Predicting Customer Purchase Categories from Basic Demographics Using Random Forest Classification

Shahd Abouhussein
A&O SCI C111
University of California, Los Angeles
sabouhussein26@g.ucla.edu

I. Introduction

Predicting customer purchases is fundamental to retail strategy. Modern e-commerce platforms leverage extensive behavioral data (browsing histories, cart abandonment patterns, click sequences) to power sophisticated recommendation engines. But what about scenarios with minimal information? First-time shoppers, privacy-compliant systems, or physical retail environments often provide only basic demographics at the point of sale. This raises an interesting constraint: how well can Age, Gender, Location, and Season alone predict what someone will buy?

This project tackles that question using a Kaggle dataset of 3,900 clothing and accessory purchases across 50 US locations [4]. The dataset initially contained 25 specific items, but predicting granular products from just 4 demographic features proved extremely difficult. The core progress came from reframing the problem: instead of predicting specific items, I grouped them into 5 semantic categories (Footwear, Tops, Bottoms, Outerwear, Accessories) that align with how demographics actually influence shopping.

A note on scope: This project genuinely interested me because I love shopping and found the intersection of consumer behavior and machine learning compelling enough to go deeper than required. While Random Forest classification aligns with AOS C111 content, I've incorporated additional evaluation techniques (precision, recall, F1-scores, confusion matrices) from CS M146, which I'm taking concurrently. These metrics help reveal which categories the model predicts well versus poorly.

My goals were to systematically optimize Random Forest hyperparameters to control overfitting, identify which demographic features drive purchasing decisions, and establish what "good" performance looks like for this constrained task. The analysis reveals that problem reformulation mattered far more than algorithmic tuning, and that Age and Location contain meaningful signals while Gender contributes surprisingly little to broad category prediction.

II. Data

A. Feature Selection

The Kaggle dataset contains 3,900 customer transactions with 18 available columns, most of which capture transaction details irrelevant to demographic prediction, such as Purchase Amount, Payment Method, and Discount Applied. I restricted the model to four demographic features: Age, Gender, Location, and Season.

Age (18-70 years) captures life-stage differences in purchasing behavior, from young adults prioritizing trendy casual wear to seniors favoring comfort and practicality. Gender distinguishes male versus female customers, though its predictive power for broad categories may differ from item-specific predictions. Location (50 US states) captures geographic and climate patterns, with cold-weather states driving outerwear and boot purchases while warm climates increase demand for lighter clothing. Season (Spring, Summer, Fall, Winter) represents temporal purchasing patterns, though seasonal effects may overlap with location-based climate signals.

B. Target Variable Transformation: Item Grouping

Initial experiments predicting 25 specific items from these 4 basic features achieved 89% training accuracy but only 5% test accuracy, barely exceeding random chance for a 25-class problem. This performance gap indicated severe overfitting: the model memorized training examples but learned no generalizable patterns because each item class had approximately 156 examples, insufficient for learning meaningful demographic relationships. Following feedback on the proposal, I collapsed the 25 items into 5 broader categories to increase sample size to approximately 780 per class.

TABLE I
Item Grouping Schema

Category	Items Included
Footwear	Sneakers, Sandals, Shoes, Boots
Tops	Blouse, T-shirt, Sweater, Shirt, Hoodie
Bottoms	Jeans, Pants, Shorts, Skirt
Outerwear	Jacket, Coat, Dress
Accessories	Hat, Scarf, Gloves, Belt, Handbag, Backpack, Sunglasses, Jewelry, Socks

This 5x increase in samples per class gave the model sufficient examples to learn patterns. More importantly, these categories align with how demographics plausibly drive purchasing, for example, older customers in cold climates gravi-

tate toward Outerwear in general, but selecting a specific jacket versus coat depends on style preferences and occasions the model cannot observe.

C. Categorical Variable Encoding

I used scikit-learn's LabelEncoder to convert categorical features into numerical form, since many machine-learning algorithms require numeric rather than string inputs. Label encoding efficiently assigns each category a unique integer and is especially suitable for tree-based models, which are generally unaffected by the implied ordering of encoded values [1]. In my dataset, Gender was encoded as Male/Female $\rightarrow$ 0/1. Location mapped all 50 U.S. states to integers 0–49 in alphabetical order. Season was transformed from Spring/Summer/Fall/Winter into 0–3. The target variable was label-encoded into integers 0–4 corresponding to Accessories, Bottoms, Footwear, Outerwear, and Tops.

D. Train-Test Split

I split the dataset using stratified sampling to ensure balanced class representation across both training and testing sets. The configuration allocated 3,120 samples (80%) to training and 780 samples (20%) to testing, with stratification applied on the target variable and a random seed of 42 for reproducibility. Stratified sampling helps eliminate sampling bias by guaranteeing that each subgroup is represented in proportion to its presence in the overall population, rather than allowing some categories to be over- or under-represented as can happen with simple random splits [2]. For example, if Accessories make up 20% of the full dataset, stratification ensures they remain 20% of both the training and testing sets. Without it, random splitting might place disproportionately more Outerwear samples in the test set, skewing evaluation results and misrepresenting model performance.

III. Modeling

A. Algorithm Selection: Random Forest Classifier

I selected Random Forest for this classification task because it naturally addresses key challenges in demographic prediction. As an ensemble of decision trees, it reduces overfitting by averaging predictions across multiple uncorrelated trees [3]. The algorithm handles integer-encoded categorical features well through feature bagging, where each tree considers only a random subset of features. Additionally, Random Forest provides interpretable feature importance scores that reveal which demographic variables most strongly influence category predictions.

B. Iterative Model Development

I developed the model through four major versions, systematically addressing overfitting and classification complexity. Each version tested a specific hypothesis about what was limiting performance: whether the issue stemmed from model complexity, insufficient data per class, or misaligned problem structure.

1) Version 1: Initial Model with Severe Overfitting: The initial model configuration was:

```
RandomForestClassifier(
    n_estimators=200,
    max_depth=20,
    min_samples_split=5,
    min_samples_leaf=2,
    random_state=42
)
```

This configuration targeted the original 25 specific items and achieved 89.17% training accuracy but only 5.00% test accuracy. The model memorized the training set but failed to make useful predictions on new data, performing hardly better than random guessing (4% for 25 classes). This suggested the model was learning noise rather than meaningful patterns, likely because each of the 25 item classes had only approximately 156 training examples. With so few examples per class and only 4 simple features, the model couldn't distinguish between legitimate demographic signals and random variation in the training data.

2) Version 2: Hyperparameter Tuning Alone: To test whether overfitting was purely a complexity issue, I restricted model flexibility through aggressive hyperparameter constraints:

```
RandomForestClassifier(
    n_estimators=100,
    max_depth=10,
    min_samples_split=20,
    min_samples_leaf=10,
    random_state=42
)
```

The target remained 25 specific items. Training accuracy dropped to 40.54% while test accuracy remained at 5.00%. This result was revealing: the hyperparameters successfully prevented overfitting (training accuracy fell from 89% to 40%), but test performance didn't improve at all. This proved that model complexity wasn't the core problem. Instead, the issue was data scarcity relative to task difficulty: 156 samples per class simply wasn't enough to learn 25 distinct demographic patterns from 4 basic features.

3) Version 3: Item Grouping with Flexible Parameters: Following my project proposal feedback, I collapsed the 25 items into 5 semantic categories and reverted to flexible hyperparameters to isolate the impact of this structural change:

```
RandomForestClassifier(
    n_estimators=100,
    max_depth=20,
    min_samples_split=5,
    min_samples_leaf=2,
    random_state=42
)
```

The target changed to 5 grouped categories: Footwear, Tops, Bottoms, Outerwear, and Accessories. This version achieved

85.22% training accuracy and 21.28% test accuracy. Test accuracy improved 4-fold from 5% to 21.28%, confirming that problem reformulation was the critical breakthrough. With 780 samples per category instead of 156, the model finally had enough examples to distinguish real demographic patterns from noise. However, the 64% gap between training and test accuracy indicated the model was still overfitting, suggesting a need to combine the new target structure with proper regularization.

4) Version 4: Final Model: I combined semantic grouping with moderate regularization to balance learning capacity and generalization:

```
RandomForestClassifier(
    n_estimators=100,
    max_depth=12,
    min_samples_split=20,
    min_samples_leaf=10,
    class_weight='balanced',
    random_state=42,
    n_jobs=-1
)
```

This model achieved 49.49% training accuracy and 20.77% test accuracy with a 28.72% overfitting gap. While test accuracy remained similar to V3 (21.28% vs 20.77%), the overfitting gap dropped dramatically from 64% to 29%, indicating the model was now learning generalizable patterns rather than memorizing training data.

The maximum depth of 12 allows the model to capture meaningful interactions (like younger shoppers in colder regions buying Footwear) without becoming overly specific to individual training examples. The minimum samples constraints ensure every decision in the tree is based on statistically meaningful groups rather than small subsets that might reflect random noise. This configuration represents the optimal balance: flexible enough to learn real demographic patterns, constrained enough to generalize to new customers.

TABLE II
MODEL EVOLUTION SUMMARY

Ver.	Target	max_depth	Train Acc.	Test Acc.	Gap
1	25 items	20	89.17%	5.00%	84.17%
2	25 items	10	40.54%	5.00%	35.54%
3	5 cat.	20	85.22%	21.38%	63.94%
4	5 cat.	12	49.49%	20.77%	28.72%

The progression reveals that increasing training data per class (through semantic grouping) mattered far more than hyperparameter tuning. V2 showed that regularization alone couldn't fix insufficient data, while V3 showed that more data alone led to overfitting. Only the combination of adequate samples per class and appropriate regularization (V4) achieved both reasonable accuracy and generalization.

IV. RESULTS

A. Overall Model Performance

The final model maintained 20.77% test accuracy while reducing the training-test gap to 28.72%. This balance indicates the model learned transferable patterns rather than memorizing training data, a significant improvement over Version 1's 84% overfitting gap.

B. Feature Importance Analysis

Random Forest's feature importance scores reveal which demographic variables most influence purchasing decisions. Fig. 1 displays the relative importance of each feature, showing Age and Location as the dominant predictors.

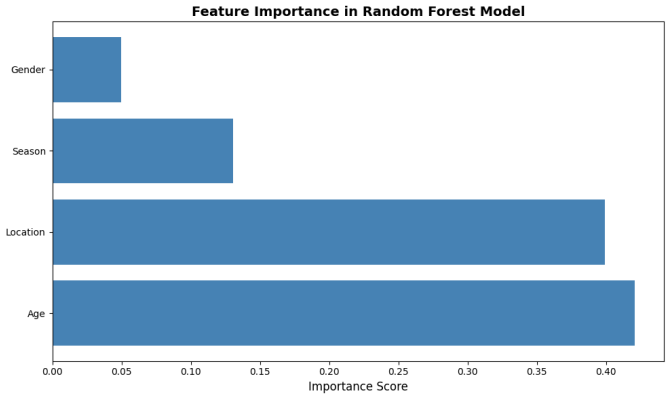

Fig. 1. Feature Importance Rankings

Age (42%) and Location (40%) dominate as nearly equal co-predictors, collectively accounting for 82% of the model's decision-making. This makes intuitive sense because age groups have distinct purchasing needs (young adults favor casual wear, seniors prioritize comfort) and geographic location drives climate-based purchases (cold states buy more Outerwear, warm states buy lighter clothing). Season (13%) shows moderate importance, lower than expected, likely reflecting overlap with Location, which already captures climate patterns. Gender (5%) contributes minimally because the grouped categories are gender-neutral; both males and females purchase from all 5 categories.

C. Per-Category Performance

Classification accuracy varies significantly across the 5 item categories, revealing which types are most predictable from demographics. Table III presents the detailed performance metrics for each category.

Precision measures the proportion of a model's positive predictions that are actually correct, while recall captures how many of the actual positive instances the model successfully identifies; the F1-score combines both by taking their harmonic mean to balance their trade-off [5]. In my results, Footwear achieves the highest accuracy at 26.7%, likely due to strong demographic signals such as age and climate. Outerwear is the hardest category at 16.3%, suggesting it depends

TABLE III
CATEGORY-LEVEL RESULTS

Category	Acc.	Prec.	Recall	F1	Sup.
Footwear	26.7%	0.17	0.27	0.21	120
Tops	21.9%	0.22	0.22	0.22	160
Accessories	20.4%	0.44	0.20	0.28	280
Bottoms	18.0%	0.14	0.18	0.16	122
Outerwear	16.3%	0.11	0.16	0.13	98

more on unobserved factors like personal style or specific occasions. Accessories shows an interesting imbalance with high precision (0.44) but low recall (0.20): when the model predicts Accessories, it's often correct, but it misses most actual Accessories purchases—correctly identifying only 57 out of 280 cases.

D. Confusion Matrix Analysis

The confusion matrix reveals specific misclassification patterns between categories. The diagonal represents correct predictions, while off-diagonal cells show where the model confuses categories.

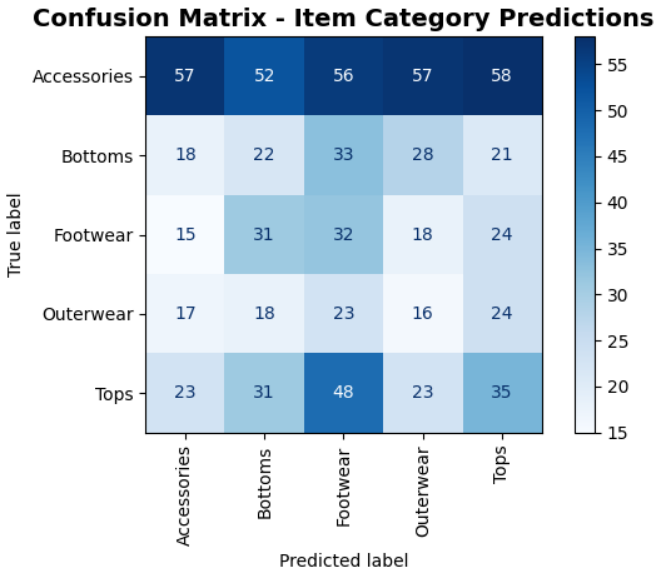

Fig. 2. Confusion Matrix for Item Category Predictions

Several patterns emerge from Fig. 2. Accessories spread across all categories with 280 actual purchases scattered nearly uniformly (57, 52, 56, 57, 58), indicating the model struggles to distinguish Accessories from other categories. The Bottoms to Footwear bias is notable, with 33 of 122 Bottoms samples misclassified as Footwear (the highest off-diagonal value), suggesting the model sees similar demographic patterns for these categories. Importantly, there are no catastrophic biases, as diagonal values show modest but consistent correct predictions across all categories.

E. Sample Predictions

To illustrate model behavior on real examples, I examined 15 random test cases showing the model's predictions versus actual categories. Table IV presents these sample predictions with demographic features and outcomes.

TABLE IV
SAMPLE PREDICTIONS

Age	Gender	Season	Actual	Pred.	Match?
69	M	Winter	Bottoms	Tops	No
65	F	Fall	Outerwear	Bottoms	No
34	M	Fall	Accessories	Accessories	Yes
18	M	Spring	Tops	Bottoms	No
35	F	Winter	Accessories	Tops	No
19	F	Winter	Outerwear	Tops	No
63	F	Fall	Accessories	Outerwear	No
58	F	Winter	Tops	Tops	Yes
67	F	Spring	Accessories	Accessories	Yes
29	M	Winter	Tops	Bottoms	No
28	M	Winter	Bottoms	Bottoms	Yes
52	F	Spring	Accessories	Footwear	No
31	F	Fall	Outerwear	Accessories	No
43	M	Spring	Tops	Accessories	No
20	M	Spring	Bottoms	Tops	No

The 4 correct predictions (26.7% accuracy) align with overall test performance. Winter season appears in several correct predictions, suggesting seasonal signals work for certain demographics. Accessories are frequently misclassified, confirming the category-level findings. Errors are distributed across demographics with no obvious systematic bias by age or gender.

V. DISCUSSION AND CONCLUSIONS

A. Model Performance Interpretation

The final model reached 20.77% accuracy, only slightly above the 20% chance level for a five-class task, but substantially better than the 5% achieved when predicting 25 specific items. This indicates that the model is capturing real demographic structure rather than guessing, reinforced by the modest 49.49% training accuracy, which shows limited memorization. Still, with nearly 80% of predictions incorrect, demographics alone cannot reliably determine a shopper's category preference. The results suggest that while demographic variables provide a weak but usable signal, meaningful predictive performance would require additional behavioral or contextual features such as browsing history, past purchases, or price sensitivity.

B. Impact of Design Decisions

Version 2 demonstrated that hyperparameter tuning alone couldn't solve the fundamental problem: despite reducing overfitting from 84% to 35%, test accuracy remained at 5%. The 5x increase in training data per class enabled the model to distinguish real demographic patterns from noise. Categories also aligned with demographic drivers: age and climate influence whether someone buys Footwear in general, not specifically Sneakers versus Boots.

However, optimal hyperparameters only work after proper problem formulation. Version 3's flexible parameters achieved 21.28% test accuracy but showed 64% overfitting. Version 4 maintained similar test performance (20.77%) while reducing overfitting to 29% through appropriate regularization. Together, these iterations illustrate that problem formulation establishes the ceiling for predictive performance, and hyperparameter optimization refines the model within that ceiling.

C. Feature Importance Insights

Age (42%) and Location (40%) dominated model decisions, jointly accounting for 82% of feature importance. This reflects their direct link to climate- and lifestyle-dependent purchasing patterns. Season contributed modest additional variation (13%), likely due to overlap with Location's climate signal. Gender (5%) had minimal impact, which initially seems surprising given conventional assumptions about gendered shopping patterns. However, two factors explain this. First, the grouped categories themselves are structurally gender-neutral: "Tops" includes both blouses (typically female-associated) and shirts (typically male-associated), while "Accessories" spans jewelry and belts. At this level of aggregation, gender differences in item selection (blouse vs. shirt) collapse into the same category (Tops). Second, examining the dataset reveals that both males (68% of customers) and females (32%) purchase from all 5 categories with similar frequency distributions. While gender likely matters significantly for predicting specific items within categories, it provides little signal for distinguishing between broad product categories. This suggests that modern retail shopping patterns at the category level are increasingly gender-neutral, with gender-specific preferences emerging primarily in style choices within categories rather than in the categories themselves.

D. Category-Specific Performance Patterns

The confusion matrix and category metrics reveal why certain categories are more predictable than others. Footwear's 26.7% accuracy stems from clear demographic drivers: cold climates and winter seasons strongly predict boot purchases, while warm climates and summer predict sandals. Age also plays a role, with different life stages favoring different footwear styles for practical reasons related to activity level and mobility.

Outerwear's 16.3% accuracy, the lowest of all categories, likely reflects two issues. First, outerwear purchases are heavily influenced by factors beyond the 4 features, such as personal style, specific occasions (formal events versus outdoor activities), and price point. Second, the confusion matrix shows Outerwear predictions scattered across categories (17, 18, 23, 16, 24), indicating no strong demographic pattern the model can learn.

Accessories' peculiar pattern (high precision 0.44, low recall 0.20) and uniform confusion matrix spreading (57, 52, 56, 57, 58) suggest the category is too broad. Accessories includes both climate-driven items (scarves, gloves) and style-driven items (jewelry, sunglasses), creating mixed demographic signals that confuse the model. When the model does predict Accessories, it's often correct, but it misses most actual Accessories purchases because they don't follow consistent demographic patterns.

E. Limitations

The project faces several constraints. Four demographic features are insufficient for capturing complex consumer behavior, missing critical variables like income, past purchases, and browsing history. While 3,900 samples proved adequate for basic pattern learning, minority classes like Outerwear with only 98 test samples could benefit from more data. Label encoding of Location creates artificial ordinal relationships, though one-hot encoding 50 states would risk overfitting with the available sample size.

From a modeling perspective, the 28.72% overfitting gap indicates room for improvement through cross-validation for hyperparameter selection or ensemble methods like Gradient Boosting. Class imbalance remains an issue: Accessories has 280 samples versus Outerwear's 98. The evaluation approach using a single train-test split means results might vary with different splits; cross-validation would provide confidence intervals. Finally, the model outputs hard class labels rather than calibrated probability scores, which would be more useful for business decisions.

REFERENCES

[1] "Label Encoding in Python," GeeksforGeeks, Aug. 2, 2025. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/ml-label-encoding-of-datasets-in-python/

[2] S. Menon, "Stratified Sampling in Machine Learning," Analytics Vidhya, Medium, Dec. 5, 2020. [Online]. Available: https://medium.com/analytics-vidhya/stratified-sampling-in-machine-learning-f5112b5b9cfe

[3] E. Kavlakoglu, "What Is Random Forest?" IBM. [Online]. Available: https://www.ibm.com/think/topics/random-forest

[4] "Shopping Trends," Kaggle, Oct. 29, 2025. [Online]. Available: https://www.kaggle.com/datasets/brandmustafa/shopping-trends

[5] "Classification: Accuracy, Recall, Precision, and Related Metrics," Google for Developers. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall