**Project Part 1**

**Assignment 7.11**

July 29, 2024

Aaron Vo

BIA-678 Big Data Technologies

Stevens Institute of Technology – Summer 2024

Table of Contents

[1 Midterm Exam 4](#_Toc173143034)

[1.1 UVA CASE STUDY QUESTIONS 4](#_Toc173143035)

[1.1.1 What were some of the main challenges and issues Instacart faced with their previous item availability prediction model and system? 4](#_Toc173143036)

[Instacart faced several challenges and issues with their previous item availability prediction model and system. The previous model lacked interpretability, making it difficult to understand why the model made specific predictions. This was problematic, especially when the model made incorrect predictions about item availability. Reliable and understandable predictions are essential for taking appropriate follow-up actions, such as making more observations in the store or informing retailers to replenish stock. Another challenge that they faced was with sparse data. With the growth of Instacart's catalog during the pandemic, data sparsity became more pronounced. Although there were millions of signals from shoppers, the catalog size grew significantly, resulting in many items having limited or no shopper signals. This made it challenging to predict availability accurately for many items. Instacart needed to generate new features and revisit existing ones to improve prediction accuracy. 4](#_Toc173143037)

[Furthermore, the previous model also did not adequately support the different use case scenarios, such as instant delivery and scheduled delivery. Customers increasingly planned, placing orders at night for delivery the next day. The previous model provided a single set of scores, leading to situations where items shown as out-of-stock at order time might be restocked by delivery time. The new pipeline needed to be context-aware and support multiple versions of scores to cater to different use cases. 4](#_Toc173143038)

[Finally, Instacart was also equipped with legacy infrastructure and was not able to handle the rapid growth of the catalog efficiently. Additionally, there was a delay of several hours in signals from the source to the model serving, leading to stale scores. The need for a new infrastructure and pipeline, using the latest technical stack like Griffin, was critical to reduce costs and improve the quality and freshness of scores. 4](#_Toc173143039)

[1.1.2 What are the three hierarchical components of Instacart's new real-time item availability prediction model and what does each component aim to capture? 4](#_Toc173143040)

[In order to address the challenges around interpretability and data sparsity, Instacart had to approach it with a new model. The 3 components of the hierarchical components of Instacarts’ new model are G-T-R. G stands for General availability, which represents typical item availability patterns. It addresses the long-term pattern of availability for an item over 7–180 days and helps solve the data sparsity challenge. T stands for Trending, and it quantifies the deviations from the general baseline. It helps Instacart identify the observed availability of a product over the last 0.5-30 days. Finally, R stands for Real-Time and it’s responsible for the real time observation that is relevant to the short-term availability status. For instance, it helps Instacart recognize patterns of the shopper data and retailer inventory daily. 5](#_Toc173143041)

[1.1.3 How does Instacart's new model try to address the problem of sparse data or limited observations for predicting availability of less frequently purchased catalog items? 5](#_Toc173143042)

[Instacart's new model addresses the problem of sparse data or limited observations for predicting the availability of less frequently purchased catalog items through a methodical sampling algorithm. The model prioritizes data that is both local and recent, as these data points are more relevant for scoring purposes. This ensures that the model uses the most pertinent and timely information available. The model employs an algorithm to identify the "most relevant sample of data." This involves searching for data points from the same product and similar items within a specific area over a certain period. By doing so, the model can draw from a broader but still relevant dataset. Furthermore, for frequently purchased items, a smaller, more localized time frame is sufficient to gather the necessary observations. However, for less frequently purchased items, the model extends the time frame and geographic scope to accumulate the required sample size. For instance, it might use data from the past 180 days across nationwide stores for tail items. 5](#_Toc173143043)

[1.1.4 What is the difference between the "general availability" and "trending" components of the new model? How do they interact? 5](#_Toc173143044)

[General availability provides a baseline probability based on long-term patterns of item availability. This component uses a large sample of observations aggregated from local and recent data, and if necessary, extends to a broader time and geographic scope to ensure statistical reliability. On the other hand, “Trending” detects and quantify near-term deviations from the long-term patterns, especially in response to recent events. This helps Instacart understand the impact of sudden change in the demand or supply. The interaction between these components ensures that the model can make both reliable and responsive predictions, providing a robust and interpretable solution for item availability prediction. 6](#_Toc173143045)

[1.1.5 How does the stratified scoring frequency structure for head, torso, and tail catalog items help reduce computational costs for the system? 6](#_Toc173143046)

[The stratified scoring frequency structure for head, torso, and tail catalog items helps reduce computational costs for the system by adjusting the scoring frequency based on the purchase frequency of the items. Head items, which are purchased frequently, receive updates every hour with real-time inference, ensuring timely and accurate availability predictions. In contrast, torso items, with moderate purchase frequency, and tail items, which are rarely purchased, are updated daily. This approach ensures that computational resources are focused where they are most needed, significantly reducing the load for less frequently purchased items. 6](#_Toc173143047)

[By applying this targeted strategy, the system achieves approximately 80% cost savings. The reduced frequency of updates for torso and tail items prevents unnecessary computational effort, optimizing overall resource use. This not only cuts operational costs but also improves system performance, allowing Instacart to manage a vast catalog efficiently while maintaining high-quality predictions for all items. 6](#_Toc173143048)

[1.1.6 Why was cross-functional collaboration across data, machine learning, infrastructure, and product teams important for tackling the inventory prediction challenge faced by Instacart? 6](#_Toc173143049)

[Cross-functional collaboration across data, machine learning, infrastructure, and product teams are crucial for tackling the inventory prediction challenge faced by Instacart. Each team brought unique expertise and resources essential for different aspects of the project. The data team provided the necessary data engineering capabilities to handle and preprocess vast amounts of data efficiently. Their work ensured that high-quality data was available for analysis and model training. The machine learning team focused on developing and refining the predictive models. Their efforts in algorithm design and model training were pivotal in creating accurate and interpretable availability predictions. Meanwhile, the infrastructure team was instrumental in implementing the technical stack and MLOps tools required to deploy and maintain the models in a scalable and efficient manner. Their contributions ensured that the models could run in real-time, handling the high volume of transactions and updates required by Instacart's operations. The product team played a key role in aligning the technical solutions with business needs, ensuring that the models and systems developed were practical and addressed real-world use cases. By collaborating closely, these teams could integrate their specialized knowledge, leading to a robust, cost-effective solution that improved prediction accuracy and operational efficiency. This synergy between different functions enabled Instacart to successfully address the inventory prediction challenge, ultimately enhancing customer experience and supporting retailer partners more effectively. 7](#_Toc173143050)

# Final Project part 1

## UVA CASE STUDY QUESTIONS

### What were some of the main challenges and issues Instacart faced with their previous item availability prediction model and system?

### Instacart faced several challenges and issues with their previous item availability prediction model and system. The previous model lacked interpretability, making it difficult to understand why the model made specific predictions. This was problematic, especially when the model made incorrect predictions about item availability. Reliable and understandable predictions are essential for taking appropriate follow-up actions, such as making more observations in the store or informing retailers to replenish stock. Another challenge that they faced was with sparse data. With the growth of Instacart's catalog during the pandemic, data sparsity became more pronounced. Although there were millions of signals from shoppers, the catalog size grew significantly, resulting in many items having limited or no shopper signals. This made it challenging to predict availability accurately for many items. Instacart needed to generate new features and revisit existing ones to improve prediction accuracy.

### Furthermore, the previous model also did not adequately support the different use case scenarios, such as instant delivery and scheduled delivery. Customers increasingly planned, placing orders at night for delivery the next day. The previous model provided a single set of scores, leading to situations where items shown as out-of-stock at order time might be restocked by delivery time. The new pipeline needed to be context-aware and support multiple versions of scores to cater to different use cases.

### Finally, Instacart was also equipped with legacy infrastructure and was not able to handle the rapid growth of the catalog efficiently. Additionally, there was a delay of several hours in signals from the source to the model serving, leading to stale scores. The need for a new infrastructure and pipeline, using the latest technical stack like Griffin, was critical to reduce costs and improve the quality and freshness of scores.

### What are the three hierarchical components of Instacart's new real-time item availability prediction model and what does each component aim to capture?

### In order to address the challenges around interpretability and data sparsity, Instacart had to approach it with a new model. The 3 components of the hierarchical components of Instacarts’ new model are G-T-R. G stands for General availability, which represents typical item availability patterns. It addresses the long-term pattern of availability for an item over 7–180 days and helps solve the data sparsity challenge. T stands for Trending, and it quantifies the deviations from the general baseline. It helps Instacart identify the observed availability of a product over the last 0.5-30 days. Finally, R stands for Real-Time and it’s responsible for the real time observation that is relevant to the short-term availability status. For instance, it helps Instacart recognize patterns of the shopper data and retailer inventory daily.

### How does Instacart's new model try to address the problem of sparse data or limited observations for predicting availability of less frequently purchased catalog items?

### Instacart's new model addresses the problem of sparse data or limited observations for predicting the availability of less frequently purchased catalog items through a methodical sampling algorithm. The model prioritizes data that is both local and recent, as these data points are more relevant for scoring purposes. This ensures that the model uses the most pertinent and timely information available. The model employs an algorithm to identify the "most relevant sample of data." This involves searching for data points from the same product and similar items within a specific area over a certain period. By doing so, the model can draw from a broader but still relevant dataset. Furthermore, for frequently purchased items, a smaller, more localized time frame is sufficient to gather the necessary observations. However, for less frequently purchased items, the model extends the time frame and geographic scope to accumulate the required sample size. For instance, it might use data from the past 180 days across nationwide stores for tail items.

### What is the difference between the "general availability" and "trending" components of the new model? How do they interact?

### General availability provides a baseline probability based on long-term patterns of item availability. This component uses a large sample of observations aggregated from local and recent data, and if necessary, extends to a broader time and geographic scope to ensure statistical reliability. On the other hand, “Trending” detects and quantify near-term deviations from the long-term patterns, especially in response to recent events. This helps Instacart understand the impact of sudden change in the demand or supply. The interaction between these components ensures that the model can make both reliable and responsive predictions, providing a robust and interpretable solution for item availability prediction.

### How does the stratified scoring frequency structure for head, torso, and tail catalog items help reduce computational costs for the system?

### The stratified scoring frequency structure for head, torso, and tail catalog items helps reduce computational costs for the system by adjusting the scoring frequency based on the purchase frequency of the items. Head items, which are purchased frequently, receive updates every hour with real-time inference, ensuring timely and accurate availability predictions. In contrast, torso items, with moderate purchase frequency, and tail items, which are rarely purchased, are updated daily. This approach ensures that computational resources are focused where they are most needed, significantly reducing the load for less frequently purchased items.

### By applying this targeted strategy, the system achieves approximately 80% cost savings. The reduced frequency of updates for torso and tail items prevents unnecessary computational effort, optimizing overall resource use. This not only cuts operational costs but also improves system performance, allowing Instacart to manage a vast catalog efficiently while maintaining high-quality predictions for all items.

### Why was cross-functional collaboration across data, machine learning, infrastructure, and product teams important for tackling the inventory prediction challenge faced by Instacart?

### Cross-functional collaboration across data, machine learning, infrastructure, and product teams are crucial for tackling the inventory prediction challenge faced by Instacart. Each team brought unique expertise and resources essential for different aspects of the project. The data team provided the necessary data engineering capabilities to handle and preprocess vast amounts of data efficiently. Their work ensured that high-quality data was available for analysis and model training. The machine learning team focused on developing and refining the predictive models. Their efforts in algorithm design and model training were pivotal in creating accurate and interpretable availability predictions. Meanwhile, the infrastructure team was instrumental in implementing the technical stack and MLOps tools required to deploy and maintain the models in a scalable and efficient manner. Their contributions ensured that the models could run in real-time, handling the high volume of transactions and updates required by Instacart's operations. The product team played a key role in aligning the technical solutions with business needs, ensuring that the models and systems developed were practical and addressed real-world use cases. By collaborating closely, these teams could integrate their specialized knowledge, leading to a robust, cost-effective solution that improved prediction accuracy and operational efficiency. This synergy between different functions enabled Instacart to successfully address the inventory prediction challenge, ultimately enhancing customer experience and supporting retailer partners more effectively.

References

He, Jack. "How Instacart Modernized the Prediction of Real Time Availability for Hundreds of Millions of

Items While Saving Costs." Tech-at-Instacart, Medium, 17 July 2023. Accessed 29 July 2024.