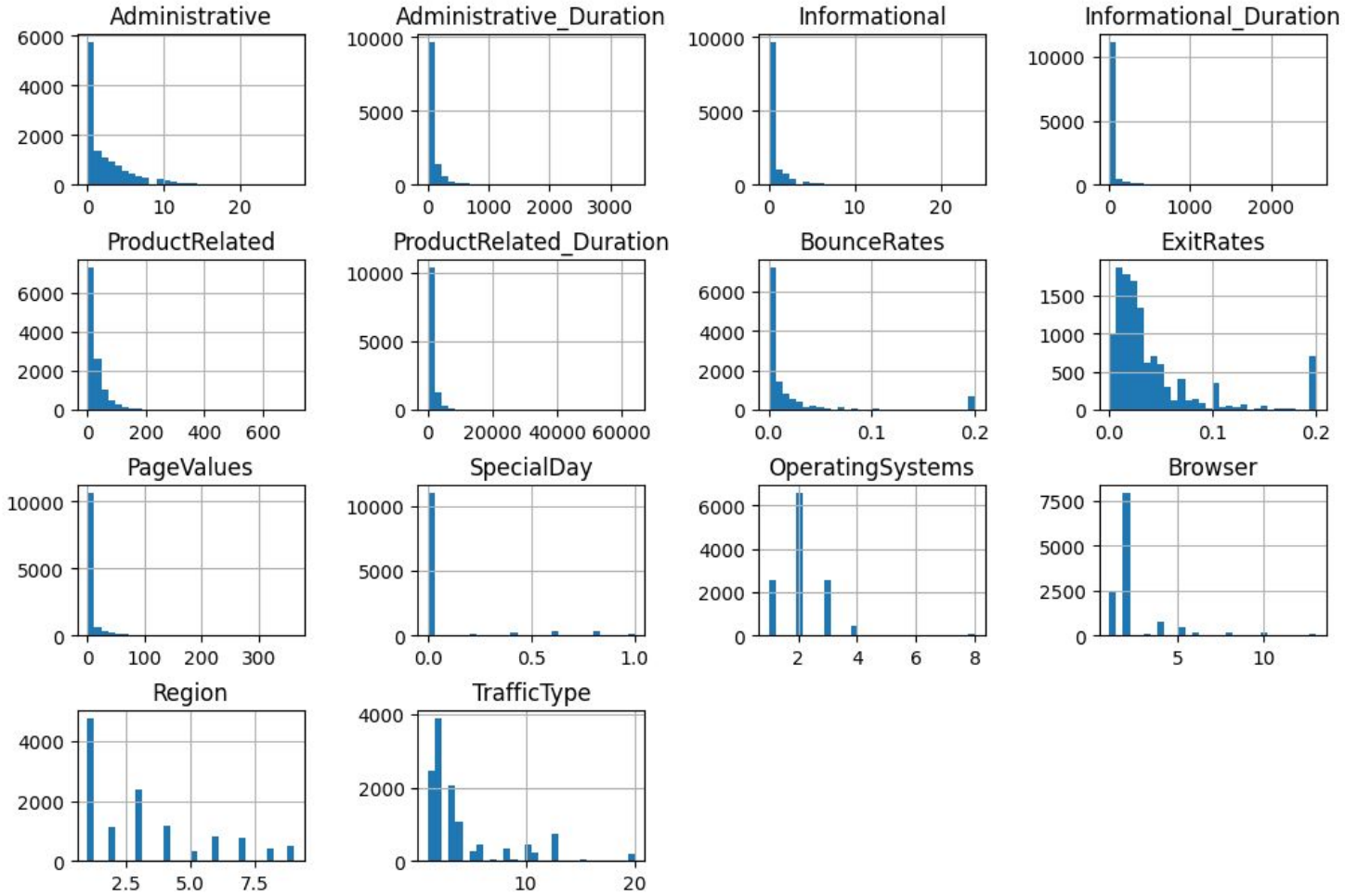
Shopper Behavior Prediction

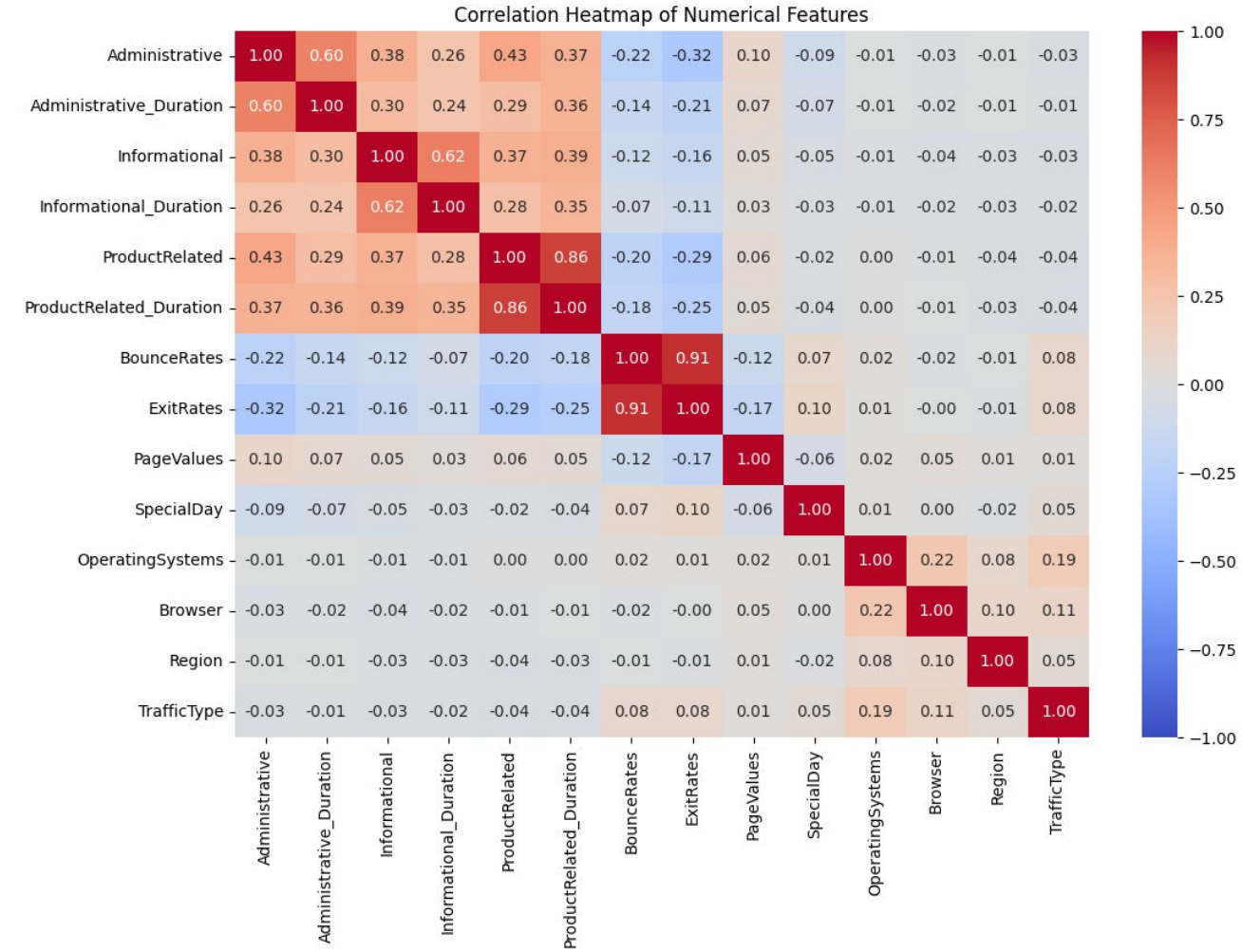
Analyzing User Behavior to Predict Revenue Generation

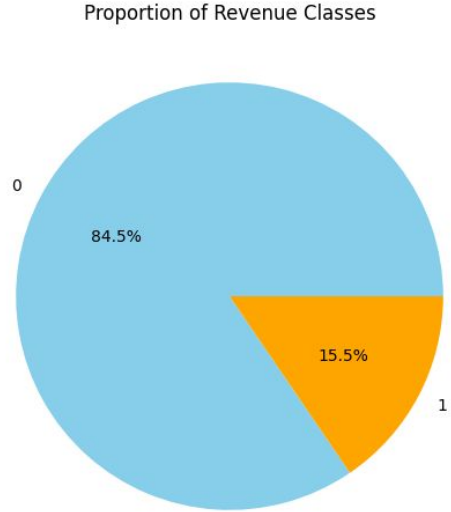
By: Matthew Neba

Introduction

- **Dataset:** UCI Online Shopper Intention Dataset
- **Objective:** Analyze user behavior to predict revenue generation
- **Data Composition:**
  - **Samples:** Online shopping sessions
  - **Features:** 17 variables including session duration, visited pages, bounce rates, and visitor type
  - **Response Variable:** Revenue (True/False)


Distribution of Numerical Features
Administrative
Administrative_Duration
Informational
Informational_Duration
ProductRelated
ProductRelated_Duration
BounceRates
ExitRates
PageValues
SpecialDay
OperatingSystems
Browser
Region
TrafficType


Correlation Heatmap of Numerical Features
Administrative 1.00 0.60 0.38 0.26 0.43 0.37 -0.22 -0.32 0.10 -0.09 -0.01 -0.03 -0.01 -0.03
Administrative_Duration 0.60 1.00 0.30 0.24 0.29 0.36 -0.14 -0.21 0.07 -0.07 -0.01 -0.02 -0.01 -0.01
Informational 0.38 0.30 1.00 0.62 0.37 0.39 -0.12 -0.16 0.05 -0.05 -0.01 -0.04 -0.03 -0.03
Informational_Duration 0.26 0.24 0.62 1.00 0.28 0.35 -0.07 -0.11 0.03 -0.03 -0.01 -0.02 -0.03 -0.02
ProductRelated 0.43 0.29 0.37 0.28 1.00 0.86 -0.20 -0.29 0.06 -0.02 0.00 -0.01 -0.04 -0.04
ProductRelated_Duration 0.37 0.36 0.39 0.35 0.86 1.00 -0.18 -0.25 0.05 -0.04 0.00 -0.01 -0.03 -0.04
BounceRates -0.22 -0.14 -0.12 -0.07 -0.20 -0.18 1.00 0.91 -0.12 0.07 0.02 -0.02 -0.01 0.08
ExitRates -0.32 -0.21 -0.16 -0.11 -0.29 -0.25 0.91 1.00 -0.17 0.10 0.01 -0.00 -0.01 0.08
PageValues 0.10 0.07 0.05 0.03 0.06 0.05 -0.12 -0.17 1.00 -0.06 0.02 0.05 0.01 0.01
SpecialDay -0.09 -0.07 -0.05 -0.03 -0.02 -0.04 0.07 0.10 -0.06 1.00 0.01 0.00 -0.02 0.05
OperatingSystems -0.01 -0.01 -0.01 -0.01 0.00 0.00 0.02 0.01 0.02 0.01 1.00 0.22 0.08 0.19
Browser -0.03 -0.02 -0.04 -0.02 -0.01 -0.01 -0.02 -0.00 0.05 0.00 0.22 1.00 0.10 0.11
Region -0.01 -0.01 -0.03 -0.03 -0.04 -0.03 -0.01 -0.01 0.01 -0.02 0.08 0.10 1.00 0.05
TrafficType -0.03 -0.01 -0.03 -0.02 -0.04 -0.04 0.08 0.08 0.01 0.05 0.19 0.11 0.05 1.00
Administrative
Administrative_Duration
Informational
Informational_Duration
ProductRelated
ProductRelated_Duration
BounceRates
ExitRates
PageValues
SpecialDay
OperatingSystems
Browser
Region
TrafficType
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Proportion of Revenue Classes
0
84.5%
15.5%
1

Feature Preprocessing:

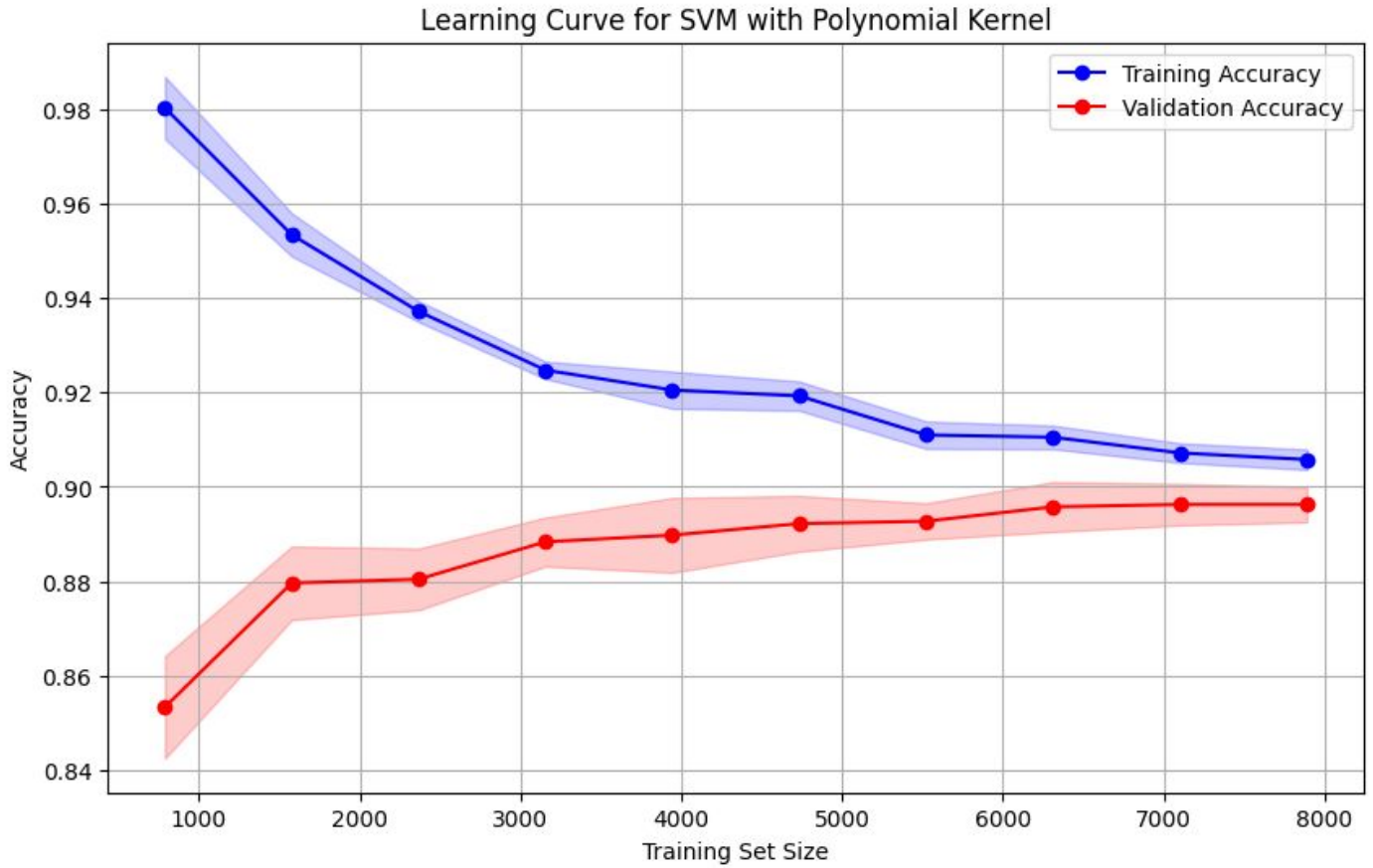
- **Numerical Features:** Imputed missing values using `SimpleImputer(strategy="mean")`, standardized using `StandardScaler()`
- **Categorical Features:** Imputed missing values using `SimpleImputer(strategy='constant', fill_value='missing')`, one-hot encoded using `OneHotEncoder(handle_unknown='ignore')`

Methodology & Tools

- **Algorithm Used:** Support Vector Machine (SVM) with a polynomial kernel
- **Hyperparameter Tuning:** 5-fold internal cross-validation to determine optimal cost and degree
- **Evaluation Metric:** Accuracy
-
- `pandas` for data manipulation
- `scikit-learn` for model training, preprocessing, and evaluation
- `datasets` to load UCI dataset

Model Training

- **Model Construction:**
  - SVM Classifier implemented using `SVC(kernel="poly")`
  -
- **Hyperparameter Optimization:**
  - Grid search performed using `GridSearchCV()` with parameter grid:
    - `classifier__degree`: [2, 3, 4, 7]
    - `classifier__C`: [0.1, 10, 100, 1000]
  - Best parameters identified using 5 fold cross-validation, scoring based on accuracy

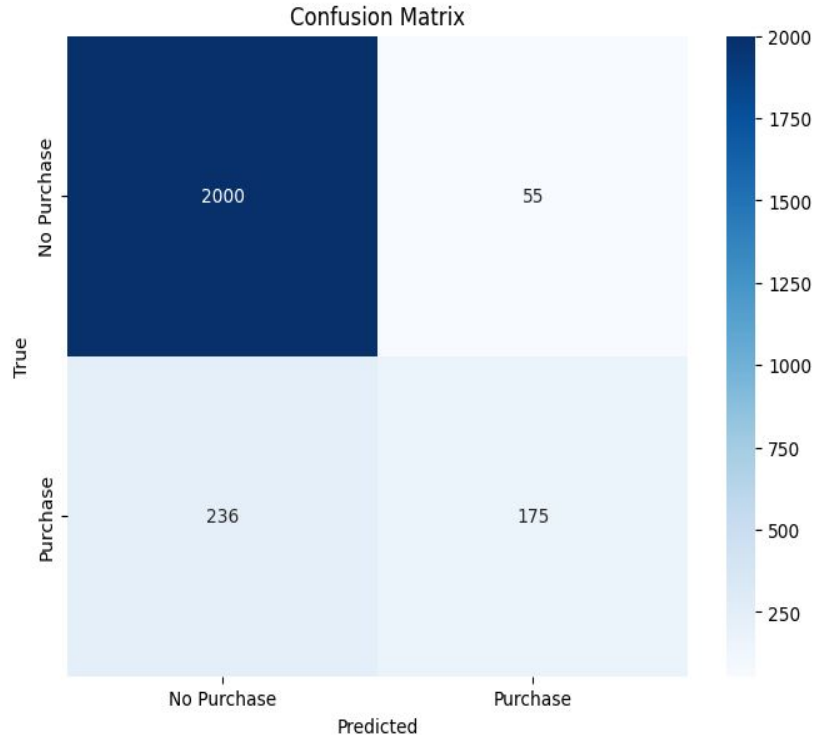

Learning Curve for SVM with Polynomial Kernel
Training Accuracy
Validation Accuracy
Accuracy
Training Set Size
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
1000
2000
3000
4000
5000
6000
7000
8000

Evaluation

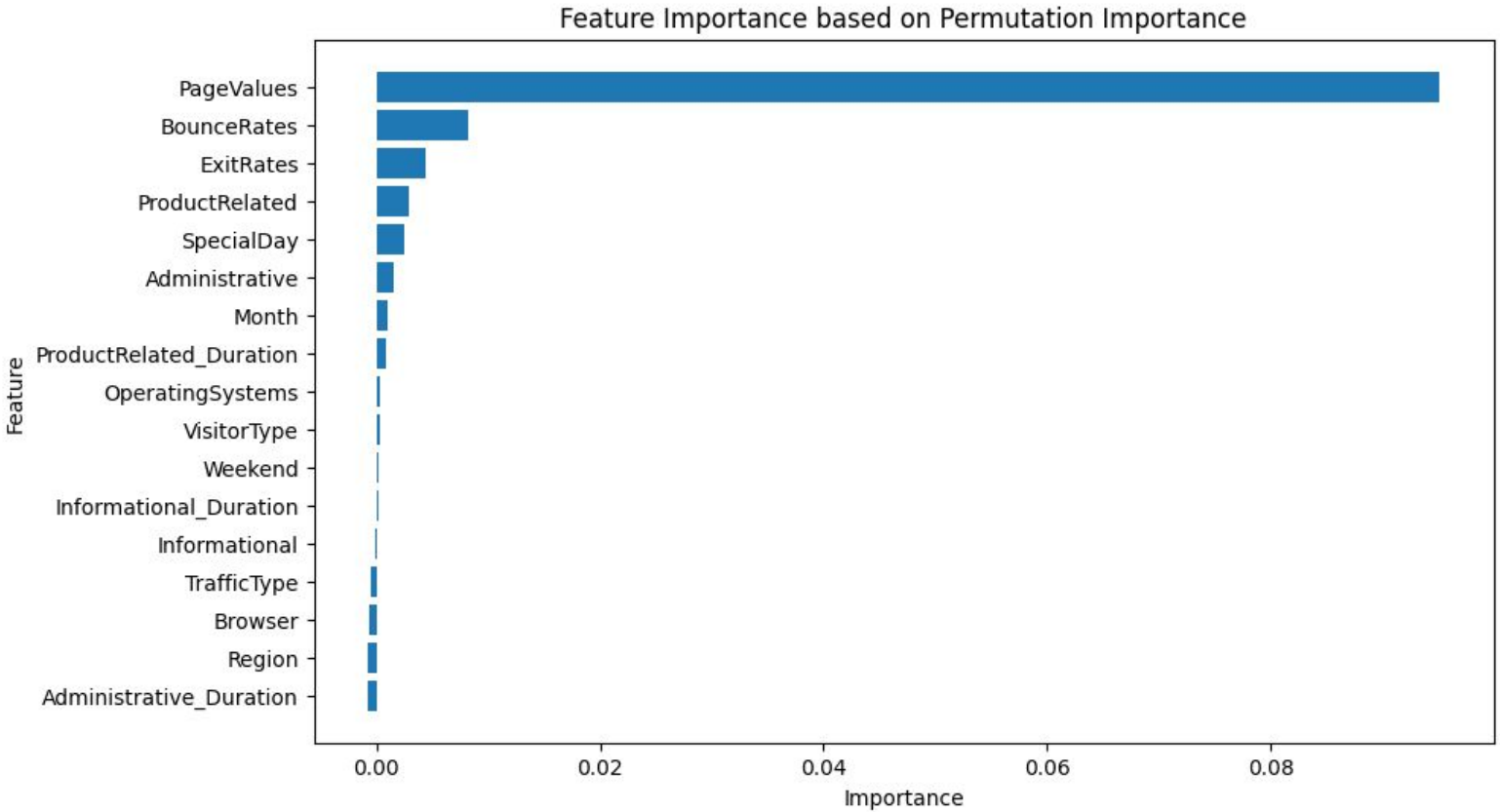
- **Accuracy Score:** Used as the primary metric to assess model performance (`accuracy_score()` from `sklearn.metrics`)
- **Confusion Matrix:** Provides a breakdown of true positives, true negatives, false positives, and false negatives (`confusion_matrix()` from `sklearn.metrics`)
- **Feature Importance (Permutation Importance Analysis):**
  - Computed using `permutation_importance()` from `sklearn.inspection`
  - Evaluates how much model accuracy decreases when a feature is randomly shuffled

Key Findings & Insights

- **Best SVM Parameters Found:** Cost: 100, Degree of Polynomial Kernel: 2
- **Training Accuracy:** 0.8963
- **Test Accuracy:** 0.8820



Feature Importance (Permutation Importance Analysis)



Conclusion

Business Implications

- **Customer engagement with high-value pages** is the strongest predictor of purchasing behavior
- **Session quality metrics** (bounce/exit rates) are more valuable than visitor demographics
- Product browsing behavior is moderately predictive of purchase intent
- Temporal factors (Month, Weekend) and regional information have minimal predictive power
- Technical attributes (Browser, OS) show little correlation with purchasing decisions