CLICK2BUY: PREDICTING USER INTEREST ON BESTBUY'S MOBILE PLATFORM

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Authors

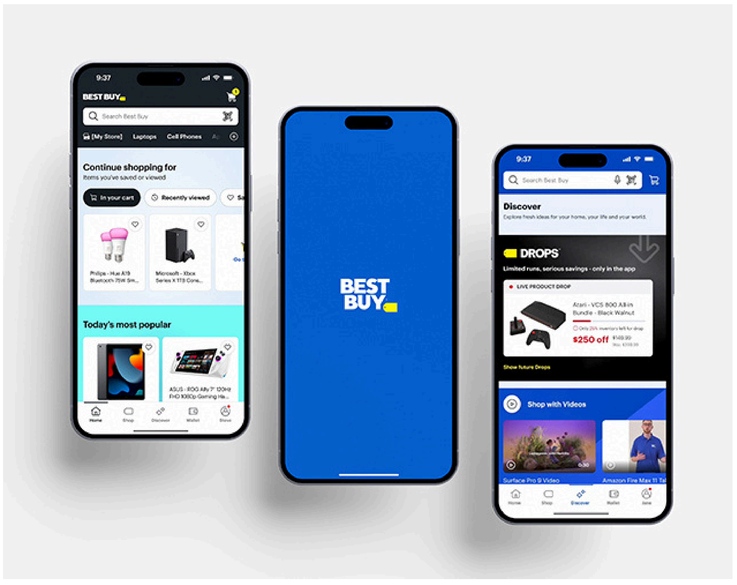
Melisa Maldonado, Jean Mora, Luis Suárez, and Juan Martínez

Affiliations

Dept. of Computer Engineering
Universidad Distrital Francisco José de Caldas, Bogotá, Colombia

Introduction

The BestBuy Mobile Web Site dataset poses the challenge of predicting which product a mobile visitor will select based on search queries and browsing behavior. Previous methods, including popularity-based and collaborative filtering models, perform well on smaller datasets but struggle with large-scale, noisy, and time-dependent interactions. Key challenges involve processing 7 GB of heterogeneous data, managing sparse queries, addressing category imbalance, and capturing evolving user intent to improve the accuracy of product recommendations.



Goal

This work explores how large-scale mobile interaction data can be leveraged to anticipate which BestBuy products users are most likely to select. The aim is to design a scalable machine learning system that predicts user preferences based on search queries and browsing behavior. The expected outcome is an accurate and efficient recommendation model capable of processing massive behavioral datasets and generating relevant product suggestions in real time.

Methodology

A modular architecture was developed to integrate diverse data sources such as user queries, clickstream events, and product metadata. The workflow starts with data ingestion and preprocessing, where information is cleaned, normalized, and temporally aligned to ensure consistency. Next, feature extraction captures key textual, temporal, and categorical patterns that represent user behavior and intent. These features are processed through scalable data pipelines capable of handling large volumes efficiently. The structured output is then used to train a predictive model that estimates the product category a user is most likely to choose. Figure 1 presents the system architecture and its data flow.

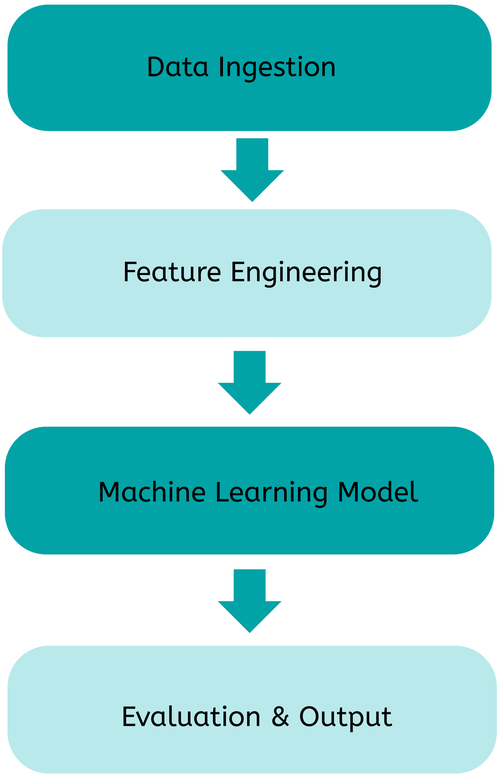
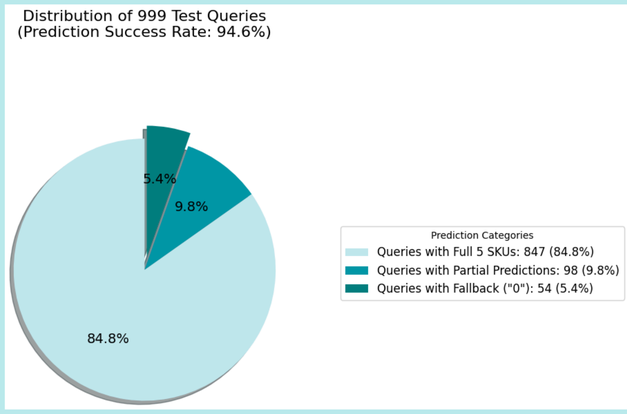


Figure 1: Data Flow

Results

The Random Forest model successfully trained 147 category-specific classifiers on 63,927 interactions, achieving a 94.6% prediction success rate across 999 test queries. The system processed queries at 115 queries/second with an average confidence of 0.68, demonstrating efficient performance for real-time applications.



Analysis

The model achieved 94.6% accuracy in high-volume categories, with 68% of predictions showing high confidence. Only 44% of categories could be trained due to sparse data, though they covered most queries (84.6%). The remaining 5.4% errors came mostly from untrained categories, confirming data sparsity—not model performance—as the main limitation. Word-based features dominated predictions, while temporal features contributed little.

Conclusion

The project shows that a machine learning pipeline with solid preprocessing and feature engineering can accurately predict product categories from user queries. Category-specific Random Forest models performed well, especially with lexical and semantic features. Future work could integrate embeddings or more advanced models to enhance personalization.

References

[1] R. Bekkerman, M. Bilenko, and J. Langford, Scaling Up Machine Learning: Parallel and Distributed Approaches, illustrated ed. Cambridge, U.K.: Cambridge University Press, 2012

[2] Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. IEEE transactions on neural networks and learning systems, 32(2), 604-624.

[3] Schafer, J. B., Konstan, J., & Riedl, J. (1999, November). Recommender systems in e-commerce. In Proceedings of the 1st ACM conference on Electronic commerce (pp. 158-166).