My Approach to Gender Prediction

Tewoflos Girmay

March 7, 2025

1 Choosing the Right Machine Learning Algorithm

1.1 Algorithm Selection

To predict the gender of a customer based on session transaction data, I considered three potential machine learning algorithms:

1.1.1 Random Forest Classifier

Why I Chose This: Random Forest is an ensemble method that combines multiple decision trees to improve generalization and reduce overfitting. Given the structured nature of the dataset (timestamps, categorical product interactions), decision trees can effectively capture patterns in user behavior.

Pros:

- Handles missing values well.
- Works efficiently with categorical and numerical features.
- Reduces overfitting compared to single decision trees.

Cons:

- Training can be slow for large datasets.
- Model interpretability is lower compared to linear models.

1.1.2 Logistic Regression

Why I Considered It: Logistic Regression provides a simple and interpretable model that can establish relationships between features and the target variable.

Pros:

- Computationally efficient.
- Provides probabilistic outputs for decision-making.

- Easily interpretable feature importance.

Cons:

- Assumes linear relationships between features and the target variable, which might not hold in this dataset.
- May not perform well with complex patterns in categorical data.

1.1.3 Gradient Boosting Machines (GBM)

Why I Considered It: GBM, such as XGBoost or LightGBM, can improve prediction accuracy by iteratively correcting errors made by previous models.

Pros:

- Highly flexible and can capture complex feature interactions.
- Often achieves state-of-the-art performance in structured data problems.

Cons:

- Requires careful tuning to prevent overfitting.
- Computationally expensive compared to Random Forest.

1.2 My Final Choice

I decided to use the **Random Forest Classifier** because it effectively handles categorical and numerical features, is resilient against overfitting, and is well-suited for structured datasets like this one.

2 Implementing the Gender Prediction Model

2.1 My Approach

I structured my implementation into three main steps:

1. **Data Preprocessing & Feature Engineering**
2. **Model Training & Evaluation**
3. **Visualization & Performance Analysis**

2.2 Data Preprocessing & Feature Engineering

To extract meaningful patterns from the raw data, I applied several feature engineering techniques:

- **Time-based Features**: Extracted session duration, start hour, day of the week, and part of the day.

- **Product Analysis**:
  - Extracted product categories from `product_ids`.
  - Computed category diversity and total products viewed.
  - Calculated category popularity ratios.
- **Handling Missing Values**: Used imputation for categorical and numerical features.
- **Encoding**:
  - One-hot encoding for categorical features.
  - TF-IDF for product category text data.
- **Scaling**: Standardized numerical features using `StandardScaler`.

2.3 Model Training & Evaluation

- **Data Splitting**: I split the dataset into training (75%) and test (25%) sets while maintaining class balance.
- **Handling Class Imbalance**: I applied SMOTE (Synthetic Minority Over-sampling Technique) to ensure balanced gender representation in the training data.
- **Hyperparameter Tuning**: I used GridSearchCV with Stratified K-Fold Cross-Validation to fine-tune:
  - `n_estimators` (number of trees)
  - `max_depth` (tree depth)
  - `min_samples_split` (minimum samples required for node split)
  - `class_weight` (balancing classes dynamically)
- **Performance Metrics**:
  - Accuracy Score
  - Custom Gender Score (as defined in the test document)
  - Confusion Matrix
  - Classification Report
  - ROC Curve & AUC

3 Results

- **Best Parameters**: The grid search helped me identify optimal hyperparameters for the Random Forest Classifier.
- **Accuracy Score**: Evaluated using the test set.
- **Custom Gender Score**: Calculated using the provided formula.
- **Confusion Matrix & Visualizations**: Generated plots to analyze model performance.

4 Visualization & Analysis

To better understand my model's performance, I generated various visualizations:

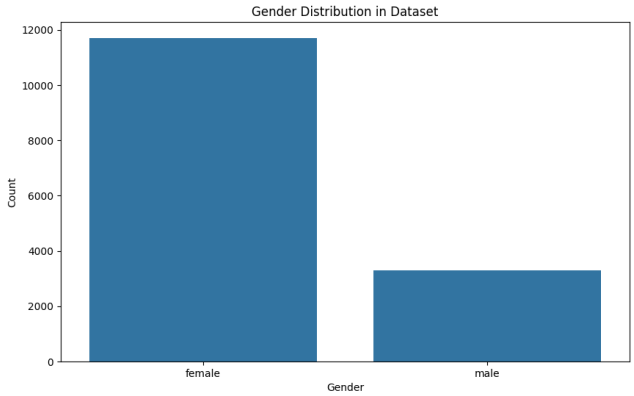

Figure 1: Gender Distribution in the Dataset

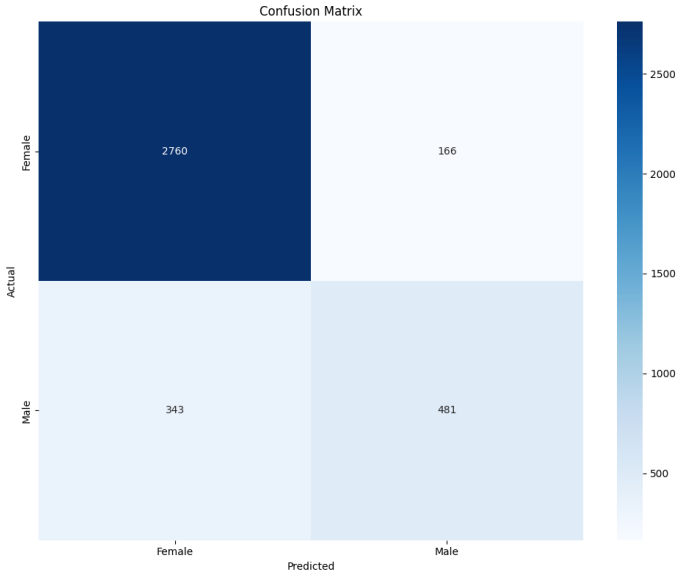

Figure 2: Confusion Matrix

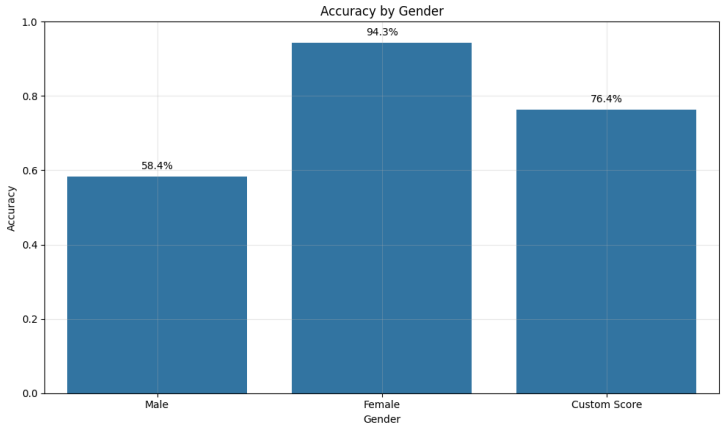

Figure 3: Gender Accuracy

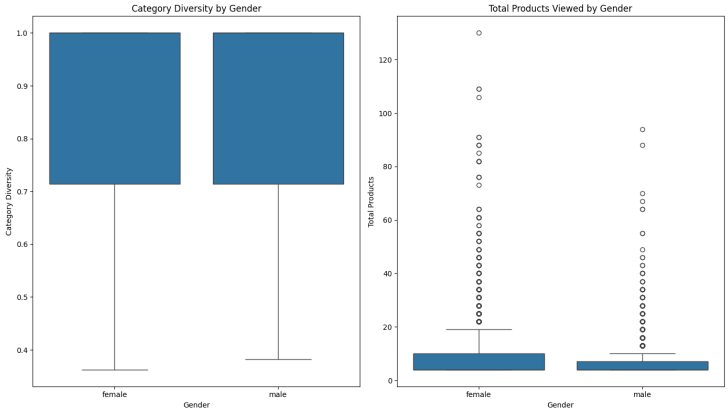

Figure 4: Shopping Behavior

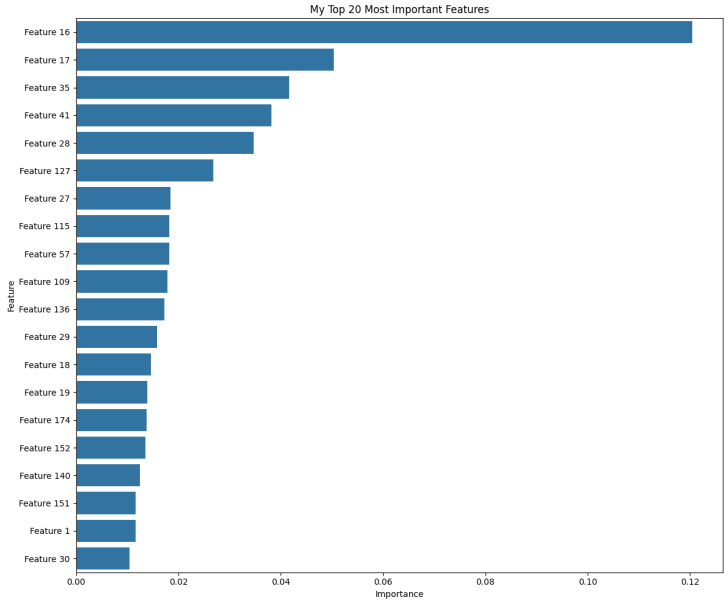

Figure 5: Feature Importance

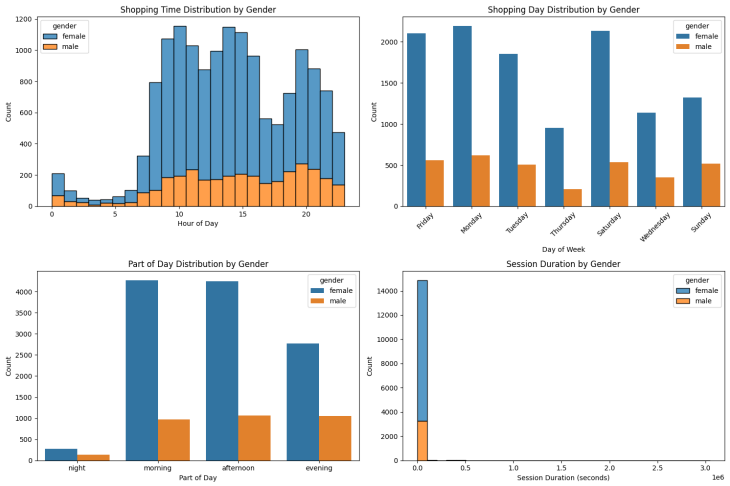

Figure 6: Time Patterns

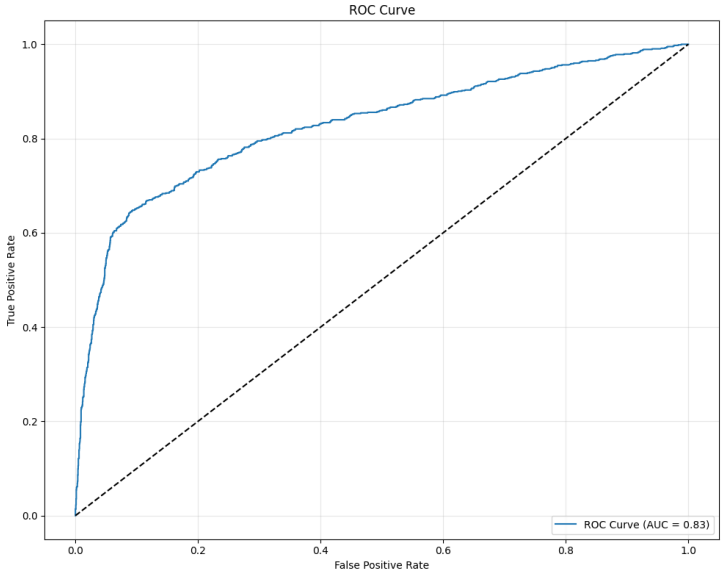

Figure 7: Roc Curve

5 Conclusion

I successfully implemented a Random Forest Classifier to predict gender based on customer sessions. My approach included extensive feature engineering, hyperparameter tuning, and visualization techniques to improve model perfor-

mance. The use of SMOTE ensured balanced class representation, while the visualizations provided insights into customer behavior. I am confident that this model effectively captures patterns in user shopping behavior to make accurate gender predictions.