Classification of Online Shopper Intention Using Python

Objective

The goal of this project was to build a machine learning model that predicts whether someone browsing an online store ends up making a purchase. Using real user behavior data, I wanted to see which model could best spot the sessions that actually lead to revenue.

Dataset Summary

The dataset came from the UCI Machine Learning Repository. It tracks user interactions on an e-commerce site on things like how many pages someone visits, how long they stay, what time of year it is, whether they're a returning visitor, and so on. The final column (Revenue) tells us whether the session ended with a purchase.

This is a binary classification task: predict 1 if the user buys something, 0 if they don't.

1. Preprocessing the Data

- Loaded the dataset using pandas.
- No missing values to worry about.
- Convert categorical features like Month and VisitorType into numbers using label encoding.
- Turn Weekend and Revenue from True/False into 1/0.
- Scaled all numeric features using StandardScaler so they're in the same range.
- Split the data into 80% training and 20% testing.

2. Model Training

I trained two models to compare:
1. Logistic Regression: A simple, fast baseline.
2. Random Forest Classifier: A more powerful model that combines multiple decision trees.

For Random Forest, I used GridSearchCV to try different hyperparameters. I kept it lightweight with just two options for max_depth and set n_estimators to 100.

3. How the Models Performed

Metric	Logistic Regression	Random Forest
Accuracy	86.9%	89.6%
Precision	72.7%	77.2%

Recall	34.3%	53.5%
F1 Score	46.6%	63.2%
ROC AUC	0.878	0.925

Random Forest beat Logistic Regression across the board especially on recall and F1 score. Higher recall means Random Forest caught way more actual buyers (53.5% vs 34.3%).
In real-world terms, that's huge. You'd rather catch more potential customers even if it means a few more false positives.
The ROC AUC also jumped from 0.878 to 0.925, showing that Random Forest is better at separating buyers from non-buyers.

Confusion Matrices

Logistic Regression:

```
[[2002   53]
 [ 270  141]]
```

It missed 270 buyers and only caught 141.

Random Forest:

```
[[1990   65]
 [ 191  220]]
```

It missed fewer buyers (191) and correctly identified 220.

4. Overall Comparison

Logistic Regression was fast and easy to implement, but it missed a lot of actual buyers. It had decent precision but weak recall.
Random Forest, on the other hand, found more of the sessions that ended in purchases, had stronger overall balance, and gave a more useful signal for real business use.
That's especially important if you're using the model for marketing or conversion optimization.

Final Thoughts

If the goal is to help a business understand which sessions are likely to turn into revenue, Random Forest is the better pick.
It doesn't just fit the data, it adds value by catching more of the behavior patterns that actually lead to sales.