during holiday weeks, the model fails to precisely capture the exact impact of these atypical weeks. It underestimates the drop during thanksgiving weeks (Nov 26) and overestimates the drop for Christmas and New Year's Eve.

Ideally, the model should exhibit greater accuracy in predicting user behavior during these holiday weeks. However, due to the limited number of historical data points and the varying impact of seasonality over the years, the current performance is considered satisfactory for our purposes.

It is important to note that the model, utilizing provided seasonal features and time-series cross-validation, has achieved the necessary performance level. In subsequent sections, we will explore recommendations for future interactions and strategies that could be explored to further improve the model's performance.

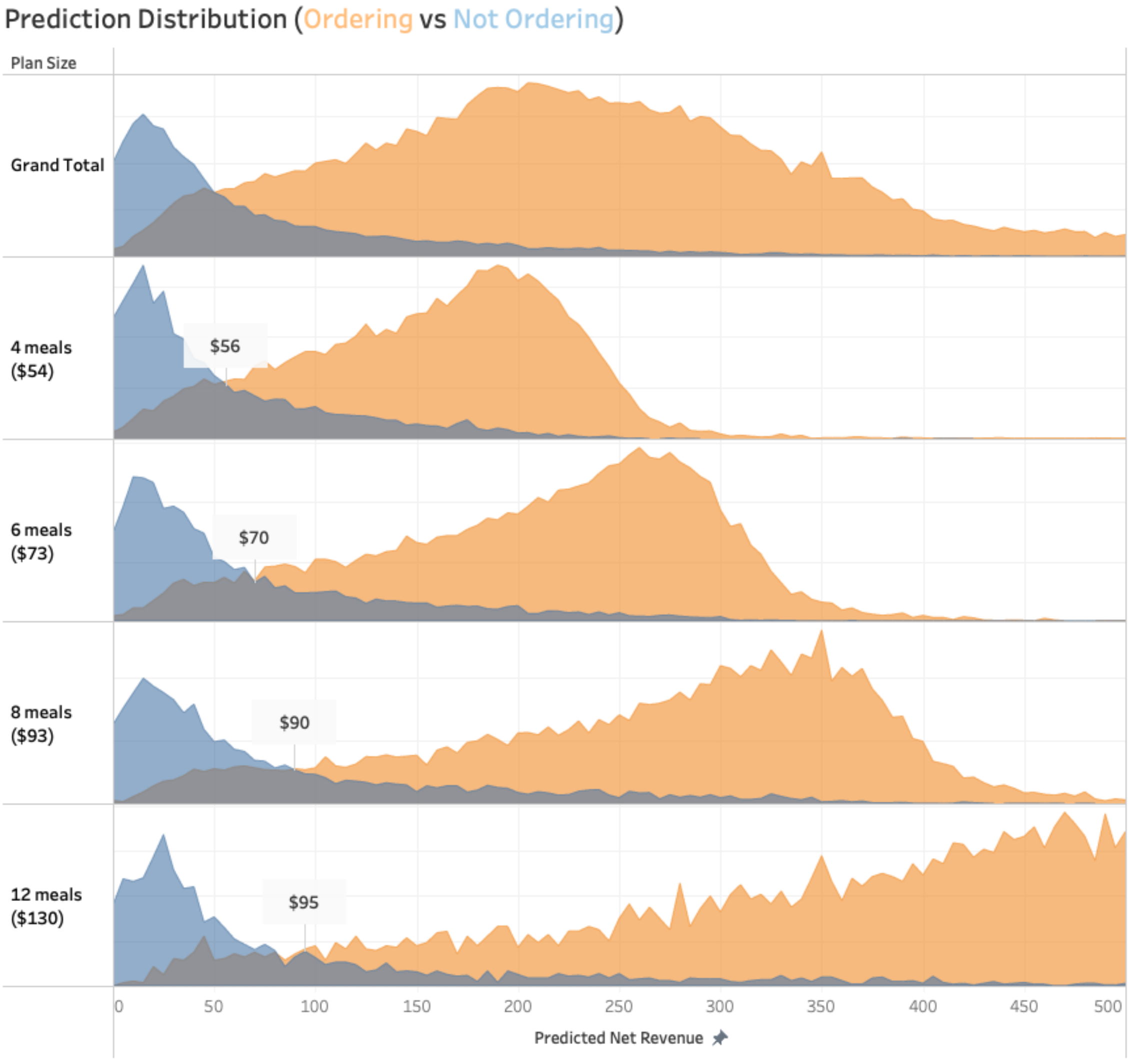
4.4.3.2. Detection of Non-Ordering Users

While exploring other solutions, we encountered a common approach in CLV estimation, which involves developing a classification model to differentiate customers who are predicted to place an order (CLV > \$0) from those who are not (CLV = \$0). However, in the context of our subscription-based business, where order values are fixed by the subscription plan costs, we propose a regression-based solution without the need for an intermediate classification step.

By directly applying regression, we can predict the CLV for each customer, and if the predicted CLV falls a certain threshold below the corresponding plan price, we assume that the customer is unlikely to place an order. This simplifies the process and removes the necessity of a classification model.

To determine whether our regression model achieves sufficient performance, we examine the density distributions of predicted CLV values. In Figure 23, we plot the density distributions of customers who actually placed an order (CLV > \$0) in orange and those who did not (CLV = \$0) in blue. The x-axis represents the predicted CLV values.

Figure 23: Prediction Distribution By Ordering and Non-Ordering Users



Upon analyzing the overall density distributions, we observe a clear differentiation between the two classes. Furthermore, when we delve deeper into specific subscription plans, we observe an even more pronounced separation, with the predicted CLV values consistently crossing the corresponding plan prices.

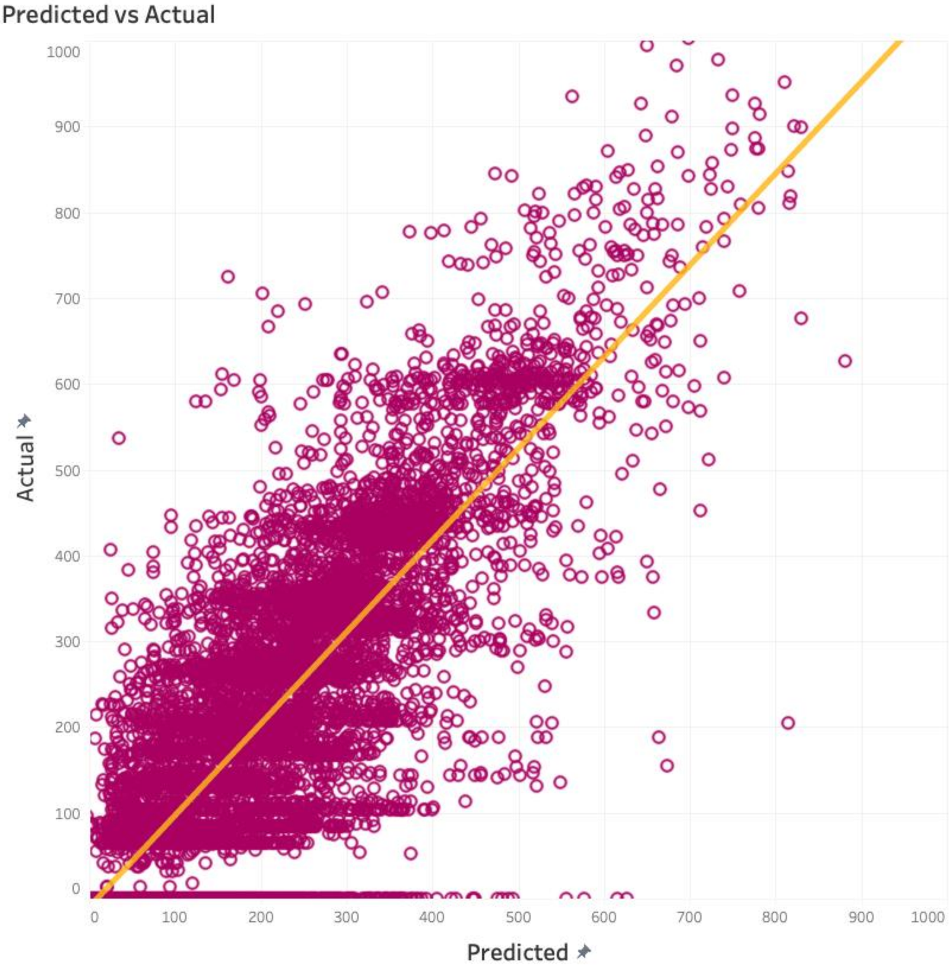
Hence, by setting a threshold such as '<IF Predicted CLV < x% PlanPrice THEN $0>', we can accurately identify customers who are unlikely to place an order. This approach effectively compensates for the absence of an intermediate classification model and aligns with the fixed order values dictated by the subscription plan prices.

The selection of the threshold for classifying customers as unlikely to place an order should be carefully evaluated based on each specific business need and the associated loss costs of error type 1 (misclassifying a potential customer) versus error type 2 (failing to identify customers who will not place an order). The optimal threshold will depend on the relative costs and priorities of each business case where this solution may be implemented, therefore not to be evaluated within this analysis.

4.4.3.3. Predicted CLV vs Actual

To dig further than just looking at the overall RMSE performance of the model, let's review a visual interpretation of the predictions by assessing the distribution of predicted vs actual scatter plot (figure 24).

Figure 24: Predicted vs Actuals Distribution



To represent the above distribution we took predicted numbers for one of the weeks in the hold-out to reduce the amount of data being displayed. This scattered plot of predicted vs actual has a RMSE of $113 that can be interpreted as the average difference between values predicted by a model and the actual values. Therefore, complementing this with the avg. plan prices reviewed in the EDA, we can say that the average difference is +- one order in a five weeks period.

We can also complement this analysis by calculating an extra measure which is the R-square which in this case is 0.69, that can be read as how well the regression model explains observed data. So we could say that 70% of the variability observed in the target variable is explained by the regression model.

To conclude this analysis, the overall performance achieved is sufficient for the objectives we have proposed to achieve. Either way, it's important to consider that this is the overall performance and most probably the model is predicting better on some customer segments. In further sections we will delve into this.

4.4.4. Modeling Conclusion

In this section, we conducted a comprehensive exploration and analysis of CookUnity's available data sources, which served as the foundation for building our feature set. Through rigorous feature engineering, we prepared the data for training four different models, including three tree-based models and one neural network model.

After careful evaluation, we determined that the Light GBM model emerged as the top performer. The results obtained from the time-series cross-validation demonstrated that the model was well-calibrated and exhibited consistent performance across different seasons, showcasing its effectiveness in distinguishing between ordering and non-ordering users.

While the overall root mean square error (RMSE) performance was commendable, we believe there is still room for improvement. Inspired by the work of Vanderveld et al. (2016), our next step in the subsequent section will involve delving into performance variations across customer segments. By identifying patterns within these segments, we aim to gain deeper insights into potential areas of enhancement and refinement.

Through this meticulous analysis and evaluation process, we have laid a solid foundation for understanding the performance and potential of our predictive models. The forthcoming section will contribute valuable insights to guide our future efforts in optimizing the model's performance and catering to the unique needs and characteristics of specific customer segments.

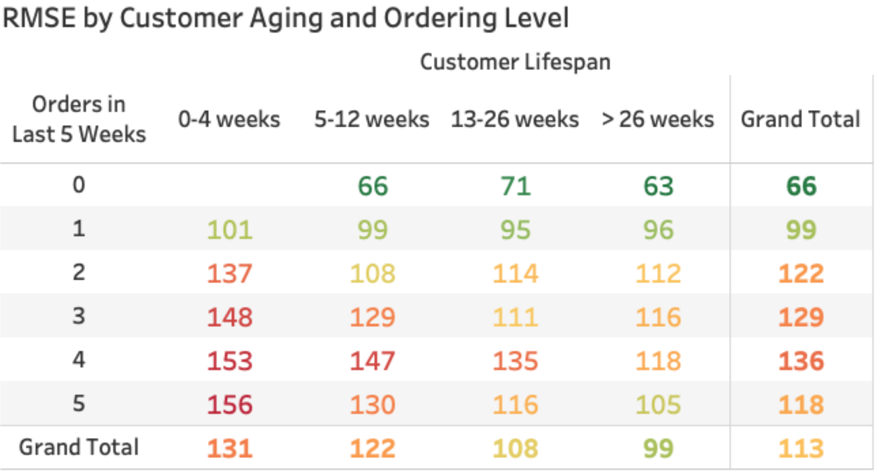
5. Discussion on results

In this section, we delve into a comprehensive discussion of the results obtained from our CLV (Customer Lifetime Value) modeling study. As we analyze the performance of our models across different customer segments, it becomes evident that the predictive accuracy and effectiveness vary significantly. This observation highlights the potential impact of additional information on certain customer segments. By exploring the role of enhanced data availability, we can gain deeper insights into the nuances of CLV modeling and uncover strategies for effectively targeting and engaging customers within these specific segments. In doing so, we aim to contribute to the advancement of CLV modeling techniques and foster a more comprehensive understanding of customer behavior and value creation.

5.1. Performance Across Customer Segments

Drawing from the insights presented in Vanderveld et al. (2016), it is evident that the performance of CLV models varies across different customer segments. This variation provides valuable insights into the strengths and weaknesses of our model. To further explore these performance dynamics, we present Table 7, which provides a breakdown of performance metrics based on the number of orders a customer made during the last 5 weeks, segmented by customer aging in weeks.

Table 9: RMSE by Customer Aging and Ordering Level



RMSE by Customer Aging and Ordering Level

	Customer Lifespan				
Orders in Last 5 Weeks	0-4 weeks	5-12 weeks	13-26 weeks	> 26 weeks	Grand Total
0		66	71	63	66
1	101	99	95	96	99
2	137	108	114	112	122
3	148	129	111	116	129
4	153	147	135	118	136
5	156	130	116	105	118
Grand Total	131	122	108	99	113

Upon analysis, we observe a discernible pattern: as the maturity of the customer increases, the model's predictive accuracy tends to improve. This improvement can be attributed to the availability of more historical data for long-standing customers, enabling the model to better capture their ordering behavior and make accurate predictions. Conversely, the model encounters challenges in accurately predicting customer behavior within the medium range of ordering frequency. This suggests that customers in this segment exhibit more erratic ordering patterns, making it difficult for the model to capture a strong signal. However, we acknowledge that further improvements could be achieved by incorporating additional data, which we will discuss in subsequent sections.

5.2. Enabling Strategic Decisions Based on Predicted CLV

By operationalizing this CLV model, CookUnity gains the ability to predict CLV for each customer on an ongoing basis, unlocking numerous strategic decision-making opportunities. While it is not feasible to explore all potential implementations within the scope of this work, we will provide examples of some key use cases.

5.2.1. Acquisition Performance Monitoring

Traditionally, CookUnity measures customer acquisition costs (CAC) to evaluate marketing performance. However, the true value lies in understanding the relationship between CAC and customer profitability (CLV). With our predictive CLV, CookUnity's marketing team can assess the value of newly acquired customers early on and adjust acquisition strategies accordingly. This enables quicker decision-making, optimizing marketing investments based on predicted CLV instead of waiting for the average payoff period.

5.2.2. Triggering Engagement Campaigns

The predicted CLV serves as an invaluable tool for the CRM team to segment the audience for engagement campaigns. By leveraging the expected profit from a customer, campaigns can be targeted in various ways. For instance:

- Targeting customers predicted to have lower CLV: Focusing on these customers allows for tailored engagement to increase their value.
- Identifying potential disengagement: Contrasting the predicted CLV with the recent revenue generated by a customer can indicate if they are disengaging. Campaigns can be triggered to re-engage these customers.
- Predicted revenue increase: Customers predicted to generate more revenue can be targeted without the need for additional engagement efforts.

5.2.3. Assessing Impact of Interactions

Predicted CLV can be used to evaluate the impact of specific interactions on customers or groups. For example, if an operational issue occurs, comparing the predicted CLV before and after the issue helps operations teams assess the monetary impact and identify affected customers. This enables proactive actions to compensate for any potential loss. Similarly, changes such as menu updates or improvements in recommendation algorithms can be evaluated by analyzing the predicted CLV before and after the updates, providing insights into their impact.

5.3. Challenges and Limitations in Model Performance

5.3.1. Enhancing Predictions for Early Customers

After conducting the analysis across distinct customer segments, it becomes evident that the model's performance significantly improves when predicting the behavior of mature customers. This improvement can be attributed to the availability of a larger volume of data and a more extensive history for each of these customers. Conversely, the model encounters challenges when predicting the behavior of customers during their initial weeks, primarily due to the lack of historical purchase behavior data, which is a crucial predictor in the model.

To address this limitation and enhance the model's predictive capabilities for early customers, it is proposed to incorporate and promote the collection of additional data points during their initial interactions. Specifically, the product team could encourage early customers to provide information through surveys regarding their eating preferences and household demographics, such as age, gender, or family size. By capturing these early predictors, the model can leverage this information to better predict the customer's future behavior and improve CLV predictions.

Moreover, the company should consider evaluating the feasibility of obtaining external data sources related to household income, purchasing behaviors, or other relevant variables. Integrating such data into the model would offer a more comprehensive understanding of early customers' characteristics and enable more accurate predictions of their CLV. Additionally, exploring other potential data sources or data partnerships that provide valuable insights into early customers' preferences and behaviors can further enhance the model's predictive capacity.

In conclusion, to overcome the limitations associated with the lack of historical purchase behavior data for early customers, it is recommended to gather additional early predictors through surveys on eating preferences and household demographics. Furthermore, exploring opportunities to acquire external data sources related to income, purchasing behaviors, and other relevant variables can further augment the model's predictive capabilities.This enhancement would result in more accurate forecasts and enable the company to make more informed decisions regarding early customer acquisition and retention strategies.

5.3.2. Addressing Seasonal Variations through Data Augmentation Techniques

Although the current model demonstrates the ability to identify overall trends using the existing feature set, accurately capturing the exact seasonality during certain weeks is challenging due to limited data points. With a dataset spanning two years and a quarter, the model lacks sufficient historical data to precisely capture the dynamics of certain dates.

The challenge of accurately capturing seasonal variations in CLV predictions can be addressed through data augmentation techniques. By assigning different weights to weeks with significant ordering fluctuations, augmenting the dataset with synthetic data, and employing ensemble learning, the model can adapt to irregularities in customer behavior and improve its ability to forecast CLV during periods with unique seasonal patterns. These techniques enhance the model's capacity to handle atypical weeks and ensure that the predictions align closely with the specific dynamics of seasonal fluctuations, leading to more precise and informed decision-making in customer relationship management.

6. Conclusion and Recommendations

6.1. Summary of the study

This study focused on unlocking customer lifetime value (CLV) in the context of CookUnity, a dining experience company. The objective was to develop a predictive CLV model and understand consumer engagement drivers. The study employed exploratory data analysis, feature engineering, and evaluated various modeling approaches such as random forest, XGBoost, Light GBM, and neural networks. The top-performing model was Light GBM, which demonstrated good performance across different customer segments. The study highlighted the importance of evaluating performance metrics across customer aging and ordering behavior. It also discussed the potential use cases of CLV models, including acquisition performance monitoring, triggering engagement campaigns, and assessing the impact of interactions. The study acknowledged challenges and limitations in model performance, particularly for early customers and seasonal variations. Recommendations were provided for enhancing predictions, addressing seasonal variations through data augmentation techniques, and further research in the field.

6.2. Conclusion and implications of the findings

The developed predictive CLV model, particularly the Light GBM approach, showed promising performance in predicting customer lifetime value across various customer segments. The study revealed that as customer maturity increases, the model's predictive accuracy improves. This highlights the importance of leveraging historical data for accurate CLV predictions. The study also emphasized the strategic implications of CLV models, such as optimizing acquisition performance and targeting engagement campaigns based on predicted CLV. Furthermore, the assessment of the impact of interactions on CLV can provide valuable insights for operational and promotional decision-making. However, the study also identified challenges in accurately predicting CLV for early customers and addressing seasonal variations.

6.3. Recommendations for further research

Based on the findings and limitations of this study, several areas of further research can be explored to enhance CLV modeling and its applications in the context of CookUnity:

- Advanced Predictive Models: Explore more advanced machine learning models and techniques to further improve the accuracy of CLV predictions. Investigate the effectiveness of ensemble models or deep learning architectures in capturing complex customer behaviors and patterns.
- Incorporating External Data: Evaluate the integration of external data sources, such as demographic data, customer feedback, or social media data, to enhance CLV predictions. Assess the impact of including these additional variables on the model's performance.
- Seasonality and Data Augmentation: Develop data augmentation techniques to address the seasonal variations observed in CLV modeling. Explore methods to synthetically generate data points to improve model performance during seasonal shifts or periods with limited data.
- Long-Term CLV and Retention Strategies: Analyze the long-term CLV patterns and retention strategies to identify key factors influencing customer ordering behavior.

Develop proactive retention strategies based on predicted CLV to minimize customer attrition and maximize customer lifetime value.

- Validation and Model Monitoring: Establish a robust validation framework and ongoing model monitoring system to assess the performance and accuracy of CLV models over time. Implement feedback loops and regular updates to ensure the models remain relevant and effective.

By pursuing these avenues of research, CookUnity can further enhance its understanding of customer lifetime value, refine its strategic decision-making processes, and continuously improve customer engagement and retention strategies.

7. References

Baxter, R. K. (2015). *The Membership Economy: Find Your Super Users, Master the Forever Transaction, and Build Recurring Revenue*. McGraw-Hill Education.

Benoit, D. F., & Van den Poel, D. (2012). Improving customer retention in financial services using kinship network information. *Expert Systems with Applications*, *39*(13), 11435-11442. https://doi.org/10.1016/j.eswa.2012.04.016

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Čermák, P. (2015). Customer Profitability Analysis and Customer Life Time Value Models: Portfolio Analysis. *Procedia Economics and Finance*, *25*, 14-25. https://doi.org/10.1016/S2212-5671(15)00708-X

Chamberlain, B. P., Cardoso, A., Bryan Liu, C. H., Pagliari, R., & Deisenroth, M. P. (2017). Customer Lifetime Value Prediction Using Embeddings. *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1753-1762. https://doi.org/10.1145/3097983.3098123

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *22*, 785–794. https://doi.org/10.1145/2939672.2939785

Crowder, M., Hand, D., & Krzanowski, W. (2007). On optimal intervention for customer lifetime value. *European Journal of Operational Research*, *183*(3), 1550-1559. https://doi.org/10.1016/j.ejor.2006.08.062

Dwyer, F. (1997). Customer lifetime valuation to support marketing decision making. *Journal of Direct Marketing*, *11*(4), 6-13. https://journals.sagepub.com/doi/abs/10.1002/%28SICI%291522-7138%28199723%2911%3A4%3C6%3A%3AAID-DIR3%3E3.0.CO%3B2-T

Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Publications.

Fleming, N. (2016). *The Customer Loyalty Loop: The Science Behind Creating Great Experiences and Lasting Impressions*. Career Press.

Gallo, A. (2014). *How Valuable Are Your Customers?* Harvard Business Review. Retrieved April 29, 2023, from https://hbr.org/2014/07/how-valuable-are-your-customers

Gupta, S. (2014). Marketing Reading: Customer Management. *Core Curriculum*, *8162*. https://hbsp.harvard.edu/product/8162-PDF-ENG

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Lin, N., Ravishanker, N., Kumar, V., & Sriram, S. (2006). Modeling Customer Lifetime Value. *Journal of Service Research*, *9*(2), 139-155. https://www.researchgate.net/publication/237287176_Modeling_Customer_Lifetime_Value

Heath, C., & Heath, D. (2017). *The Power of Moments: Why Certain Experiences Have Extraordinary Impact*. Simon & Schuster.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

Ke, G., & Meng, Q. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Conference on Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Kumar, V., & Wiener, M. (2007). Measuring and maximizing customer equity: a critical analysis. *Journal of the Academy of Marketing Science*, *35*(2), 157-171. https://link.springer.com/article/10.1007/s11747-007-0028-2

Mehta, N., Steinman, D., & Murphy, L. (2016). *Customer Success: How Innovative Companies Are Reducing Churn and Growing Recurring Revenue*. Wiley.

Ries, E. (2011). *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*. Crown.

Rosset, S., Neumann, E., Eick, U., & Vatnik, N. (2003). Customer Lifetime Value Models for Decision Support. *Data Mining and Knowledge Discovery*, *7*(3), 321-339. https://doi.org/10.1023/A:1024036305874

Shmueli, G., & Lichtendahl, K. C. (2016). *Practical Time Series Forecasting with R: A Hands-on Guide*. Axelrod Schnall Publishers. https://www.pdfdrive.com/practical-time-series-forecasting-with-r-a-hands-on-guide-2nd-edition-e187475800.html

Stoffer, D. S., & Shumway, R. H. (2006). *Time series analysis and its applications : with R examples*. Springer.

Tzuo, T., & Weisert, G. (2018). *Subscribed: Why the Subscription Model Will Be Your Company's Future - and What to Do About It*. Penguin Publishing Group.

Vanderveld, A., Pandey, A., Han, A., & Parekh, R. (2016). An Engagement-Based Customer Lifetime Value System for E-commerce. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 293-302. https://doi.org/10.1145/2939672.2939693

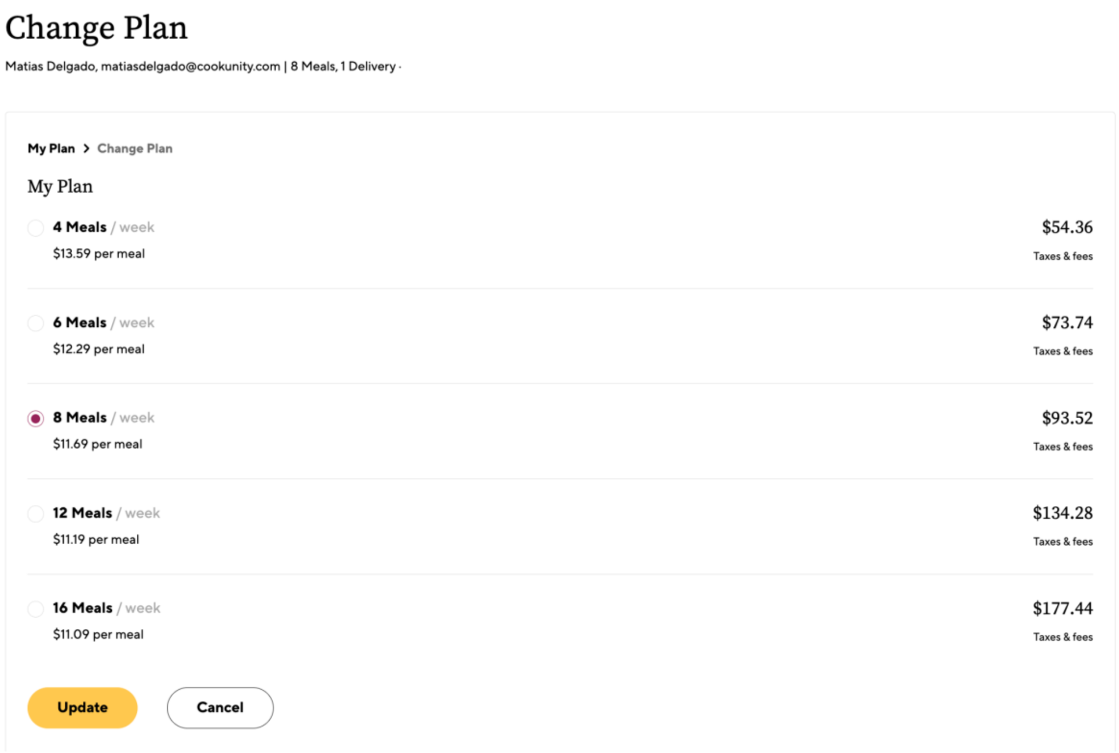
Warrillow, J. (2015). *The Automatic Customer: Creating a Subscription Business in Any Industry*. Penguin Publishing Group.

Zheng, A., & Casari, A. (2017). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly.

Zhen-Yu, C., & Peng, S. (2012). Distributed customer behavior prediction using multiplex data: A collaborative MK-SVM approach. *Knowledge-Based Systems*, *35*(1), 11-119. https://doi.org/10.1016/j.knosys.2012.04.023

Appendix A: CookUnity's Platform

Subscription Plans Offered:



Change Plan

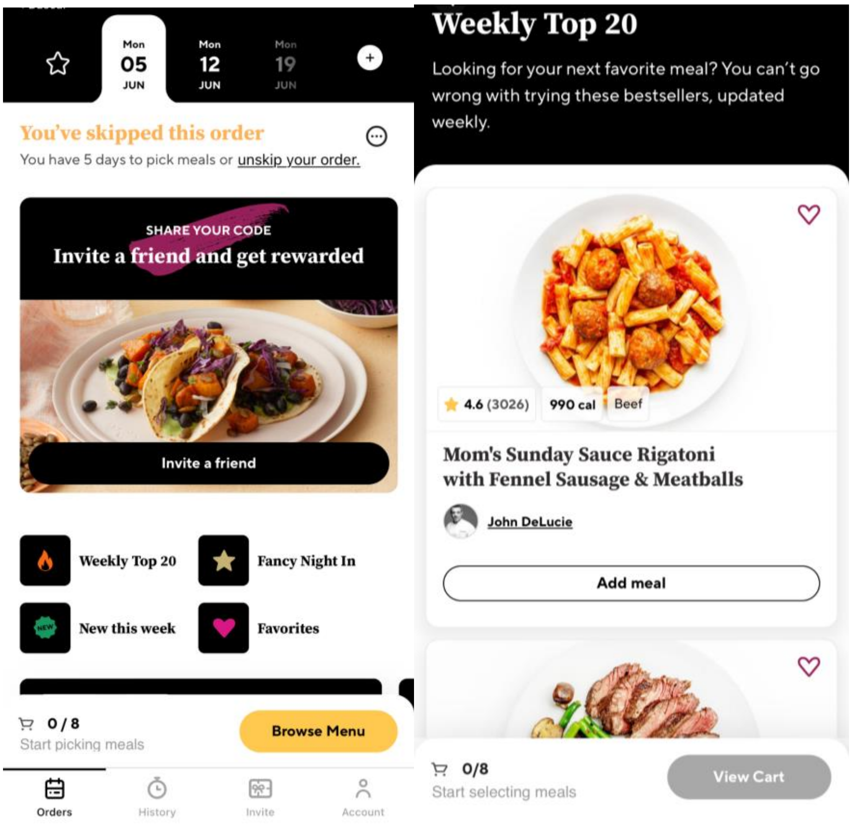
Matias Delgado, matiasdelgado@cookunity.com | 8 Meals, 1 Delivery ·

My Plan > Change Plan

My Plan

○ **4 Meals** / week $13.59 per meal	$54.36 Taxes & fees
○ **6 Meals** / week $12.29 per meal	$73.74 Taxes & fees
● **8 Meals** / week $11.69 per meal	$93.52 Taxes & fees
○ **12 Meals** / week $11.19 per meal	$134.28 Taxes & fees
○ **16 Meals** / week $11.09 per meal	$177.44 Taxes & fees

Mobile Landing Page:



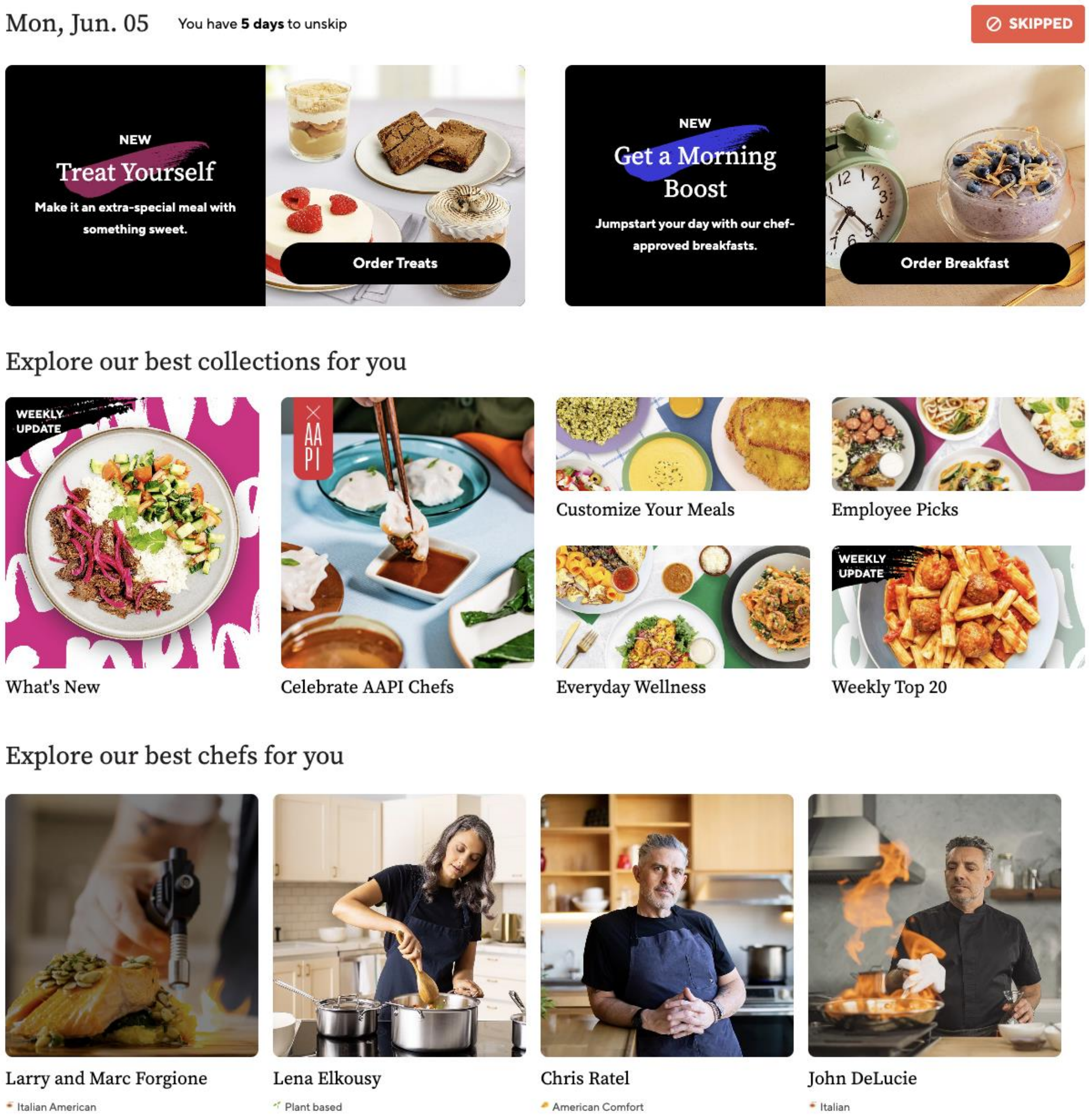
You've skipped this order

You have 5 days to pick meals or unskip your order.

Weekly Top 20

Looking for your next favorite meal? You can't go wrong with trying these bestsellers, updated weekly.

Web Landing Page:



Mon, Jun. 05 You have **5 days** to unskip

NEW

Treat Yourself

Make it an extra-special meal with something sweet.

NEW

Get a Morning Boost

Jumpstart your day with our chef-approved breakfasts.

Explore our best collections for you

What's New

Celebrate AAPI Chefs

Customize Your Meals

Employee Picks

Everyday Wellness

Weekly Top 20

Explore our best chefs for you

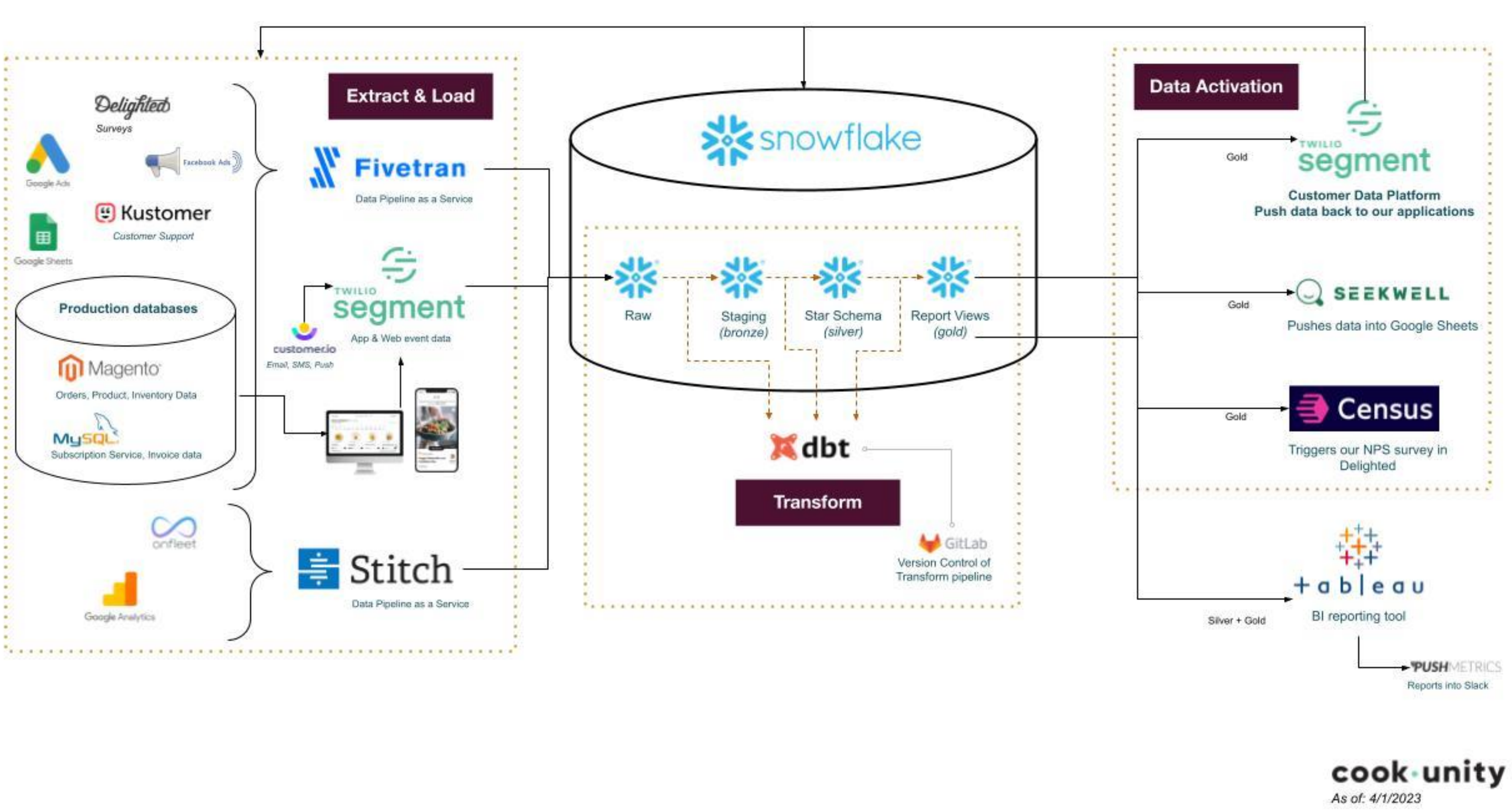
Larry and Marc Forgione – Italian American

Lena Elkousy – Plant based

Chris Ratel – American Comfort

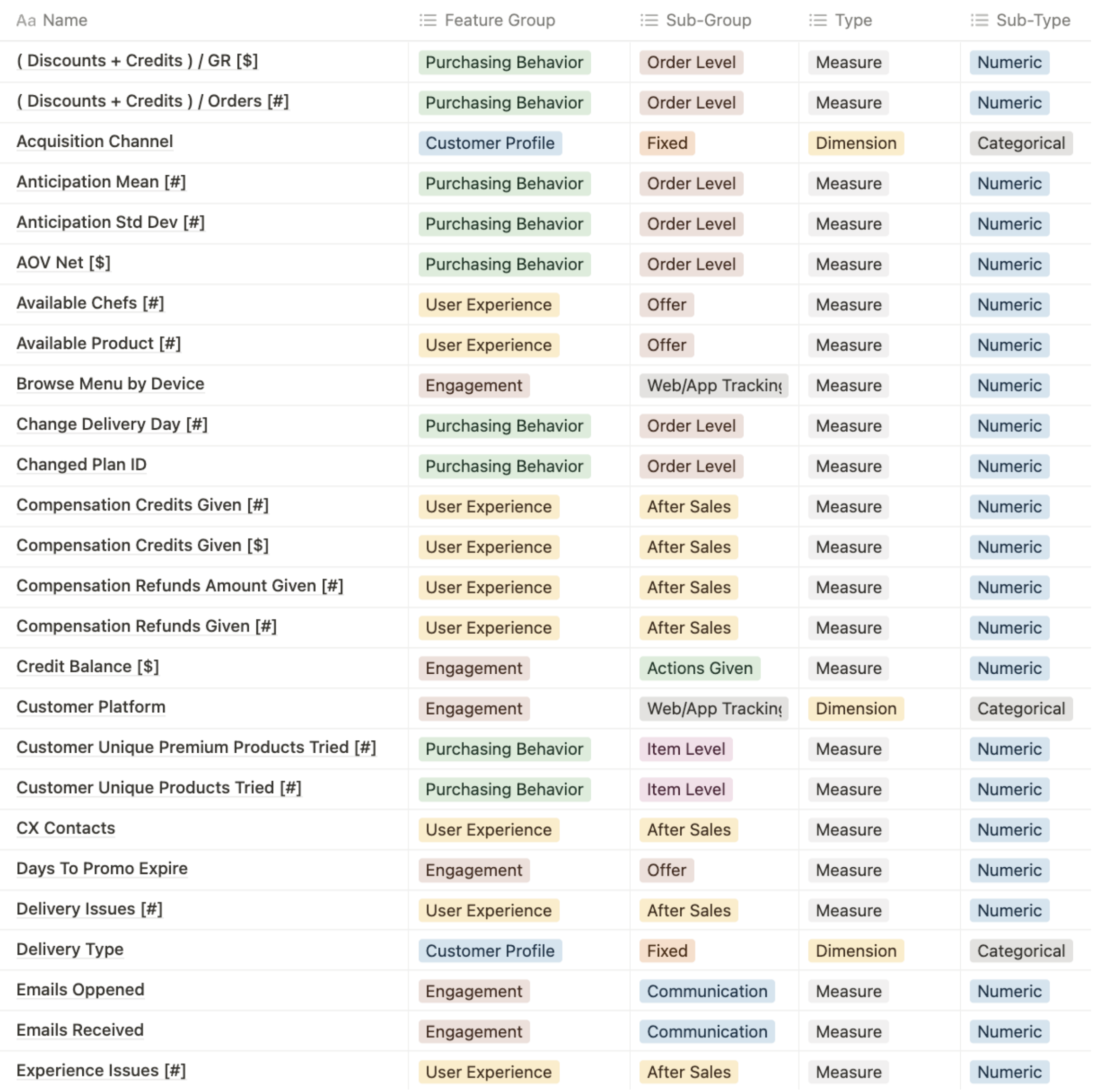
John DeLucie – Italian

Appendix B: CookUnity's Data Stack

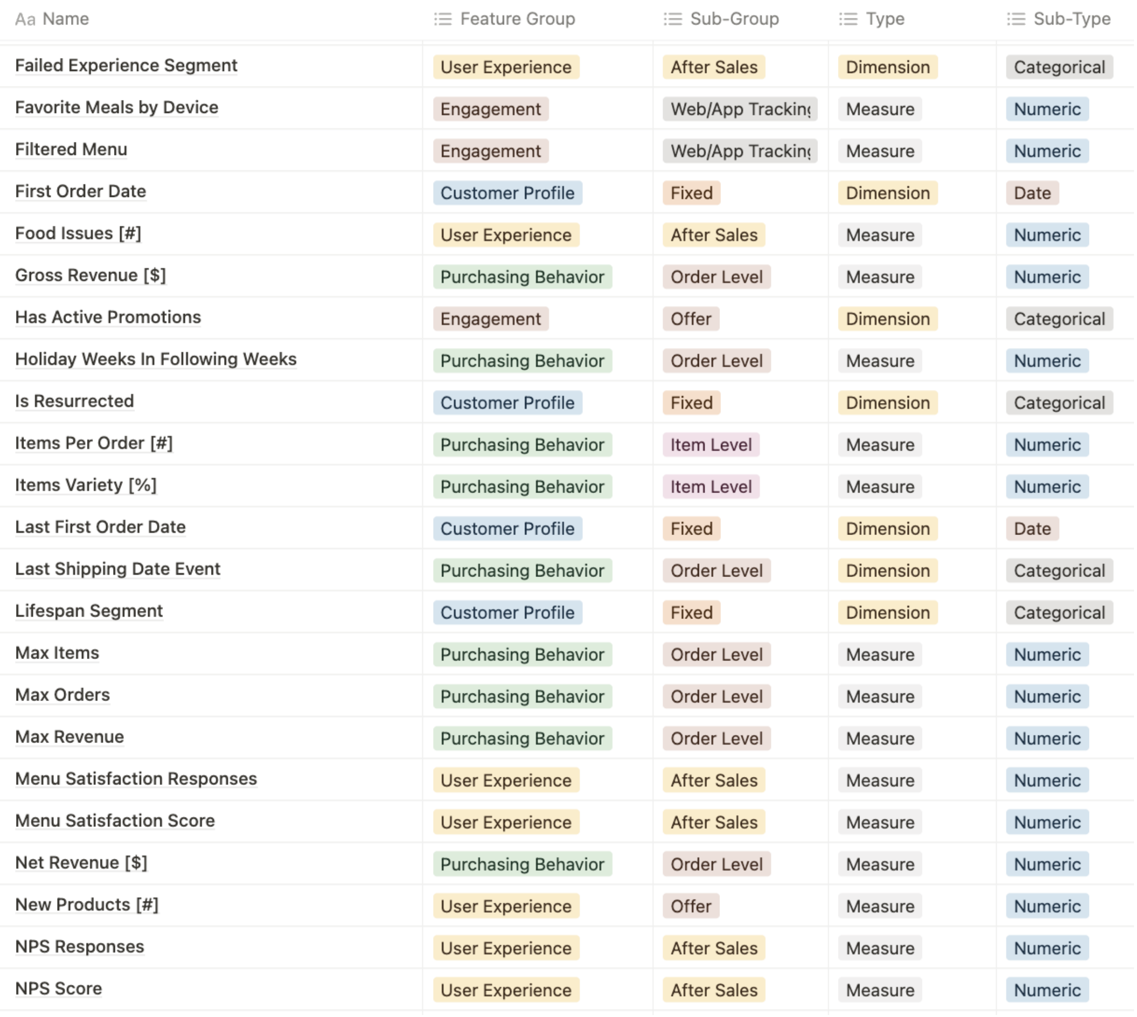


As of: 4/1/2023

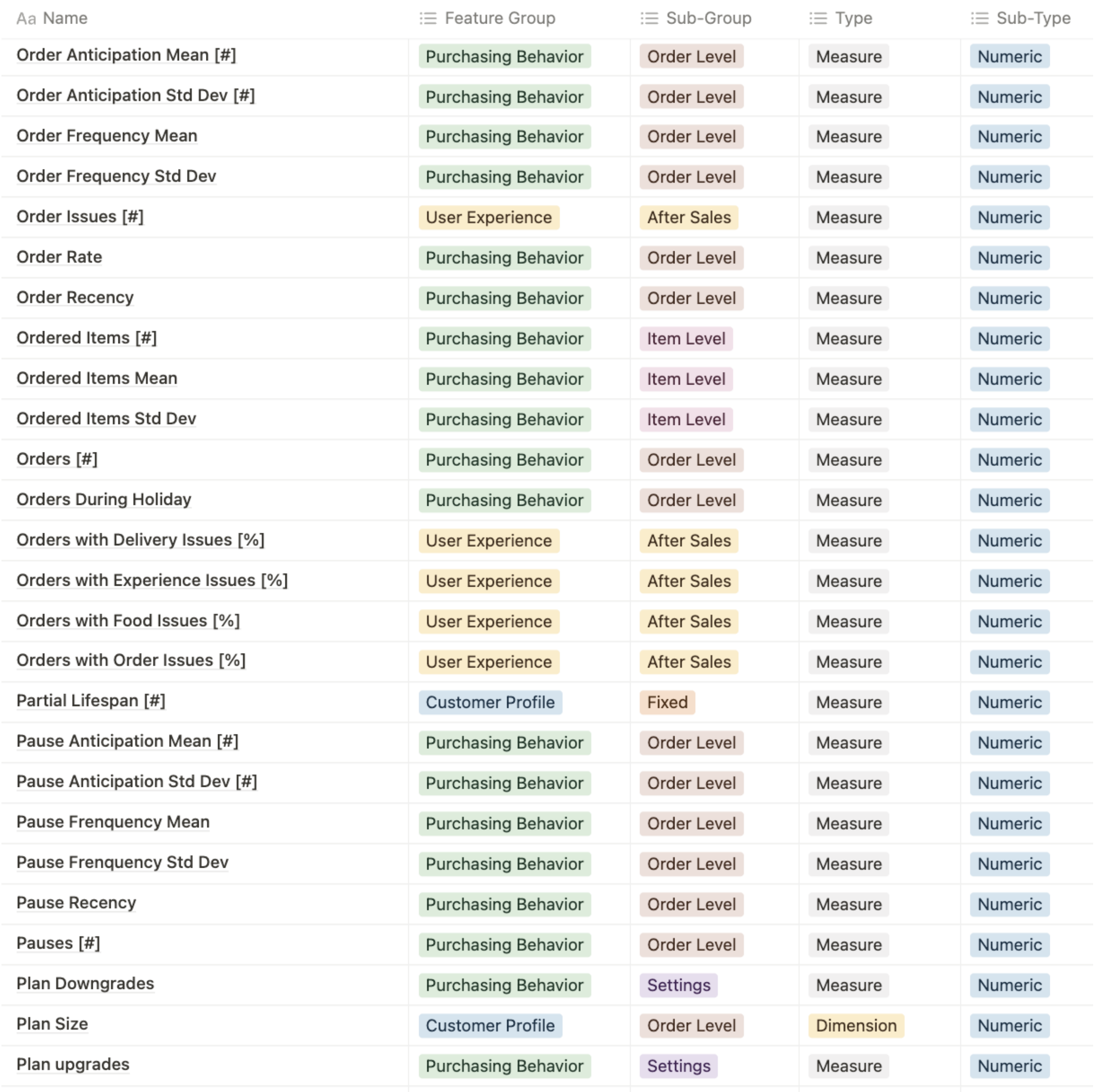
Appendix C. Predictive Feature Set



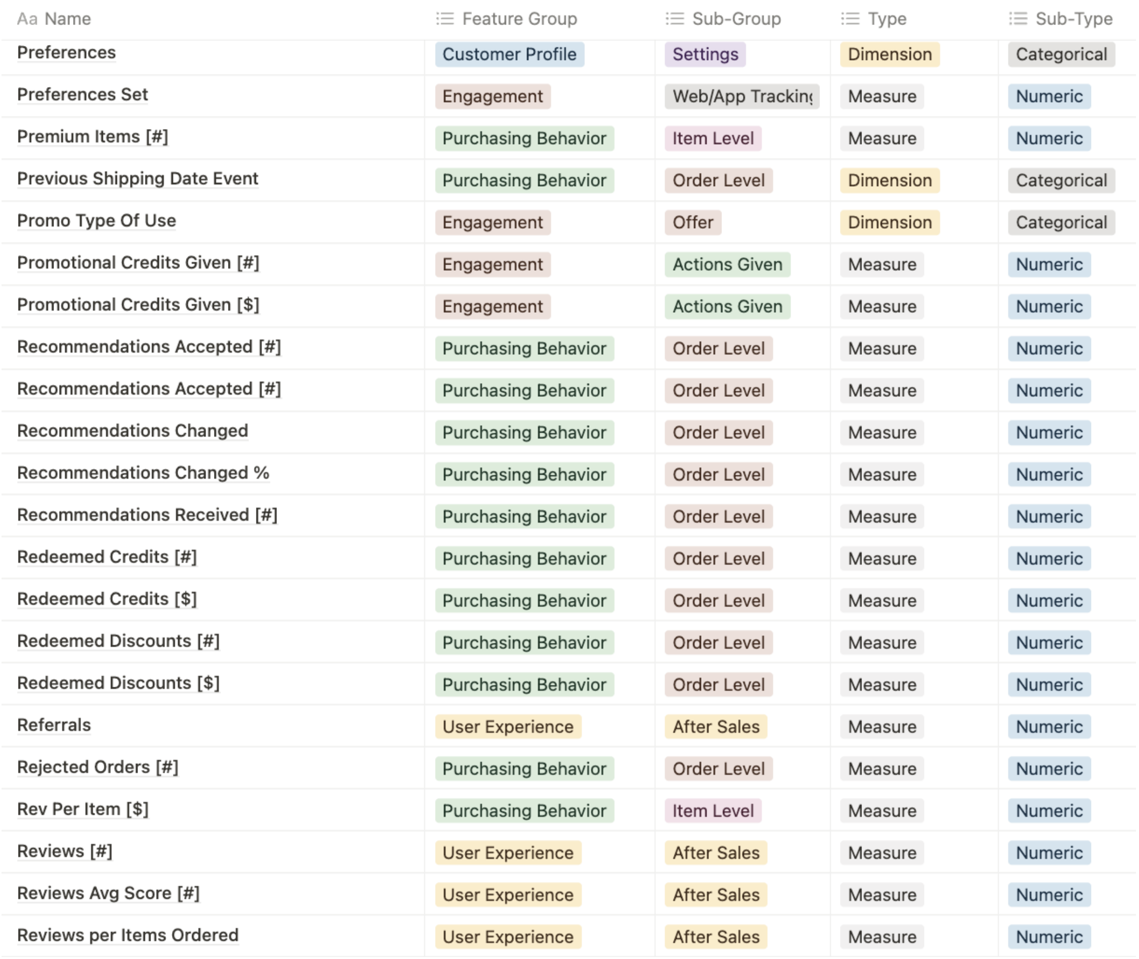
Aa Name	Feature Group	Sub-Group	Type	Sub-Type
(Discounts + Credits) / GR [$]	Purchasing Behavior	Order Level	Measure	Numeric
(Discounts + Credits) / Orders [#]	Purchasing Behavior	Order Level	Measure	Numeric
Acquisition Channel	Customer Profile	Fixed	Dimension	Categorical
Anticipation Mean [#]	Purchasing Behavior	Order Level	Measure	Numeric
Anticipation Std Dev [#]	Purchasing Behavior	Order Level	Measure	Numeric
AOV Net [$]	Purchasing Behavior	Order Level	Measure	Numeric
Available Chefs [#]	User Experience	Offer	Measure	Numeric
Available Product [#]	User Experience	Offer	Measure	Numeric
Browse Menu by Device	Engagement	Web/App Trackin	Measure	Numeric
Change Delivery Day [#]	Purchasing Behavior	Order Level	Measure	Numeric
Changed Plan ID	Purchasing Behavior	Order Level	Measure	Numeric
Compensation Credits Given [#]	User Experience	After Sales	Measure	Numeric
Compensation Credits Given [$]	User Experience	After Sales	Measure	Numeric
Compensation Refunds Amount Given [#]	User Experience	After Sales	Measure	Numeric
Compensation Refunds Given [#]	User Experience	After Sales	Measure	Numeric
Credit Balance [$]	Engagement	Actions Given	Measure	Numeric
Customer Platform	Engagement	Web/App Trackin	Dimension	Categorical
Customer Unique Premium Products Tried [#]	Purchasing Behavior	Item Level	Measure	Numeric
Customer Unique Products Tried [#]	Purchasing Behavior	Item Level	Measure	Numeric
CX Contacts	User Experience	After Sales	Measure	Numeric
Days To Promo Expire	Engagement	Offer	Measure	Numeric
Delivery Issues [#]	User Experience	After Sales	Measure	Numeric
Delivery Type	Customer Profile	Fixed	Dimension	Categorical
Emails Oppened	Engagement	Communication	Measure	Numeric
Emails Received	Engagement	Communication	Measure	Numeric
Experience Issues [#]	User Experience	After Sales	Measure	Numeric



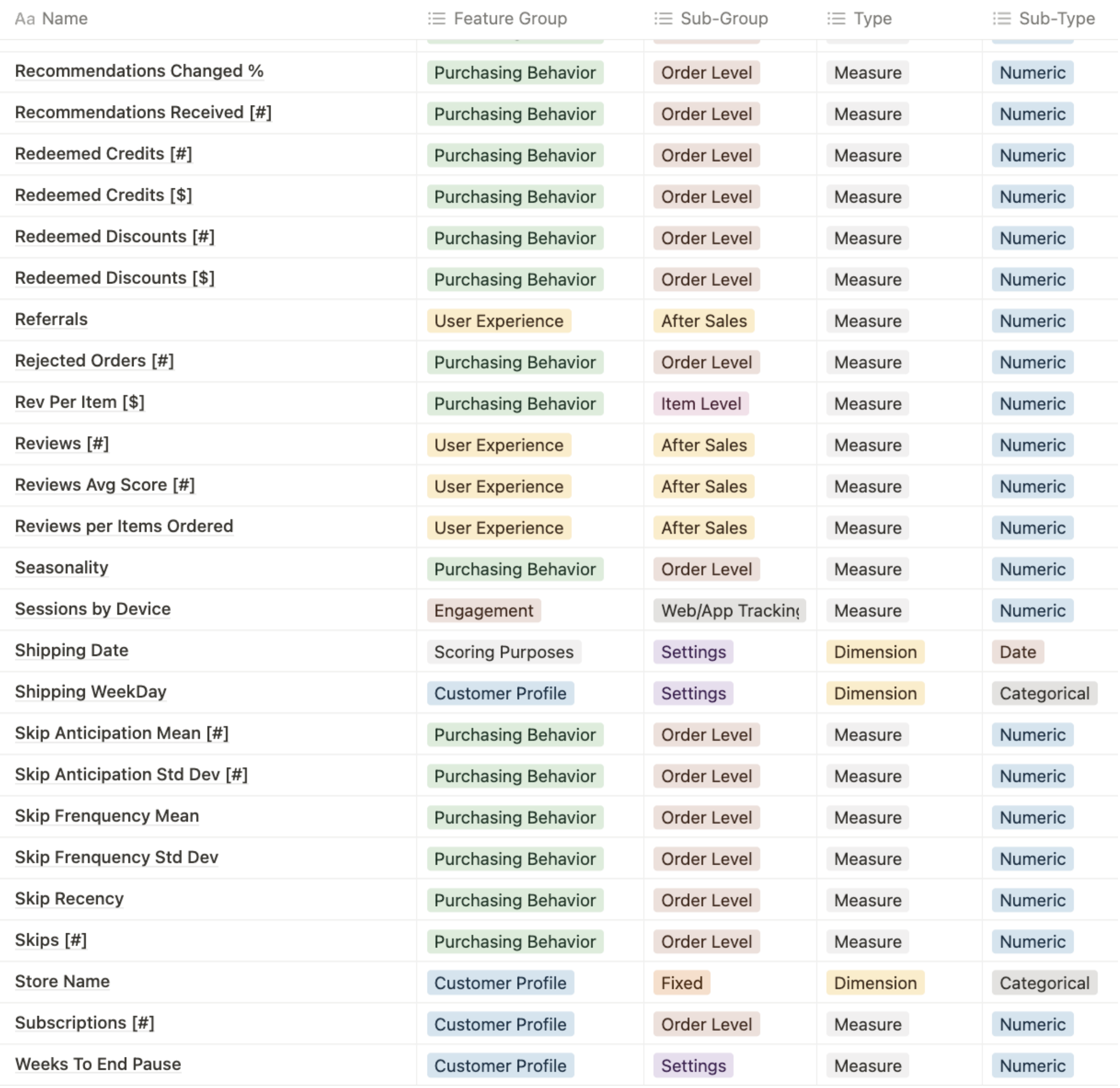
Name	Feature Group	Sub-Group	Type	Sub-Type
Failed Experience Segment	User Experience	After Sales	Dimension	Categorical
Favorite Meals by Device	Engagement	Web/App Trackin	Measure	Numeric
Filtered Menu	Engagement	Web/App Trackin	Measure	Numeric
First Order Date	Customer Profile	Fixed	Dimension	Date
Food Issues [#]	User Experience	After Sales	Measure	Numeric
Gross Revenue [$]	Purchasing Behavior	Order Level	Measure	Numeric
Has Active Promotions	Engagement	Offer	Dimension	Categorical
Holiday Weeks In Following Weeks	Purchasing Behavior	Order Level	Measure	Numeric
Is Resurrected	Customer Profile	Fixed	Dimension	Categorical
Items Per Order [#]	Purchasing Behavior	Item Level	Measure	Numeric
Items Variety [%]	Purchasing Behavior	Item Level	Measure	Numeric
Last First Order Date	Customer Profile	Fixed	Dimension	Date
Last Shipping Date Event	Purchasing Behavior	Order Level	Dimension	Categorical
Lifespan Segment	Customer Profile	Fixed	Dimension	Categorical
Max Items	Purchasing Behavior	Order Level	Measure	Numeric
Max Orders	Purchasing Behavior	Order Level	Measure	Numeric
Max Revenue	Purchasing Behavior	Order Level	Measure	Numeric
Menu Satisfaction Responses	User Experience	After Sales	Measure	Numeric
Menu Satisfaction Score	User Experience	After Sales	Measure	Numeric
Net Revenue [$]	Purchasing Behavior	Order Level	Measure	Numeric
New Products [#]	User Experience	Offer	Measure	Numeric
NPS Responses	User Experience	After Sales	Measure	Numeric
NPS Score	User Experience	After Sales	Measure	Numeric



Aa Name	Feature Group	Sub-Group	Type	Sub-Type
Order Anticipation Mean [#]	Purchasing Behavior	Order Level	Measure	Numeric
Order Anticipation Std Dev [#]	Purchasing Behavior	Order Level	Measure	Numeric
Order Frequency Mean	Purchasing Behavior	Order Level	Measure	Numeric
Order Frequency Std Dev	Purchasing Behavior	Order Level	Measure	Numeric
Order Issues [#]	User Experience	After Sales	Measure	Numeric
Order Rate	Purchasing Behavior	Order Level	Measure	Numeric
Order Recency	Purchasing Behavior	Order Level	Measure	Numeric
Ordered Items [#]	Purchasing Behavior	Item Level	Measure	Numeric
Ordered Items Mean	Purchasing Behavior	Item Level	Measure	Numeric
Ordered Items Std Dev	Purchasing Behavior	Item Level	Measure	Numeric
Orders [#]	Purchasing Behavior	Order Level	Measure	Numeric
Orders During Holiday	Purchasing Behavior	Order Level	Measure	Numeric
Orders with Delivery Issues [%]	User Experience	After Sales	Measure	Numeric
Orders with Experience Issues [%]	User Experience	After Sales	Measure	Numeric
Orders with Food Issues [%]	User Experience	After Sales	Measure	Numeric
Orders with Order Issues [%]	User Experience	After Sales	Measure	Numeric
Partial Lifespan [#]	Customer Profile	Fixed	Measure	Numeric
Pause Anticipation Mean [#]	Purchasing Behavior	Order Level	Measure	Numeric
Pause Anticipation Std Dev [#]	Purchasing Behavior	Order Level	Measure	Numeric
Pause Frenquency Mean	Purchasing Behavior	Order Level	Measure	Numeric
Pause Frenquency Std Dev	Purchasing Behavior	Order Level	Measure	Numeric
Pause Recency	Purchasing Behavior	Order Level	Measure	Numeric
Pauses [#]	Purchasing Behavior	Order Level	Measure	Numeric
Plan Downgrades	Purchasing Behavior	Settings	Measure	Numeric
Plan Size	Customer Profile	Order Level	Dimension	Numeric
Plan upgrades	Purchasing Behavior	Settings	Measure	Numeric



Name	Feature Group	Sub-Group	Type	Sub-Type
Preferences	Customer Profile	Settings	Dimension	Categorical
Preferences Set	Engagement	Web/App Trackin	Measure	Numeric
Premium Items [#]	Purchasing Behavior	Item Level	Measure	Numeric
Previous Shipping Date Event	Purchasing Behavior	Order Level	Dimension	Categorical
Promo Type Of Use	Engagement	Offer	Dimension	Categorical
Promotional Credits Given [#]	Engagement	Actions Given	Measure	Numeric
Promotional Credits Given [$]	Engagement	Actions Given	Measure	Numeric
Recommendations Accepted [#]	Purchasing Behavior	Order Level	Measure	Numeric
Recommendations Accepted [#]	Purchasing Behavior	Order Level	Measure	Numeric
Recommendations Changed	Purchasing Behavior	Order Level	Measure	Numeric
Recommendations Changed %	Purchasing Behavior	Order Level	Measure	Numeric
Recommendations Received [#]	Purchasing Behavior	Order Level	Measure	Numeric
Redeemed Credits [#]	Purchasing Behavior	Order Level	Measure	Numeric
Redeemed Credits [$]	Purchasing Behavior	Order Level	Measure	Numeric
Redeemed Discounts [#]	Purchasing Behavior	Order Level	Measure	Numeric
Redeemed Discounts [$]	Purchasing Behavior	Order Level	Measure	Numeric
Referrals	User Experience	After Sales	Measure	Numeric
Rejected Orders [#]	Purchasing Behavior	Order Level	Measure	Numeric
Rev Per Item [$]	Purchasing Behavior	Item Level	Measure	Numeric
Reviews [#]	User Experience	After Sales	Measure	Numeric
Reviews Avg Score [#]	User Experience	After Sales	Measure	Numeric
Reviews per Items Ordered	User Experience	After Sales	Measure	Numeric



Aa Name	Feature Group	Sub-Group	Type	Sub-Type
Recommendations Changed %	Purchasing Behavior	Order Level	Measure	Numeric
Recommendations Received [#]	Purchasing Behavior	Order Level	Measure	Numeric
Redeemed Credits [#]	Purchasing Behavior	Order Level	Measure	Numeric
Redeemed Credits [$]	Purchasing Behavior	Order Level	Measure	Numeric
Redeemed Discounts [#]	Purchasing Behavior	Order Level	Measure	Numeric
Redeemed Discounts [$]	Purchasing Behavior	Order Level	Measure	Numeric
Referrals	User Experience	After Sales	Measure	Numeric
Rejected Orders [#]	Purchasing Behavior	Order Level	Measure	Numeric
Rev Per Item [$]	Purchasing Behavior	Item Level	Measure	Numeric
Reviews [#]	User Experience	After Sales	Measure	Numeric
Reviews Avg Score [#]	User Experience	After Sales	Measure	Numeric
Reviews per Items Ordered	User Experience	After Sales	Measure	Numeric
Seasonality	Purchasing Behavior	Order Level	Measure	Numeric
Sessions by Device	Engagement	Web/App Trackin	Measure	Numeric
Shipping Date	Scoring Purposes	Settings	Dimension	Date
Shipping WeekDay	Customer Profile	Settings	Dimension	Categorical
Skip Anticipation Mean [#]	Purchasing Behavior	Order Level	Measure	Numeric
Skip Anticipation Std Dev [#]	Purchasing Behavior	Order Level	Measure	Numeric
Skip Frenquency Mean	Purchasing Behavior	Order Level	Measure	Numeric
Skip Frenquency Std Dev	Purchasing Behavior	Order Level	Measure	Numeric
Skip Recency	Purchasing Behavior	Order Level	Measure	Numeric
Skips [#]	Purchasing Behavior	Order Level	Measure	Numeric
Store Name	Customer Profile	Fixed	Dimension	Categorical
Subscriptions [#]	Customer Profile	Order Level	Measure	Numeric
Weeks To End Pause	Customer Profile	Settings	Measure	Numeric