Supervised Learning on Instacart Dataset: A Comprehensive Model for Product Reorder Prediction

Anjisnu Roy

November 18, 2024

Abstract

This comprehensive study investigates the application of machine learning models to the Instacart dataset for predicting customer product reorder behavior. The research encompasses extensive data preprocessing, feature engineering, and evaluation of multiple classification algorithms through a systematic approach. We implemented a two-phase model selection process: initial screening through cross-validation on a 30% stratified sample, followed by comprehensive evaluation of promising models on the full dataset. Performance assessment utilized multiple metrics including accuracy, precision, recall, F1-score, and ROC-AUC. The study revealed Logistic Regression with L1 regularization as the optimal model, achieving an F1-score of 0.8851 and ROC-AUC of 0.9365, demonstrating superior balanced performance across all metrics while maintaining computational efficiency.

Contents

1 Introduction

1.1 Problem Statement

The primary objective of this project is to develop an accurate prediction model for customer product reorder behavior using the Instacart dataset. This presents a binary classification challenge where the target variable 'reordered' indicates whether a customer will reorder a specific product in their future purchases.

1.2 Business Context

Accurate reorder prediction offers several business advantages:

- Enhanced customer experience through personalized recommendations
- Reduced operational costs through better demand forecasting

2 Dataset and Preprocessing

2.1 Dataset Characteristics

The Instacart dataset contains rich customer-product interaction data including:

- Order history and timing information
- Product details and categories
- User purchase patterns and preferences
- Temporal features such as days since last order

2.2 Preprocessing Steps

1. **Data Cleaning**
2. **Feature Engineering**
3. **Data Transformation**
    - Label encoding for categorical variables
    - Standardization of numerical features
    - Feature scaling for optimal model performance

3 Initial Model Selection and Rationale

3.1 Model Selection Criteria

The initial model selection was based on:

- Suitability for binary classification
- Ability to handle large datasets
- Computational efficiency
- Interpretability requirements
- Performance on similar e-commerce problems

3.2 Selected Models and Rationale

1. **Logistic Regression Variants**
    - **L1 Regularization**: Selected for feature selection capabilities and handling sparse data
    - **L2 Regularization**: Chosen for preventing overfitting
    - **ElasticNet**: Included for combining L1 and L2 benefits
2. **Tree-Based Models**
    - **Decision Stump**: Basic tree model for baseline performance
    - **Random Forest (Shallow)**: Selected for ensemble benefits while maintaining computational efficiency
3. **Linear Models**
    - **Linear Discriminant Analysis**: Chosen for handling class separation effectively
    - **Ridge Classifier**: Selected for stability with correlated features
    - **Linear SVC**: Included for margin-based classification
4. **Other Classifiers**
    - **Naive Bayes**: Selected for baseline probabilistic classification
    - **QDA**: Chosen for handling non-linear class boundaries
    - **SGD Classifier**: Included for handling large-scale learning

4 Initial Model Evaluation

4.1 Cross-Validation Results

Initial evaluation using k-fold cross-validation on a 30% stratified sample yielded:

Model	**Average Accuracy**	**Time (seconds)**
Logistic Regression (L1)	0.8592	509.92
Logistic Regression (L2)	0.6568	53.91
Logistic Regression (ElasticNet)	0.6585	5210.46
Naive Bayes	0.5886	4.91
Decision Stump	0.8355	13.45
Linear Discriminant Analysis	0.8523	13.25
SGD Classifier (Log Loss)	0.6292	1458.59
SGD Classifier (Hinge Loss)	0.5930	1409.09
Random Forest (Shallow)	0.8723	197.55
Ridge Classifier	0.8523	5.75
Linear SVC	0.5938	5322.89
QDA	0.8242	13.30

Table 1: Cross-Validation Results on 30% Sample

4.2 Model Selection Criteria

Models were selected for further evaluation based on:

- Accuracy threshold >80%
- Computational efficiency
- Convergence stability
- Performance-time trade-off

5 Detailed Model Evaluation

5.1 Selected Models Performance

Comprehensive evaluation of selected models on the full dataset:

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Logistic Regression (L1)	0.8594	0.8537	0.9190	0.8851	0.9365
LDA	0.8522	0.8253	0.9506	0.8835	0.9328
Random Forest (Shallow)	0.8726	0.8252	0.9946	0.9020	0.9239
QDA	0.8242	0.8824	0.8097	0.8445	0.9135

Table 2: Final Model Performance Metrics

5.2 Performance Analysis

1. **Logistic Regression (L1)**
   - Best balance of precision (0.8537) and recall (0.9190)
   - Highest ROC-AUC score (0.9365)
   - Excellent computational efficiency
   - Strong feature selection through L1 regularization
2. **Linear Discriminant Analysis**
   - High recall (0.9506) but lower precision
   - Good ROC-AUC score (0.9328)
   - Efficient computation time
3. **Random Forest**
   - Highest recall (0.9946) but lower precision
   - Signs of overfitting despite shallow depth
   - Higher computational overhead
4. **QDA**
   - Best precision (0.8824) but lower recall
   - Lower overall performance metrics
   - Limited scalability for large datasets

6 Model Selection and Justification

6.1 Final Model Selection

Logistic Regression with L1 regularization was chosen as the optimal model based on:

- **Performance Metrics**
  - Balanced F1-score of 0.8851
  - Highest ROC-AUC score of 0.9365
  - Strong precision-recall trade-off
- **Technical Considerations**
  - Efficient computation time
  - Stable convergence
  - Built-in feature selection through L1 regularization

- **Practical Advantages**
  - High interpretability
  - Easy deployment
  - Scalable for large datasets

6.2 Model Limitations

- Limited capacity to capture non-linear relationships
- Sensitivity to feature scaling
- Potential for underfitting complex patterns

7 Conclusions

7.1 Key Findings

- Linear models demonstrated superior performance for this specific problem
- L1 regularization proved effective for feature selection
- Simple models outperformed complex ones in terms of generalization
- Computation time varied significantly across models

7.2 Practical Implications

- Model suitable for production deployment
- Balance between accuracy and computational efficiency
- Interpretable results for business stakeholders
- Scalable solution for large-scale implementation

8 Future Work

8.1 Model Improvements

- Explore advanced ensemble methods (XGBoost, LightGBM)
- Implement neural network architectures
- Investigate feature importance analysis
- Perform hyperparameter tuning