Final Report

Introduction Problem Definition Methods Results and Discussion References Gantt Chart Member Contribut

Introduction

- Understanding and predicting customer behavior is essential for e-commerce companies to stay competitive. By analyzing vast amounts of user interaction data, machine learning can reveal customer preferences, predict future actions, and deliver personalized product recommendations.
- This project will apply ML techniques to data from an online electronics store, aiming to predict customer behavior and segment users based on their activities, such as browsing or purchasing.
- These insights could help businesses make informed decisions about marketing strategies, inventory management, and resource allocation, ultimately improving customer satisfaction and operational efficiency.
- In a broader societal context, these advancements can improve the consumer shopping experience and contribute to sustainability by optimizing resource usage.

Dataset Description 📊

The dataset contains behavior data for **5 months** (October 2019 – February 2020) from a large **electronics online store**. It consists of a total of **845,041 unique events**.

- Each row in the dataset represents an event.
- All events are related to products and users.
- Each event acts as a **many-to-many relation** between products and users.
- Some of the features include the event type (view, add to cart, purchase), product ID, category ID, brand, price, user ID, and user session.

Dataset Link

	event_time	event_type	product_id	category_id	category_code
0	2020-09-24 11:57:06 UTC	view	1,996,170	⚠ 2144420000000000000	electronics.telephone
1	2020-09-24 11:57:26 UTC	view	139,905	⚠ 2144420000000000000	computers.components.c
2	2020-09-24 11:57:27 UTC	view	215,454	⚠ 2144420000000000000	None
3	2020-09-24 11:57:33 UTC	view	635,807	⚠ 2144420000000000000	computers.peripherals.pri
4	2020-09-24 11:57:36 UTC	view	3,658,723	⚠ 2144420000000000000	None
5	2020-09-24 11:57:59 UTC	view	664,325	⚠ 2144420000000000000	construction.tools.saw
6	2020-09-24 11:58:23 UTC	view	3,791,349	⚠ 2144420000000000000	computers.desktop
7	2020-09-24 11:58:24 UTC	view	716,611	⚠ 2144420000000000000	computers.network.route
8	2020-09-24 11:58:25 UTC	view	657,859	⚠ 2144420000000000000	None
9	2020-09-24 11:58:31 UTC	view	716,611	⚠ 2144420000000000000	computers.network.route

Literature Review 📚

- Haque applied four ML algorithms (Gaussian Naive Bayes, Random Forest, Logistic Regression, Decision Tree) to e-commerce product recommendation [1]. The performance was evaluated using metrics like accuracy, precision, recall, and F1-score, with Random Forest performing best overall.
- Nguyen et al. proposed a personalized product recommendation model based on a retrieval strategy, combining collaborative filtering and content-based filtering techniques [2]. Their two-stage approach, involving candidate generation and ranking, showed improved performance over baseline methods when evaluated on real-world datasets.
- Loukili et al developed an e-commerce recommendation system using the FP-Growth algorithm to generate association rules based on customer purchase history [3]. This approach effectively addresses challenges like data sparsity and cold-start problems, showing promising results in predicting future purchases and improving recommendation accuracy.

Final Report

Introduction Problem Definition Methods Results and Discussion References Gantt Chart Member Contribut

Problem Definition 🔎

Problem ❓

The online electronics store currently lacks insights into their customer base, leading to inefficiencies throughout the customer journey—from product exploration to potential purchase. Without this understanding, it's difficult for the store to optimize their strategies, resulting in missed opportunities to improve customer engagement and increase sales.

Motivation 🎯

We recognized that online stores could greatly benefit from a **data-driven characterization** of their customer base. By focusing on online electronics stores, we aim to use **machine learning** to predict and segment online customer behavior on the e-commerce site.

This approach will provide valuable insights, allowing the store to enhance customer experience and improve overall business efficiency.

Final Report

Introduction Problem Definition Methods Results and Discussion References Gantt Chart Member Contribut

Methods 💻

Data Preprocessing Methods 🔧

We identified **3+ data preprocessing methods** to prepare the dataset for analysis:

1. **Feature Engineering**: Define and create features that are suitable for the clustering solution.
2. **Group Data**: Group data by **user-product-score** for more effective analysis.
3. **Data Cleaning**: Handle inconsistent data by addressing **null values** and **duplicates**.

Supervised and Unsupervised Learning Methods 💡

We explored different methods. Below are the identified methods, with both supervised and unsupervised learning** approaches. For this phase of the project, we are implementing **K-Means**:

Final Report

duction Problem Definition Methods Results and Discussion References Gantt Chart Member Contributions

Results & Discussions 📝

Data Preprocessing Methods

KMeans Clustering

Feature Engineering

We created a new feature called `total_interaction_score`, combining user interactions (purchases, view, cart) to quantify user engagement levels.
Each interaction was assigned a specific weight based on its importance:

- **Purchase**: Weight 3
- **Add to Cart**: Weight 2
- **View**: Weight 1

Grouping Data By User

We grouped data at a user level, aggregating interaction scores to obtain a holistic view of each user's behavior across interactions.
This approach made it easier to analyze users rather than individual transactions, allowing the clustering model to focus on long-term user habits rather than one-time actions.

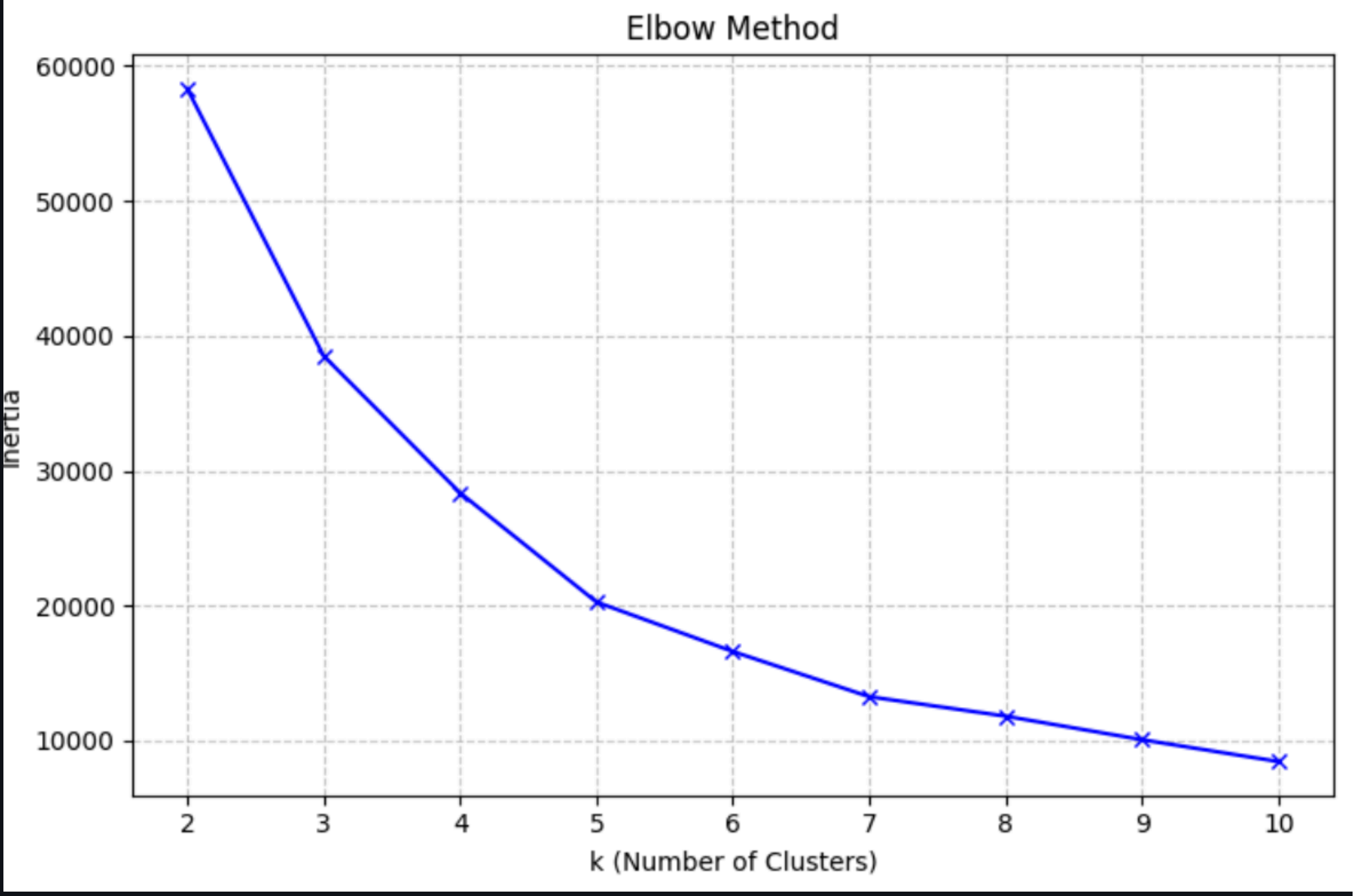
Data Cleaning

- Addressed inconsistencies in the dataset by handling null values and duplicates.

Determining Optimal Number of Clusters

Elbow Method

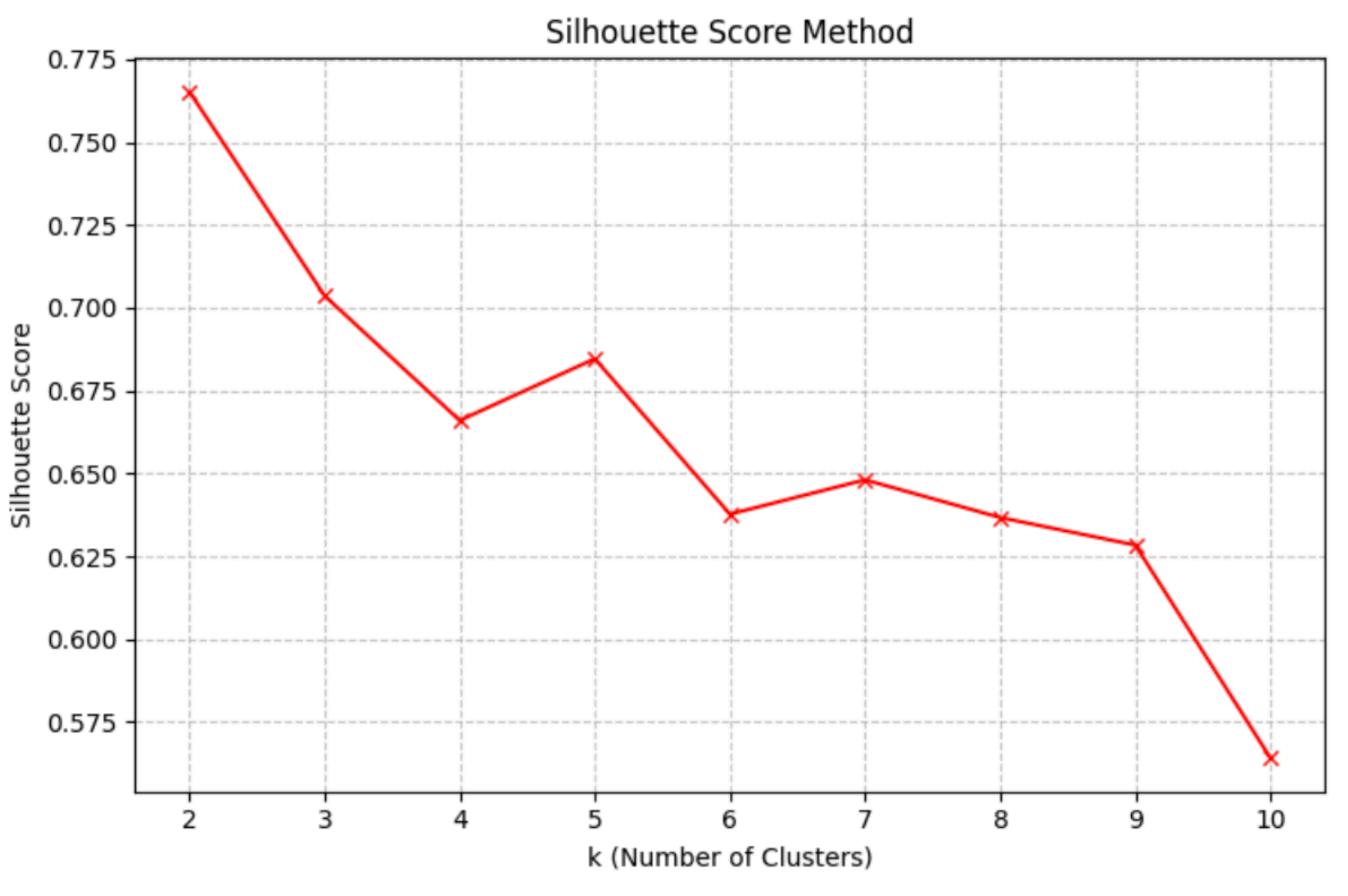
- Shows changes in **inertia** (sum of squared distances between data points and their respective cluster centroids) depending on the number of clusters used.
- Significant drop in inertia up to **k = 5**, indicating 5 clusters may be optimal.



Elbow Method

Silhouette Score Method

- The silhouette score plot shows at which points clusters are more cohesive and separated.
- A higher silhouette score means clusters are better defined, and points within a cluster are more similar to each other.
- While **k = 2** has the highest score, **k = 5** also provides a reasonable score and captures more granular user segmentation, making it an ideal candidate.
- After this point, silhouette scores decrease.



Silhouette Score

Conclusion

Based on the outputs provided by the Elbow Method and Silhouette Method, the model performs optimally with **5 clusters**.
This choice balances cohesiveness and compactness of clusters while capturing patterns in user behavior.

Quantitative Metrics

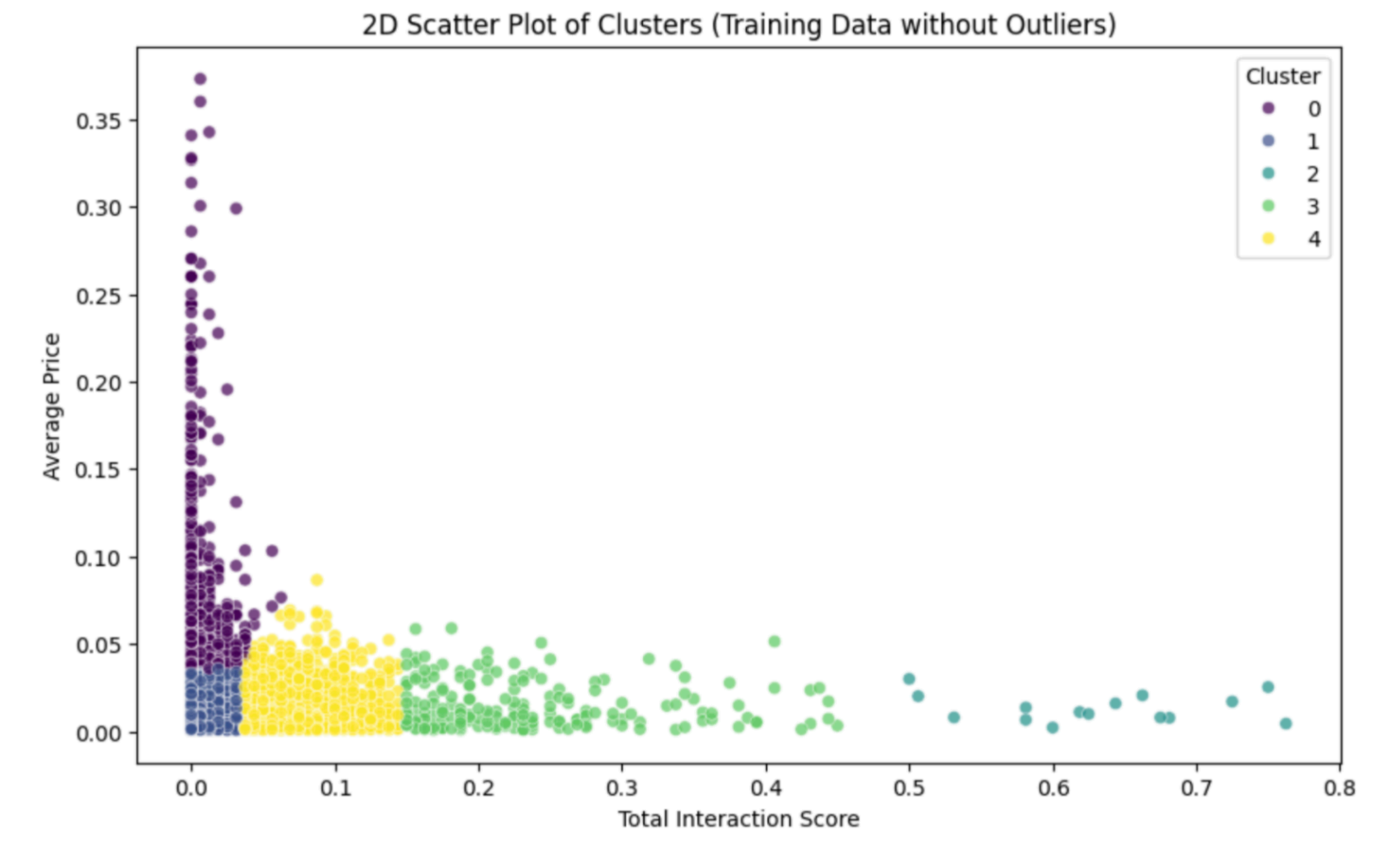
- **Training Set Silhouette Score**: 0.717
  Indicates users are grouped with other similar users and well-separated from different user groups.
- **Test Set Silhouette Score**: 0.720
  Demonstrates consistent model performance on new data.
  These similar scores indicate good generalization and well-defined, separate clusters.

2D Scatter Plot of Clusters

The scatter plot displays the clusters formed by the **k-means algorithm**.

- Y-axis: `total_interaction_score`
- X-axis: `avg_price`

  Filtered extreme values and outliers for better visualization, centering on important data points.



Scatter Plot of Clusters

User Segments

1. **Premium Buyers (Purple - Cluster 0):**
   Users with higher average prices but lower interaction scores. Likely occasional, high-value buyers.
2. **Basic Browsers (Blue - Cluster 1):**
   Users with low average prices and interaction scores. Likely new or occasional customers.
3. **Super Users (Turquoise - Cluster 2):**
   Users with the highest interaction scores but low-medium prices. Make frequent, lower-value purchases.
4. **Engaged Regulars (Green - Cluster 3):**
   Users with moderate prices and higher interaction scores. Likely regular customers with frequent engagement.
5. **Value Seekers (Yellow - Cluster 4):**
   Users with low-medium prices and interaction scores. Likely price-conscious, careful buyers.

Model Strengths

- **Consistent performance across training and test sets** with high silhouette scores (both above 0.7), indicating well-separated, distinct clusters.
- Captures diverse shopping behaviors and spending patterns. The scatter plot reveals distinct groupings of users.

Reasons for Success

1. **Good Feature Engineering**:

- Used `total_interaction_score` weighted by action type.
- Normalized prices to handle different price ranges.
  These features created meaningful distinctions in user behavior.

2. **Effective Data Preprocessing**:

- Outlier removal using quartile filtering (2nd to 99.99th percentile).
- Proper scaling of features to prevent distortion by extreme values.

3. **Appropriate Number of Clusters**:

- **k = 5** balances detail and interpretability.
- Supported by both Elbow and Silhouette methods.

ALS Algorithm

Data Preprocessing Methods

- **Grouping Data By User**:
  Aggregated interaction scores to obtain a holistic view of each user's behavior across interactions. This approach made it easier to analyze users rather than individual transactions, allowing the clustering model to focus on long-term user habits rather than one-time actions.
- **Data Cleaning**:
  Addressed inconsistencies in the dataset by handling null values and duplicates. Filtered unknown categories.
- **Index encoding for ALS Compatibility**:
  Converted categorical variables into integer indices.

Facilitated the creation of a sparse matrix suitable for the ALS algorithm.

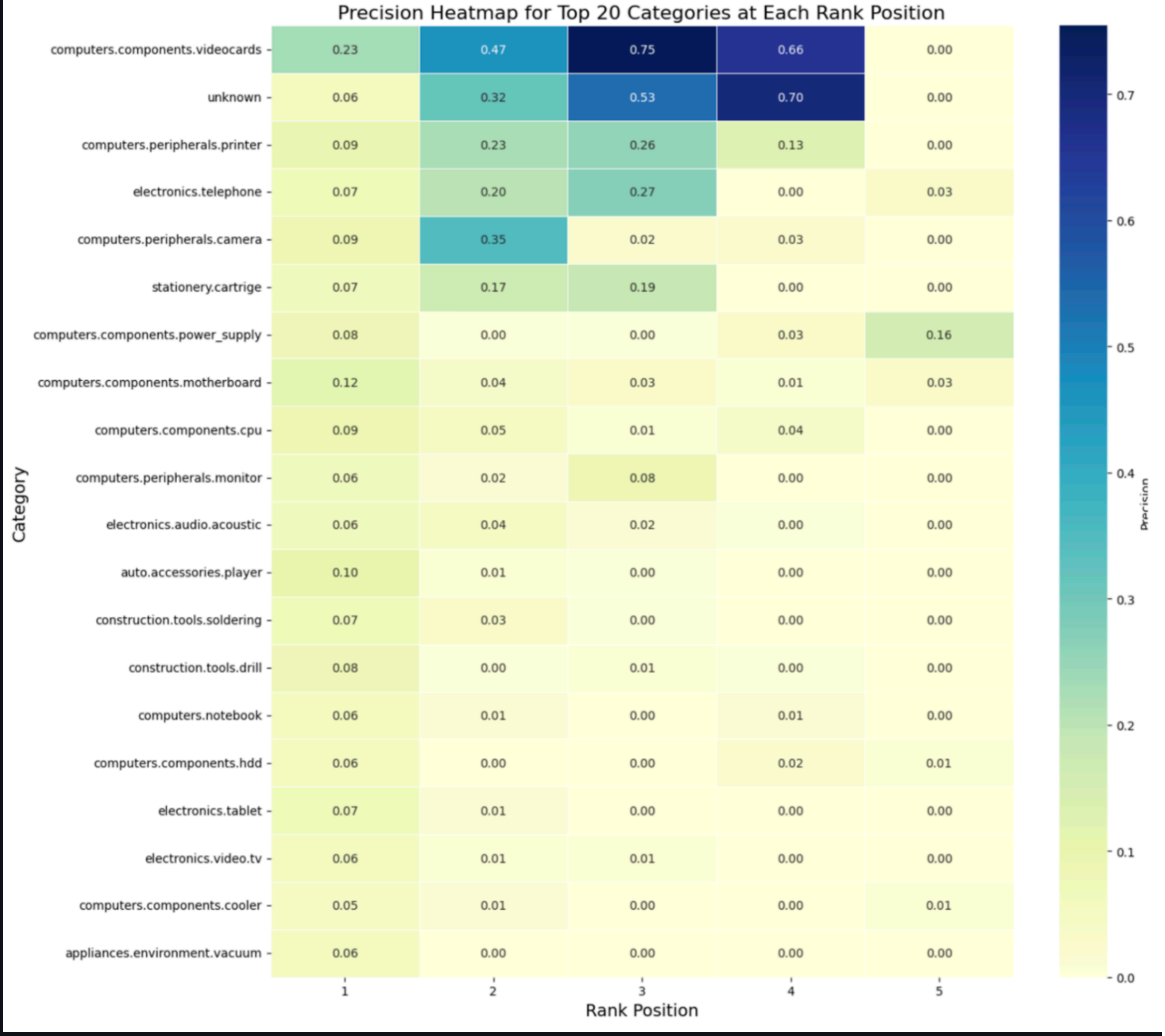
- **Sparse Matrix Generation**:
  Constructed a sparse matrix to optimize ALS performance.
  Leveraged efficient computation by representing the data with many zeroes.
- **BM25 weighting for normalization**:
  Applied BM25 weighting to normalize interaction scores.
  Mitigated the disproportionate influence of frequently interacted categories, ensuring balanced impact across all categories.

Quantitative Metrics

- **Precision@2**: 0.4624
  Indicates the top two recommended items to users will be accurately recommended mostly half of the time.
- **Precision@3**: 0.3083
- **Precision@5**: 0.1850
- **Recall@2**: 0.8761
  Indicates when we recommend two items to the users almost 9 out of 10 times those items are relevant to the users.
  These high scores indicate accurate recommendations to users.

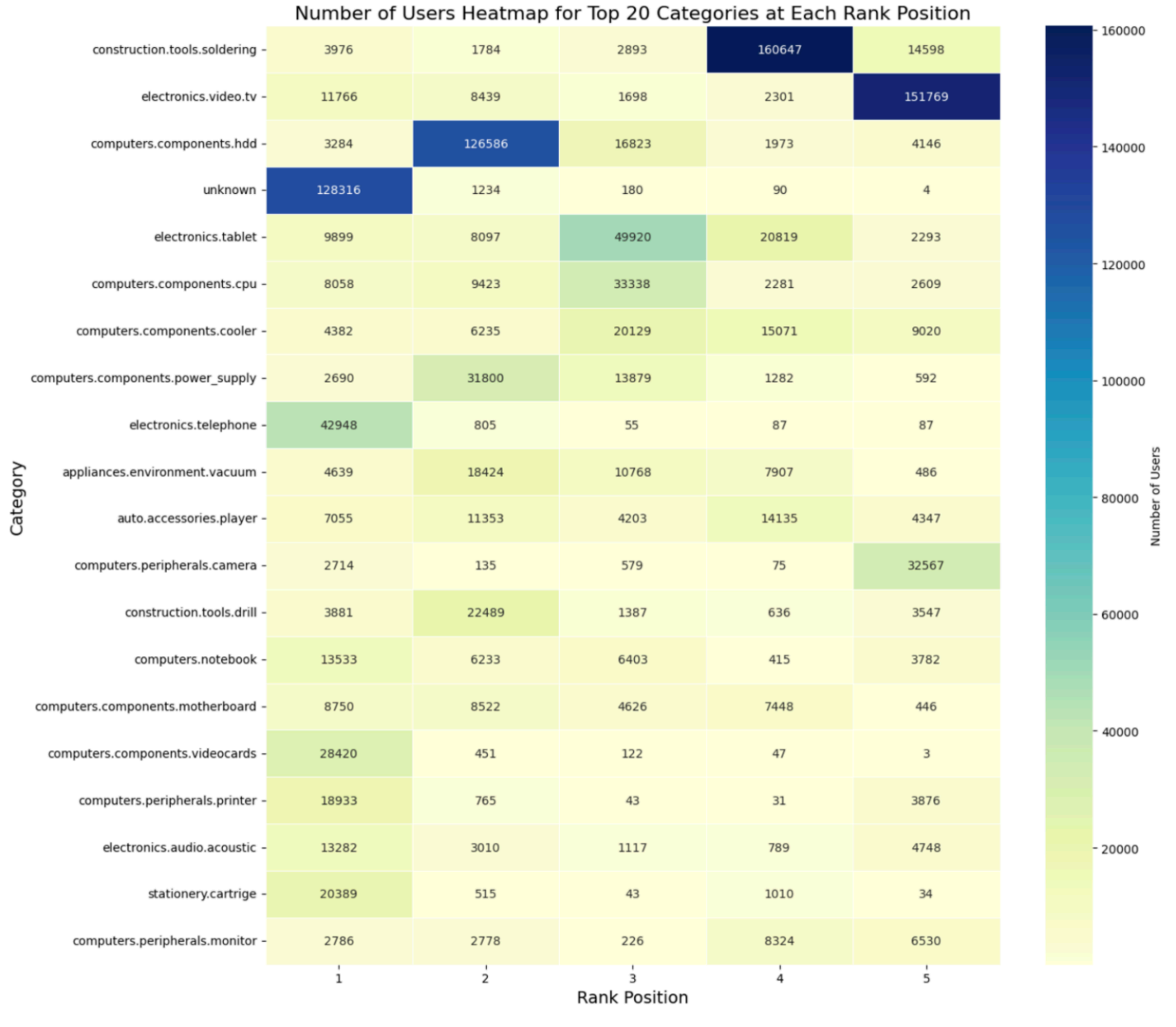
Visualizations

- **Precision Heat Map**:
- Illustrates the average Precision@5 for each category-ranking pair.
- Top 20 categories with the highest number of interactions are filtered for clarity.
- Darker colors indicate higher precision.



Precision heat map

- **User Amount Heat Map:**
- Displays the number of users for each category-ranking pair.
- Top 20 categories are filtered for clarity.
- Darker colors indicate a greater number of users.



Category	1	2	3	4	5
construction.tools.soldering	3976	1784	2893	160647	14598
electronics.video.tv	11766	8439	1698	2301	151769
computers.components.hdd	3284	126586	16823	1973	4146
unknown	128316	1234	180	90	4
electronics.tablet	9899	8097	49920	20819	2293
computers.components.cpu	8058	9423	33338	2281	2609
computers.components.cooler	4382	6235	20129	15071	9020
computers.components.power_supply	2690	31800	13879	1282	592
electronics.telephone	42948	805	55	87	87
appliances.environment.vacuum	4639	18424	10768	7907	486
auto.accessories.player	7055	11353	4203	14135	4347
computers.peripherals.camera	2714	135	579	75	32567
construction.tools.drill	3881	22489	1387	636	3547
computers.notebook	13533	6233	6403	415	3782
computers.components.motherboard	8750	8522	4626	7448	446
computers.components.videocards	28420	451	122	47	3
computers.peripherals.printer	18933	765	43	31	3876
electronics.audio.acoustic	13282	3010	1117	789	4748
stationery.cartrige	20389	515	43	1010	34
computers.peripherals.monitor	2786	2778	226	8324	6530

User amount heat map

Both heat maps show darker colors in the higher rankings, indicating that our algorithm effectively and accurately recommends users' preferred categories.

Model Strengths

- **High Precision and Recall Metrics:**
- The model demonstrates strong performance with high precision and recall scores.
- Superior precision in the top 3 recommendations indicates effective prioritization of user preferences.
- **Reasons for Success:**
- Good preprocessing ALS setup.
- Sparse matrix generation for efficient computation.
- BM25 normalization reduced high-impact categories.

- Efficient indexing and filtering of low-user categories.

Supervised Methods

Random Forest

For the supervised methods, we want to predict whether a customer will buy (1) or not (0).

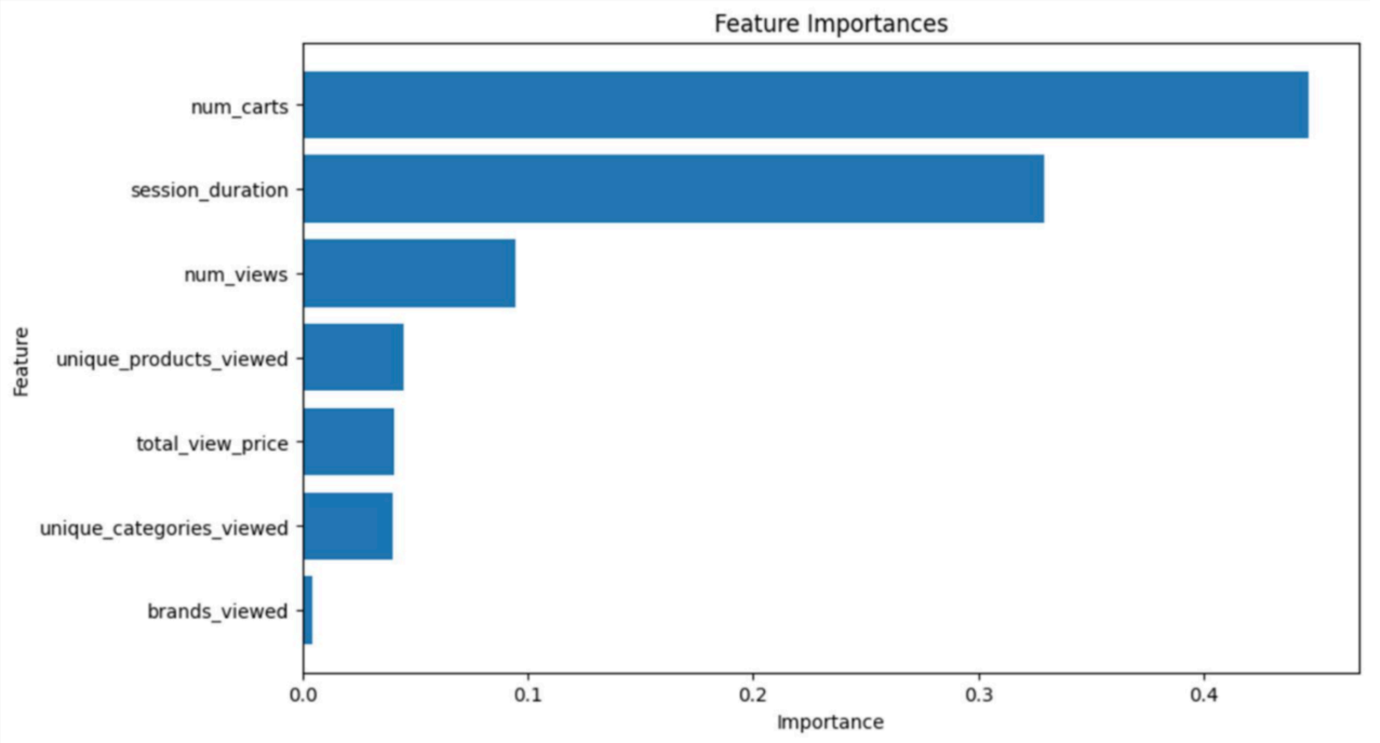
Data Sampling

- Sampled 20% from the total dataset, amounting to 177,572 rows across 98,079 sessions.

Feature Engineering

We tracked user behavior across complete sessions. The top three most important features were:

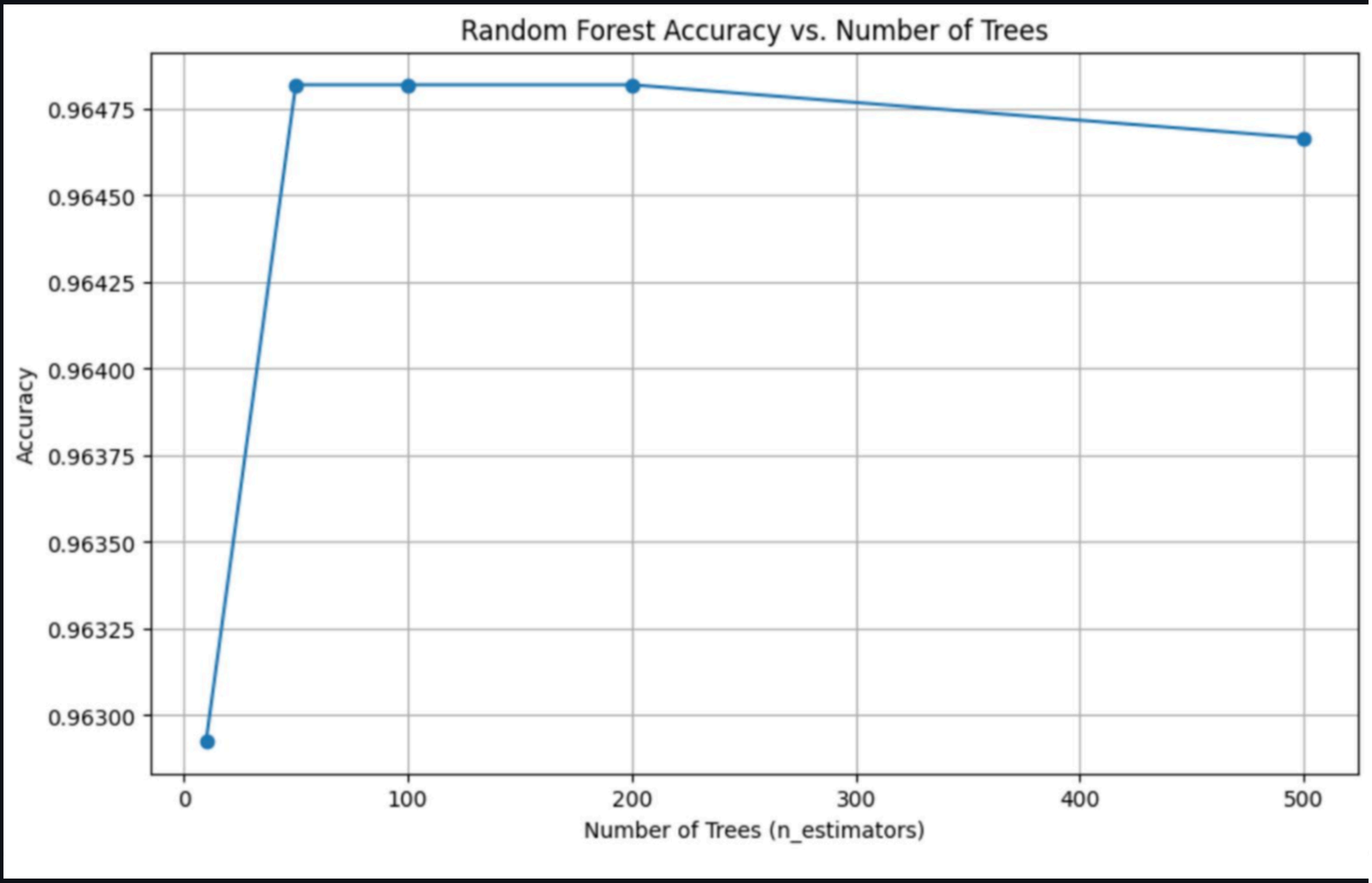
- **Number of items added to cart (num_carts)**: 0.446713
- **Total price of items viewed (total_view_price)**: 0.329522
- **Number of unique categories viewed (unique_categories_viewed)**: 0.094378



Feature Importance

Training & Evaluation

- The model was trained with 80% of the data while 20% was used for testing.
- Parameters:
- Number of trees: 100
- Max depth: 5
- **Accuracy:** 0.9648 on the test set.
- Optimal number of trees was found to be 100, as increasing beyond this showed no significant gains.



Accuracy vs Number of Trees

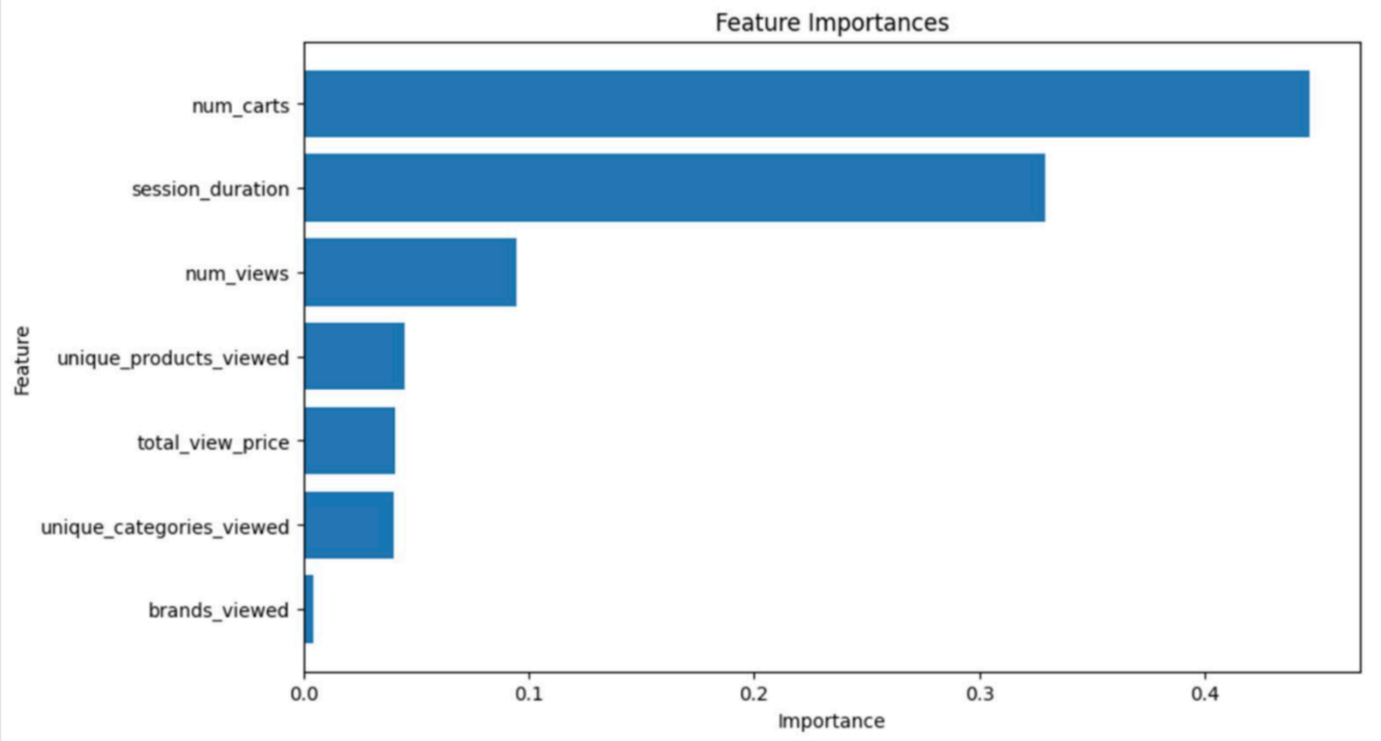
Logistic Regression

Setup & Evaluation

- Set max iterations to 1000 to ensure convergence.
- Accuracy:
- Test: 0.9556
- Mean accuracy across folds: 0.9630

Visualization of Feature Importance

The graph below highlights the dominance of `num_carts` and `session_duration` as the primary predictors of purchase behavior, emphasizing the importance of cart interactions and session duration in influencing purchasing decisions.



Feature Importance

Evaluation

- **Random Forest** slightly outperformed **Logistic Regression** due to its ability to model non-linear relationships.
- Random Forest identified `num_carts` as the most important feature, followed by `total_view_price` and `unique_categories_viewed`.
- Despite the higher accuracy of Random Forest, Logistic Regression offers simplicity with minimal performance trade-offs. Depending on the use case, the simplicity of Logistic Regression can make it an attractive choice, especially given the minimal difference in accuracy.

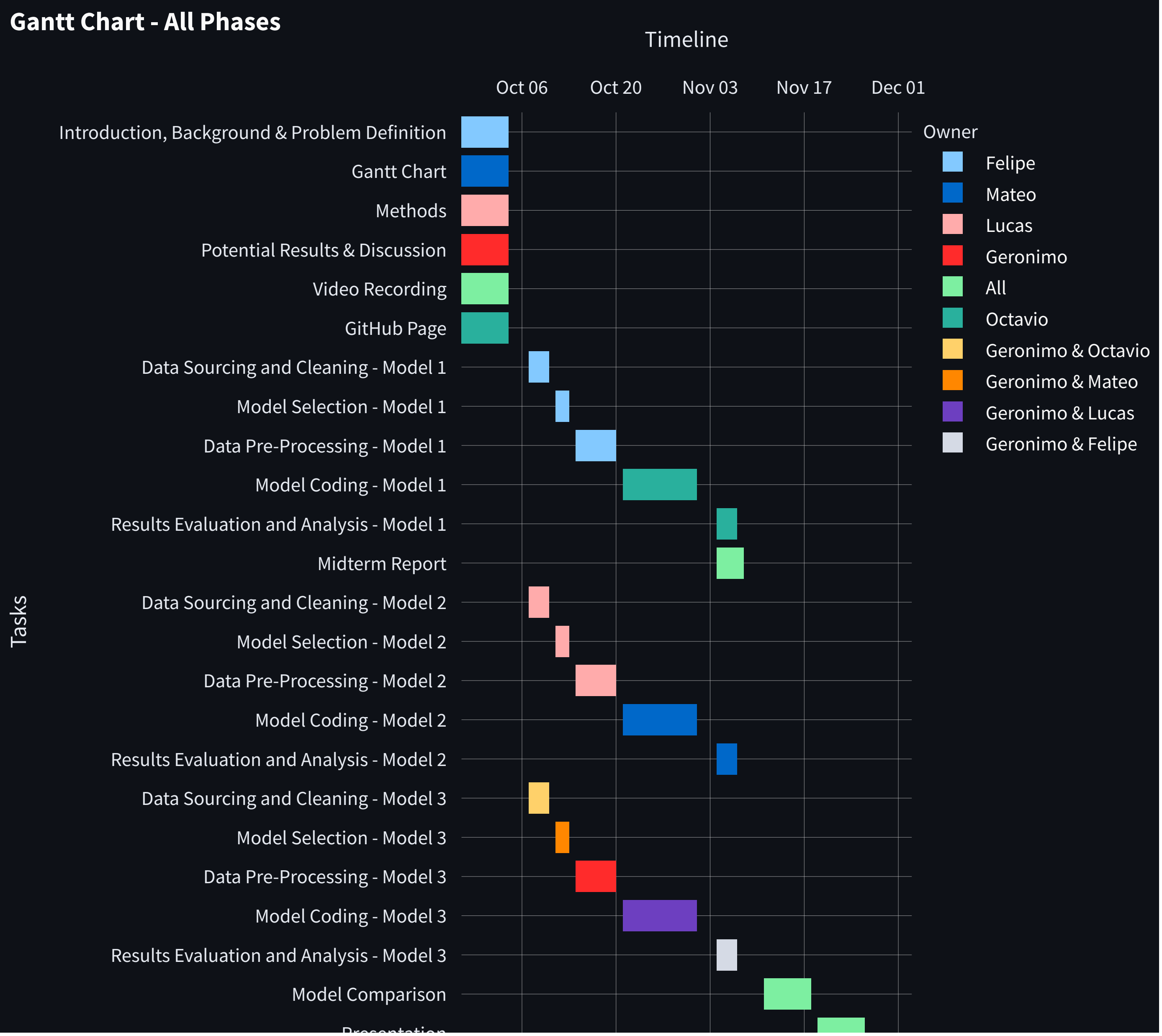
Final Report

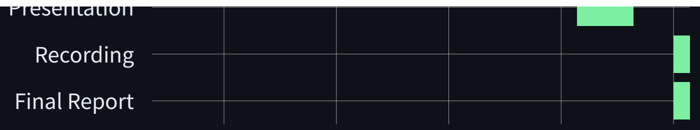
Introduction Problem Definition Methods Results and Discussion References Gantt Chart Member Contribut

Gantt Chart

Select Phase

All Phases




Presentation
Recording
Final Report

Final Report

Introduction Problem Definition Methods Results and Discussion References Gantt Chart Member Contribut

References 📄

1. M. Z. Haque, "E-Commerce Product Recommendation System based on ML Algorithms," *arXiv preprint*, vol. arXiv:2407.21026, 2024.
2. D.-N. Nguyen, V.-H. Nguyen, T. Trinh, T. Ho, and H.-S. Le, "A personalized product recommendation model in e-commerce based on retrieval strategy," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 10, no. 2, p. 100303, 2024.
3. M. Loukili, F. Messaoudi, and M. El Ghazi, "Machine learning based recommender system for e-commerce," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 3, pp. xxx-xxx, 2023.

Final Report

duction Problem Definition Methods Results and Discussion References Gantt Chart Member Contributions

Member Contributions 👥

	Task owner	Task Title
0	Felipe	Data Preprocessing
1	Mateo	K-means Implementation
2	Lucas	Data Preprocessing
3	Geronimo	Results & Discussion
4	Octavio	K-means Implementation