YZV 311E Data Mining Project Proposal

Hasan Taha Bağcı
Artificial Intelligence and Data Engineering
Istanbul Technical University
Istanbul, Turkey
bagcih21@itu.edu.tr

Selman Turan Toker
Artificial Intelligence and Data Engineering
Istanbul Technical University
Istanbul, Turkey
toker22@itu.edu.tr

***Abstract*—This proposal presents a predictive analytics solution for enhancing customer experience on a global e-commerce platform. The objective is to implement a data mining system that predicts both the specific products a customer is likely to repurchase and the timing of these purchases. By employing models such as Pareto/NBD, BG/NBD, RNN with LSTM, and Gradient Boosting Machines, we aim to provide timely and relevant recommendations for essential household items. Evaluation metrics, including MAE, RMSE, precision, and AUC-ROC, will assess model performance. This project will improve personalized recommendations, increasing customer satisfaction and optimizing stock management.**

***Index Terms*—predictive analytics, e-commerce, repurchase prediction, customer experience, machine learning, time-aware recommendations, data mining**

I. Introduction

Consumers all over the world spend a considerable amount of time to fulfill their household needs. However, questions such as which of the dozens of products in the house should be repurchased or which additional products should be purchased keep people's minds busy. With this project, we aim to relieve consumers of this burden and improve their shopping experience. We also aim to facilitate inventory management for companies that provide consumer products.

II. Dataset

The dataset for this project contains information on a total of 32,775 products, sourced from 1,513 manufacturers, across 4,299 categories and 3,898 parent categories. The dataset further includes 46,137 unique customer IDs, 29,136 distinct purchase dates, and details on the number of units purchased per transaction. The data is provided in four CSV files as follows:

- **product_catalog.csv** (size: 1.32 MB)
  This file contains 8 columns: `product_id`, `manufacturer_id`, `attribute_1`, `attribute_2`, `attribute_3`, `attribute_4`, `attribute_5`, and `categories`. The first seven columns show a validity rate of 100% and a mismatch rate of 0%. However, the `categories` column has a 79% validity rate, with 21% null values. Each attribute has distinct ranges of values.
- **product_category_map.csv** (size: 41.22 kB)
  This file contains two columns: `category_id` and `parent_category_id`. Both columns have a 100% validity rate.
- **transactions.csv** (size: 26.18 MB)
  This file includes information in 4 columns: `customer_id`, `product_id`, `purchase_date`, and `quantity`. All columns in this file have a 100% validity rate.
- **test.csv** (size: 173.32 kB)
  This file is intended to be used for testing the model after training. It includes four columns: `id`, `customer_id`, `product_id`, and `prediction`. The `prediction` column will be populated with one of the following values: [0, 1, 2, 3, 4].

Identify applicable funding agency here. If none, delete this.

III. Literature Review

Enhancing customer experience through predictive analytics has become a focal point for e-commerce platforms, especially in predicting repurchase behavior and the timing of replenishment for essential items. The literature offers various methodologies that address these challenges, focusing on machine learning models, time-series analysis, and time-aware recommendation systems.

A. Predictive Modeling of Repurchase Behavior

Fader and Hardie introduced the Pareto/NBD (Negative Binomial Distribution) model, which predicts customer purchasing behavior over time by estimating the probability of repeat purchases [1]. This model utilizes transaction data to forecast future buying patterns, aiding businesses in identifying customers who are likely to make subsequent purchases.

B. Time-Series Analysis and Deep Learning

Hidasi et al. proposed a session-based recommendation approach using Recurrent Neural Networks (RNNs), specifically designed for handling sequential data in e-commerce settings [2]. Their model captures the temporal dynamics of user behavior within sessions, enabling more accurate predictions of not only what customers will buy next but also when they will make the purchase.

C. Time-Aware Recommendation Systems

Koren incorporated temporal dynamics into collaborative filtering models in his seminal work [3]. By acknowledging that user preferences and item popularity change over time,

the model adapts recommendations accordingly. This approach improves the relevance and timeliness of product suggestions, which is crucial for consumable goods that require regular replenishment.

D. Survival Analysis for Purchase Timing

Schmittlein, Morrison, and Colombo developed the BG/NBD (Beta Geometric/Negative Binomial Distribution) model to estimate the timing of customer repeat purchases using survival analysis techniques [4]. This probabilistic model accounts for the time until a customer becomes inactive, providing insights into when they might need to restock products.

E. Machine Learning Techniques for Purchase Prediction

Xue et al. explored the use of Gradient Boosting Machines (GBM) to predict customer repurchase behavior [5]. By integrating various features such as purchase history, product categories, and customer demographics, their model achieved high accuracy in predicting both the likelihood and timing of repeat purchases.

IV. Candidate Models

To predict both products and timing for repurchase, we propose the following models:

1. **Pareto/NBD Model**: Estimates expected repeat purchases over time using customer dropout and purchase rates:

$$\mathrm{P}(X(t)=x)=\int_0^\infty\int_0^\infty \frac{(\lambda t)^x e^{-\lambda t}}{x!}\cdot\frac{\Gamma(r+\alpha)}{\Gamma(\alpha)\Gamma(r)}\lambda^{r-1}(1+\lambda/\beta)^{-(r+\alpha)}\,d\lambda\,d\beta$$

2. **BG/NBD Model**: Extends Pareto/NBD by modeling heterogeneity in purchase frequency.

3. **RNN with LSTM**: Captures temporal dependencies in purchase behavior.

4. **Gradient Boosting Machine (GBM)**: Combines decision trees to predict next purchases.

V. Model Evaluation

Models are evaluated on their ability to predict products and timing. Key metrics include:

For Timing:

- **MAE**:

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|$$

- **RMSE**:

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

For Product Prediction:

- **Precision**, **Recall**, and **F1 Score**.
- **AUC-ROC** for classification.

Cross-validation ensures robustness, and a test set assesses performance.

VI. Time Plan and Roles

- **Weeks 1-2**: **Hasan** - Data collection and cleaning. **Selman** - Exploratory data analysis.
- **Weeks 3-4**: **Hasan** - Implement Pareto/NBD and BG/NBD. **Selman** - Develop RNN and GBM.
- **Week 5**: Joint model evaluation.
- **Week 6**: **Hasan** - Hyperparameter tuning. **Selman** - Model refinement.
- **Week 7**: **Hasan** - Draft methodology and results. **Selman** - Finalize report and prepare presentation.
- **Week 8**: Joint review and submission.

VII. GitHub Repository and Kaggle Competition

This project is part of the **YZV 311E Term Project Fall 24-25** competition. All code, data preprocessing steps, model implementations, and documentation for this project are available in our GitHub repository. The repository is structured to include scripts for data loading, model training, evaluation, and visualization, providing a comprehensive overview of our approach and methodology.

- **GitHub Repository**
- **Kaggle Competition Name**: YZV 311E Term Project Fall 24-25

Please refer to the GitHub repository for detailed instructions on setting up the environment, running the models, and reproducing the results presented in this report.

References

[1] P. S. Fader and B. G. S. Hardie, "A Note on Deriving the Pareto/NBD Model and Related Expressions," *Marketing Science*, vol. 24, no. 3, pp. 408–413, 2005.

[2] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based Recommendations with Recurrent Neural Networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[3] Y. Koren, "Collaborative Filtering with Temporal Dynamics," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 447–456.

[4] D. C. Schmittlein, D. G. Morrison, and R. Colombo, "Counting Your Customers: Who Are They and What Will They Do Next?," *Management Science*, vol. 33, no. 1, pp. 1–24, 1987.

[5] H. Xue, X. Dai, J. Zhang, S. Huang, and J. Chen, "Deep Matrix Factorization Models for Recommender Systems," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 3203–3209.