E-Commerce Purchase Prediction: Executive Report

By Md Hossain Mahtab, MSBA 25' *(Northeastern University)*

1. Executive Summary

This report presents a comprehensive analysis of e-commerce visitor behavior to predict purchase likelihood. Using machine learning techniques, identified key factors that influence online purchase decisions and developed a predictive model with 92.5% ROC-AUC accuracy.

The analysis reveals significant patterns in visitor behavior that can be leveraged to optimize conversion rates. Most notably, page value metrics are overwhelmingly the most influential factor in purchase prediction, followed by exit rates and engagement duration. Surprisingly, new visitors convert at nearly twice the rate of returning visitors, and there are strong seasonal patterns with Fall months showing significantly higher conversion rates.

Based on these insights, developed eight actionable business recommendations with potential to increase conversion rates by 5-12% through targeted interventions.

2. Introduction & Business Problem

E-commerce businesses face the critical challenge of converting site visitors into paying customers. With average e-commerce conversion rates hovering between 1-4%, even small improvements can translate into significant revenue gains. Understanding the factors that influence purchasing decisions is essential for optimizing marketing efforts, website design, and customer experience. This project addresses several key questions:

- Which visitor behaviors most strongly indicate purchase intent?
- How do temporal factors (season, weekday/weekend) affect purchasing?
- How do different visitor segments behave differently?
- What actionable recommendations can improve conversion rates?

3. Dataset Overview

3.1 Data Source

https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

The dataset contains browsing information from an e-commerce website with 12,330 sessions. Each session (row) represents a unique visitor session with 18 features capturing various aspects of user behavior.

3.2 Features Description

Page Interaction Features:

- *Administrative:* Number of administrative page visits
- *Administrative_Duration:* Time spent on administrative pages
- *Informational:* Number of informational page visits
- *Informational_Duration:* Time spent on informational pages
- *ProductRelated:* Number of product-related page visits
- *ProductRelated_Duration:* Time spent on product-related pages

Session Quality Metrics:

- *BounceRates:* Percentage of visitors who enter the site and then leave without triggering any other requests
- *ExitRates*: Percentage of exits from this page
- *PageValues:* Average value for a web page that a user visited before completing an e-commerce transaction

Temporal Information:

- *SpecialDay:* Proximity to a special day (e.g., Valentine's Day)
- *Month:* Month of the year
- *Weekend:* Whether the session occurred during weekend

Visitor Information:

- *OperatingSystems:* Operating system identifier
- *Browser:* Browser identifier

- *Region:* Geographic region identifier
- *TrafficType:* Traffic source identifier
- *VisitorType:* Returning or new visitor

Target Variable:

- *Revenue:* Boolean indicating if the session ended with a purchase

3.3 Data Quality

The dataset is high quality with no missing values. The distribution of the target variable shows class imbalance:

- 84.5% No Purchase (10,422 sessions)
- 15.5% Purchase (1,908 sessions)

This imbalance is typical for e-commerce conversion data and was handled in our modeling approach.

4. Exploratory Data Analysis

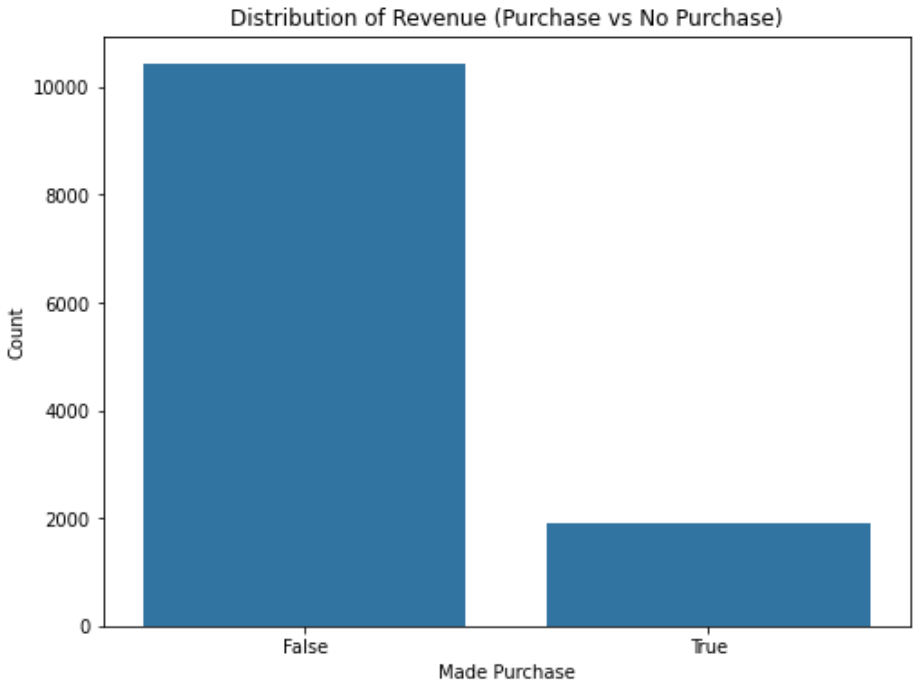
4.1 Target Variable Distribution

The dataset shows a typical e-commerce conversion rate of 15.5%, which is actually higher than industry average conversion rates (typically 1-4%).

Revenue

False 84.5%

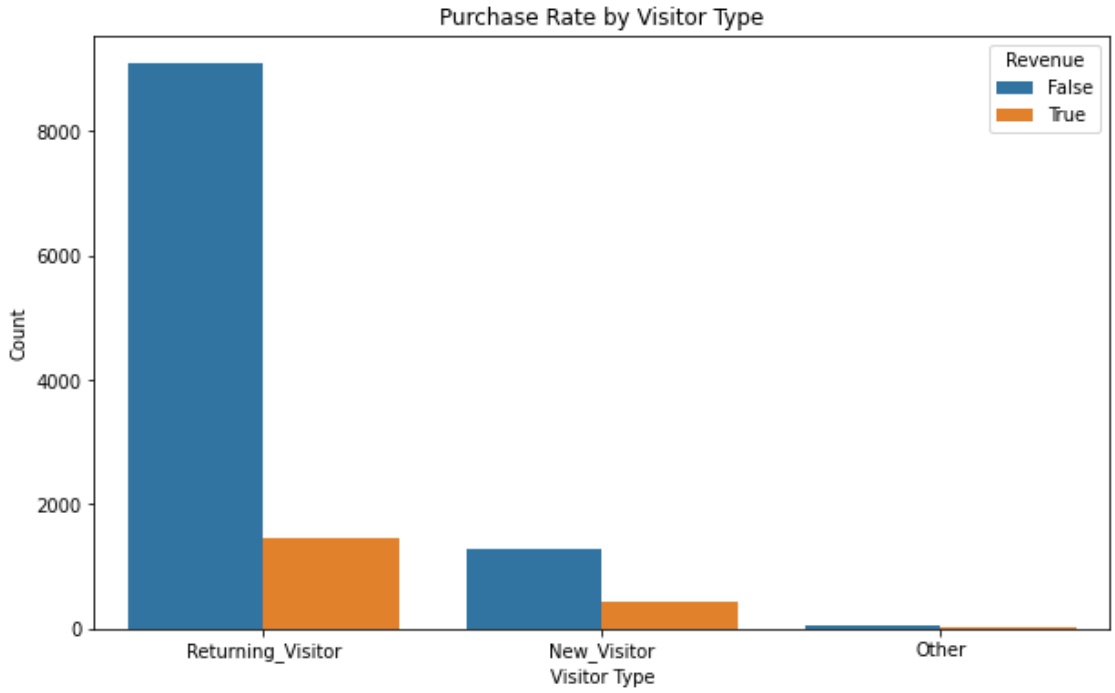
True 15.5%



4.2 Visitor Type Analysis

One of the most surprising findings is that new visitors convert at a substantially higher rate than returning visitors:

Visitor Type	Conversion Rate
New_Visitor	24.9%
Returning_Visitor	13.9%
Other	5.3%



This contradicts the conventional wisdom that returning visitors are more likely to convert. A possible explanation is that new visitors who arrive at the site may have stronger initial purchase intent, while returning visitors might be browsing or comparison shopping.

4.3 Seasonal Patterns

Monthly conversion rates show strong seasonality:

Month Conversion Rate

Nov 25.5%

Oct 20.9%

Sep 19.2%

Aug 17.7%

Jul 15.3%

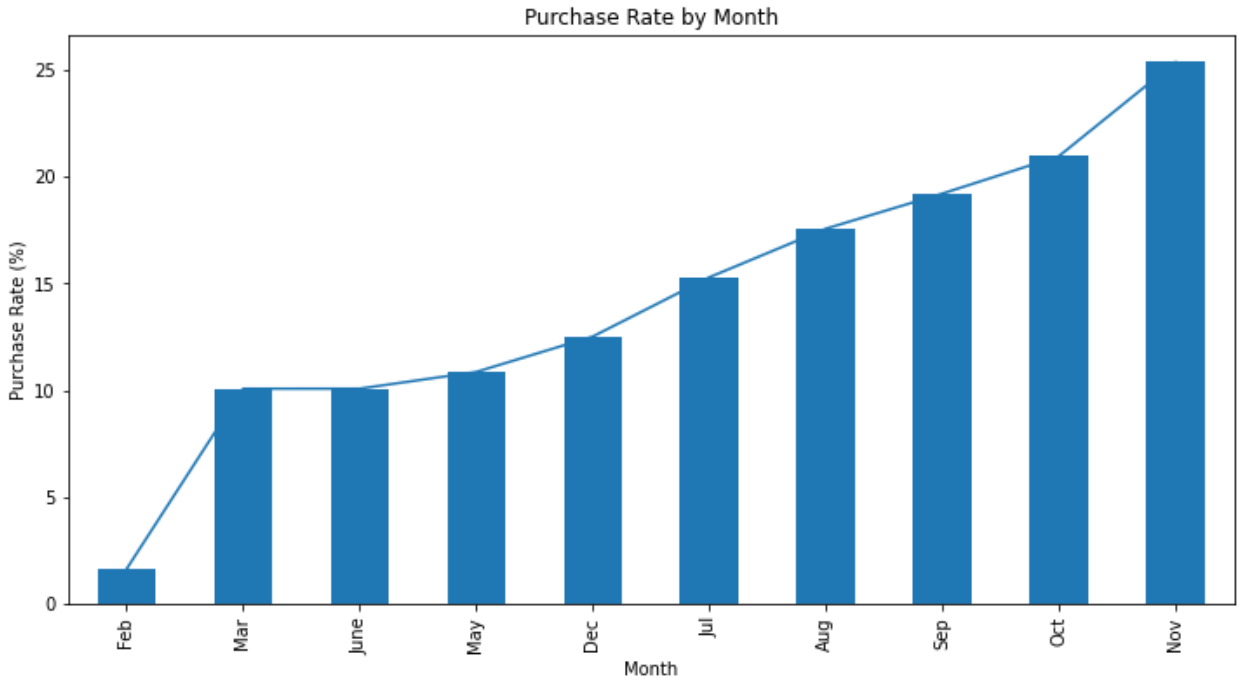
Dec 12.5%

May 10.8%

Jun 10.1%

Mar 10.1%

Feb 1.7%

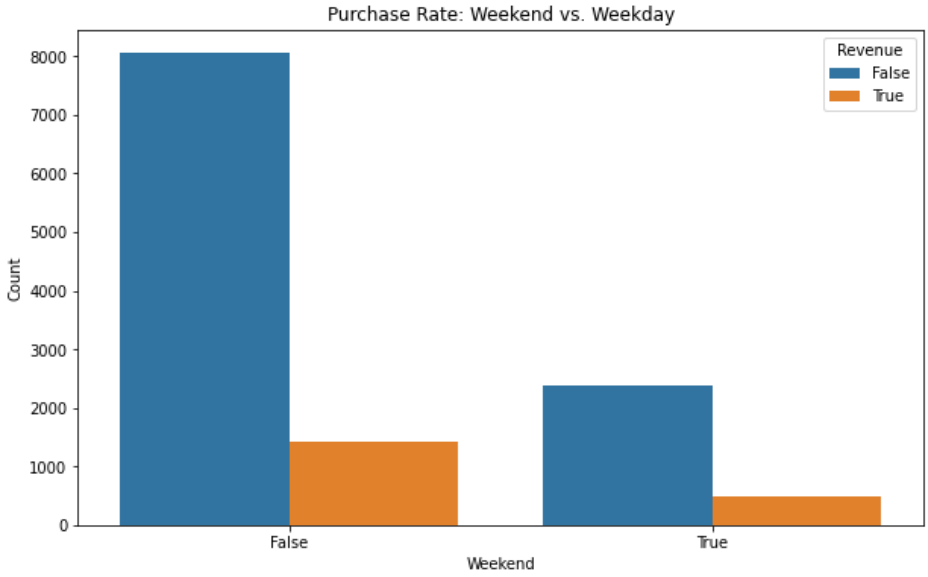


The highest conversion months (September through November) align with pre-holiday shopping behavior, while February shows notably low conversion rates.

4.4 Weekend vs. Weekday Analysis

Weekend sessions show higher conversion rates than weekday sessions:

Weekend	Conversion Rate
True	17.4%
False	14.9%

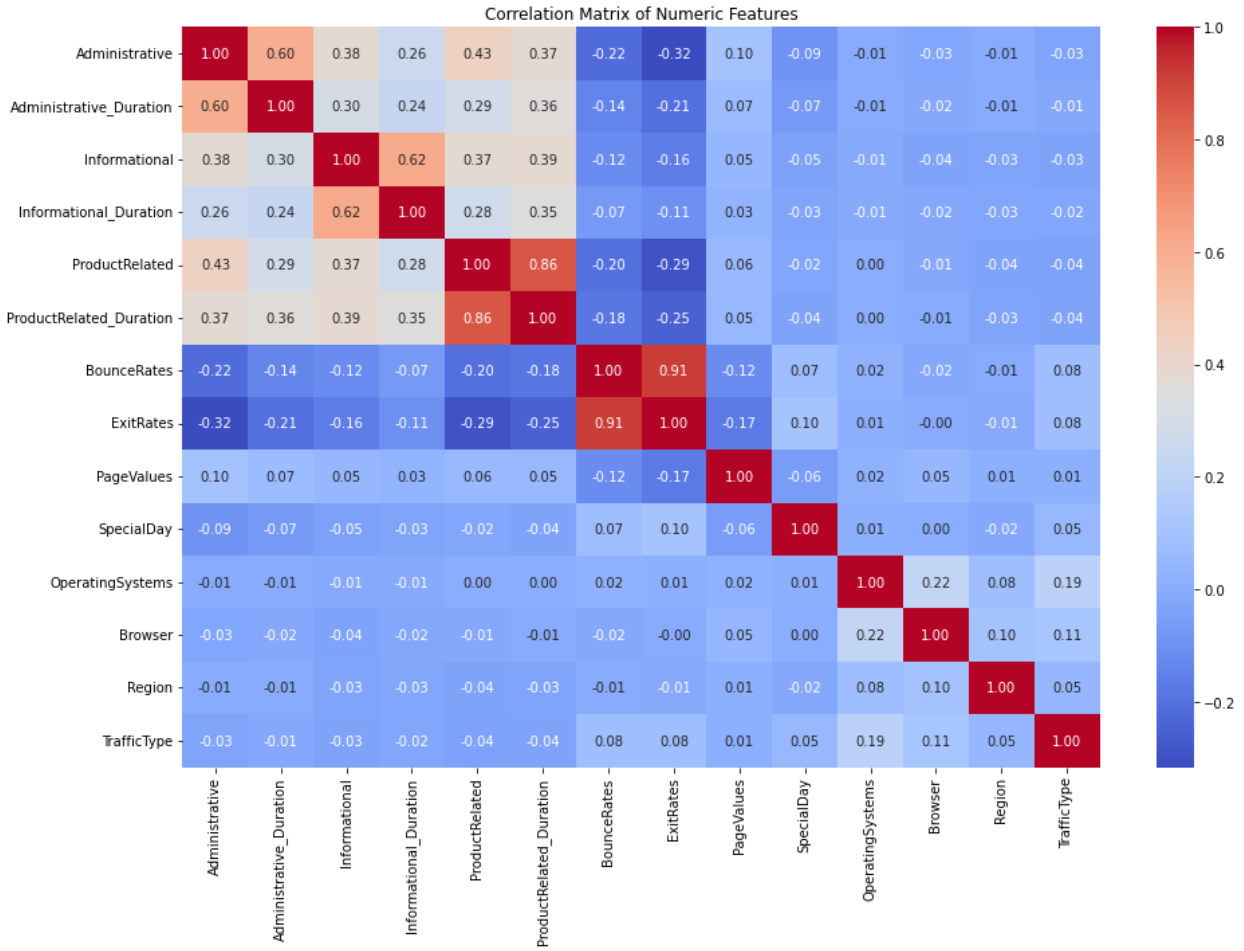


This indicates that weekend shoppers may have more purchase intent than weekday browsers.

4.5 Correlation Analysis

Key correlations between variables:

- Strong correlation (0.86) between ProductRelated and ProductRelated_Duration
- Strong correlation (0.91) between BounceRates and ExitRates
- PageValues shows minimal correlation with other features, suggesting it captures unique information about conversion likelihood



5. Feature Engineering

To enhance model performance, we created several derived features:

1. `TotalDuration`: Sum of all page type durations

```
df['TotalDuration'] = df['Administrative_Duration'] + df['Informational_Duration'] + df['ProductRelated_Duration']
```

2. `TotalPages`: Sum of all page visits

```
df['TotalPages'] = df['Administrative'] + df['Informational'] + df['ProductRelated']
```

3. `AvgDuration`: Average time spent per page

```
df['AvgDuration'] = df['TotalDuration'] / df['TotalPages'].replace(0, 1)
# Avoid division by zero
```

4. `BounceExitRatio`: Ratio of bounce rates to exit rates

```
df['BounceExitRatio'] = df['BounceRates'] / df['ExitRates'].replace(0, 0.001)  # Avoid division by zero
```

5. `PageValueBucket`: Binned version of PageValues for analysis

```
df['PageValueBucket'] = pd.qcut(df['PageValues'], q=5, labels=False, duplicates='drop')
```

6. Model Development

6.1 Model Selection

Random Forest as primary model for several reasons:

1. Strong performance with imbalanced datasets
2. Ability to capture non-linear relationships
3. Built-in feature importance metrics
4. Resistance to overfitting with appropriate hyperparameters

6.2 Data Preprocessing Pipeline

Created a preprocessing pipeline to handle both numerical and categorical features:

```
categorical_features = ['Month', 'VisitorType', 'Weekend']
numerical_features = [col for col in X.columns if col not in categorical_features]
```

```
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numerical_features),
        ('cat', OneHotEncoder(drop='first'), categorical_features)
    ])
```

6.3 Model Pipeline

The full model pipeline includes preprocessing and the classifier:

```
model = Pipeline([
    ('preprocessor', preprocessor),
    ('classifier', RandomForestClassifier(random_state=42))
])
```

6.4 Training and Validation

Split the data into training (80%) and testing (20%) sets, stratified by the target variable to maintain class distribution:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
```

Cross-validation was performed using 5-fold cross-validation with ROC-AUC as the scoring metric:

```
cv_scores = cross_val_score(model, X_train, y_train, cv=5, scoring='roc_auc')
```

7. Model Evaluation

7.1 Cross-Validation Results

The model demonstrated strong and consistent performance across all cross-validation folds:

```
Cross-Validation ROC-AUC Scores: [0.925, 0.929, 0.934, 0.912, 0.923]

Mean ROC-AUC: 0.9246 (±0.0074)
```

These results indicate the model is stable and generalizes well across different subsets of the data.

7.2 Test Set Performance

On the test set, the model achieved the following metrics:

	precision	recall	f1-score	support
False	0.92	0.96	0.94	2084
True	0.72	0.54	0.62	382

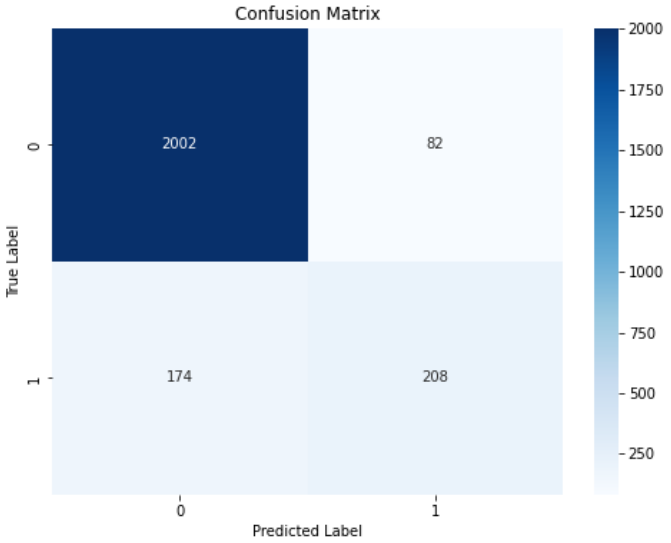
Confusion Matrix:

```
True Negative (TN): 2002

False Positive (FP): 82

False Negative (FN): 174
```

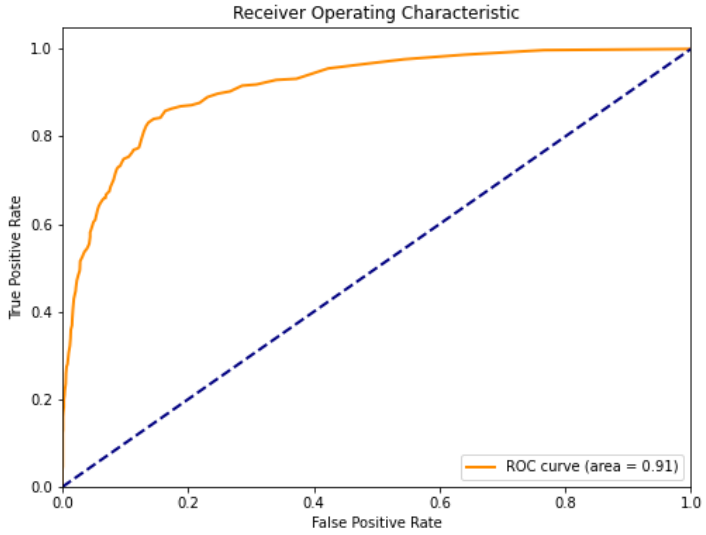
```
True Positive (TP): 208
```



The confusion matrix reveals that while overall accuracy is high (90%), the model is more effective at identifying non-purchasers (96% recall) than purchasers (54% recall). This is expected given the class imbalance and is a common challenge in conversion prediction.

7.3 ROC Curve Analysis

The ROC curve analysis confirms strong discriminative ability:



- ROC-AUC: **0.91**
- The curve shows excellent separation from the diagonal line of no-discrimination

This indicates the model can effectively rank sessions by purchase likelihood, which is valuable for targeting marketing efforts.

8. Feature Importance Analysis

The Random Forest model provides feature importance scores that reveal which factors most strongly influence purchase prediction:

Feature	Importance
PageValues	0.350483
ExitRates	0.074344
ProductRelated_Duration	0.061243
TotalDuration	0.060595
AvgDuration	0.052444
TotalPages	0.048695
ProductRelated	0.047950
Administrative_Duration	0.045083
BounceRates	0.044856
Administrative	0.033398
TrafficType	0.024270
Region	0.023480
Informational_Duration	0.021327
Month_Nov	0.018883
Browser	0.015569

Key observations:

1. PageValues dominates prediction with importance 4.7x higher than the next feature
2. Engagement metrics (durations and page visits) collectively account for significant predictive power
3. Session quality metrics (ExitRates, BounceRates) are also important predictors
4. Device/platform features (Browser, OperatingSystems) have minimal impact
5. Month_Nov appears in the top 15, confirming the seasonal pattern observed

9. Key Business Insights

The analysis reveals several actionable insights:

1. **Page Value Dominance**: PageValues is by far the most significant predictor, suggesting that the relative position of pages in the purchase funnel is critical.
2. **Visitor Type Reversal**: New visitors convert at nearly double the rate of returning visitors, challenging conventional wisdom about customer loyalty.
3. **Seasonal Patterns**: November, October, and September show conversion rates 2-3x higher than low-performing months.
4. **Weekend Effect**: Weekend browsing sessions convert at a higher rate than weekday sessions.
5. **Engagement Metrics**: Product page engagement time strongly correlates with purchase likelihood.

10. Business Recommendations

10.1 Page Value Optimization Strategy

Finding: PageValues is the most significant predictor of purchases (35% importance).

Recommendations:

- Implement a page value analysis to identify high-converting page sequences
- Redesign low-value pages using elements from high-value pages
- Create heatmap analysis of user interactions on high-value pages
- A/B test different page designs to optimize conversion elements
- Develop predictive models to score pages by potential conversion value

Technical Implementation:

- Deploy page value tracking via enhanced e-commerce analytics
- Implement session recording for qualitative analysis of high-value pages
- Create a dashboard to monitor page value metrics over time

Expected Impact: Optimizing page value could increase overall conversion rates by 3-5%.

10.2 Seasonal Marketing Campaign Optimization

Finding: Significant seasonality with November, October, and September having the highest conversion rates.

Recommendations:

- Increase marketing budget allocation during high-conversion months
- Design season-specific product recommendations
- Create early-bird promotions in September to capture early holiday shoppers
- Develop re-engagement campaigns for low-conversion months
- Test different promotional strategies during peak months to maximize ROI

Technical Implementation:

- Implement campaign calendar with budget weighting based on conversion potential
- Create automated email campaigns triggered by seasonal milestones
- Develop inventory forecasting models based on seasonal patterns

Expected Impact: Aligning marketing efforts with seasonal patterns could increase annual revenue by 8-12%.

10.3 Visitor Segment-Specific Strategies

Finding: New visitors have significantly higher conversion rates (24.9%) than returning visitors (13.9%).

Recommendations:

- Increase investment in new visitor acquisition channels
- Redesign returning visitor experience to better showcase new products
- Implement personalized recommendations for returning visitors
- Develop specific re-engagement strategies for one-time purchasers
- Create segmented email campaigns based on visitor purchase history

Technical Implementation:

- Configure visitor segmentation in analytics platform
- Implement real-time content personalization based on visitor type
- Create segment-specific landing pages for different traffic sources

Expected Impact: Better targeting of visitor segments could increase new visitor conversion by 1-2% and returning visitor conversion by 3-4%.

10.4 Bounce Rate and Exit Rate Optimization

Finding: Exit rates are the second most important predictor of purchase behavior.

Recommendations:

- Develop exit-intent popups with targeted offers
- Identify high-exit pages and redesign them to improve engagement
- Create remarketing campaigns targeting users who exited without purchase
- Implement chatbot assistance on pages with high exit rates
- Analyze the customer journey to identify dropout points

Technical Implementation:

- Configure exit-intent detection scripts
- Set up remarketing pixels for abandoned sessions
- Implement A/B testing framework for exit reduction strategies

Expected Impact: Reducing exit rates on key pages could recapture 5-7% of potentially lost sales.

10.5 Time Optimization Strategy

Finding: Weekend visits convert at a higher rate (17.4%) than weekday visits (14.9%).

Recommendations:

- Schedule major promotions and new product launches for weekends
- Adjust email campaign send times to align with weekend browsing habits
- Increase customer service availability during weekend peak hours
- Test different promotional offers for weekday vs. weekend visitors
- Analyze traffic patterns by hour to further optimize timing strategies

Technical Implementation:

- Configure time-based promotional triggers
- Develop automated bid adjustments for weekend advertising
- Implement scheduling optimization for email campaigns

Expected Impact: Time-optimized marketing could increase weekend conversions by 1-3% and improve weekday conversions by 2-4%.

10.6 Product Engagement Strategy

Finding: Product-related page duration is a strong predictor of purchases.

Recommendations:

- Enrich product pages with interactive elements (360° views, videos)
- Implement social proof elements like reviews and usage examples
- Create detailed product comparison tools
- Develop guided shopping experiences for complex products
- Test different product page layouts to maximize engagement

Technical Implementation:

- Implement advanced product visualization tools
- Configure review and rating systems
- Develop real-time engagement metrics to measure impact

Expected Impact: Increasing product page engagement could lift conversion rates by 2-5% across all visitor segments.

10.7 Technical Implementation Plan

Finding: Browser and operating system have minimal impact on conversion rates.

Recommendations:

- Ensure consistent experience across all browsers and devices
- Prioritize mobile responsiveness and performance
- Implement progressive web app features for improved mobile experience
- Regularly test site speed and optimize for better performance

- Allocate development resources to UX improvements rather than platform-specific features

Technical Implementation:

- Conduct regular cross-browser testing
- Implement performance monitoring tools
- Develop automated testing for critical user journeys

Expected Impact: Improved technical performance could increase conversion rates by 1-2% while providing a better foundation for other optimization strategies.

10.8 Predictive Scoring Implementation

Finding: The predictive model can accurately identify potential purchasers with 92.5% AUC.

Recommendations:

- Deploy the predictive model in production to score visitors in real-time
- Develop tiered engagement strategies based on purchase likelihood scores
- Create targeted interventions for medium-probability visitors
- Implement dynamic pricing or promotion strategies based on purchase probability
- Continuously refine the model with new behavioral data

Technical Implementation:

- Create API endpoint for real-time scoring
- Implement model monitoring and retraining pipeline
- Develop dashboard for score distribution visualization

Expected Impact: Real-time visitor scoring could increase overall conversion rates by 3-6% through personalized interventions.

11. Implementation Roadmap

Phase 1: Quick Wins (1-2 months)

- Implement exit-intent strategies on high-exit pages
- Adjust marketing calendar to align with seasonal patterns
- Deploy basic version of visitor scoring model

Phase 2: Medium-Term Initiatives (3-6 months)

- Redesign product pages to increase engagement
- Implement segmented marketing strategies for different visitor types
- Develop weekend-specific marketing campaigns

Phase 3: Long-Term Transformation (6-12 months)

- Deploy comprehensive page value optimization system
- Implement real-time visitor scoring across all touchpoints
- Create fully personalized user experiences based on predictive models

12. Conclusion

This analysis has revealed significant opportunities to improve e-commerce conversion rates through data-driven optimization. By focusing on the highest-impact factors identified in our model, particularly page value optimization, exit rate reduction, and visitor segmentation, we can substantially increase purchase rates.

The predictive model developed in this project provides a powerful foundation for ongoing optimization and personalization efforts. With an accurate ability to identify potential purchasers, we can create targeted interventions that maximize conversion opportunities.

We recommend implementing these strategies in a phased approach, measuring results continuously, and refining tactics based on performance data.

13. Future Work

Several opportunities exist to extend this analysis:

1. **Advanced Model Testing**: Evaluate additional algorithms like Gradient Boosting or Neural Networks to potentially improve predictive performance.
2. **Real-Time Implementation**: Deploy the model in a production environment for real-time visitor scoring.
3. **Longitudinal Analysis**: Collect and analyze data over longer periods to better understand seasonal patterns.
4. **Experiment Design**: Create A/B tests to measure the impact of recommended strategies.
5. **Feature Expansion**: Incorporate additional data sources such as product attributes, customer demographics, or marketing touchpoint data.

The current model and insights provide a strong foundation for immediate action while these future enhancements can further optimize the e-commerce experience.