CS313 DATA SCIENCE PROJECT

Faisal Al-Yahya  
CCIS  
Prince Sultan UniversityRiyadh, Saudi Arabia  
fyalyahya@gmail.com

***Abstract*—**This paper investigates the impact of various user behavior features on revenue generation in an e-commerce setting. Utilizing a comprehensive dataset, we employ descriptive and predictive analytics to uncover significant trends and correlations. Key findings indicate that user engagement metrics, such as page visit duration and bounce rates, significantly influence revenue. Predictive models, including Logistic Regression, Decision Tree, SVM, and KNN, were evaluated to forecast revenue generation, with optimization enhancing model performance. The study provides actionable insights for optimizing marketing strategies and user experience to boost conversions.

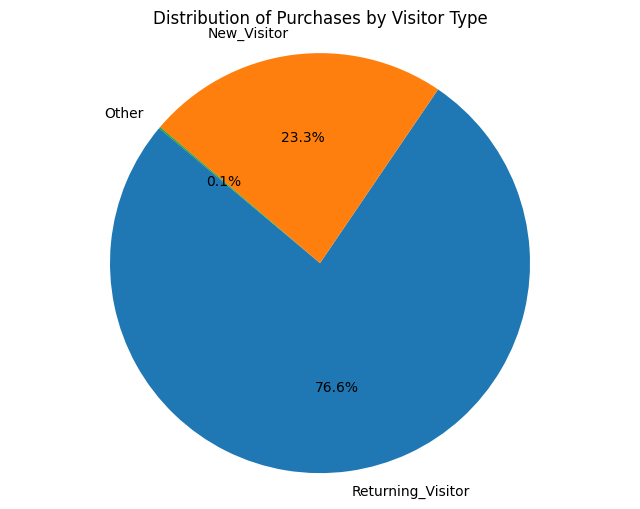
***Keywords—***E-commerce, user behavior, predictive analytics, logistic regression, decision tree, SVM, KNN, revenue optimization

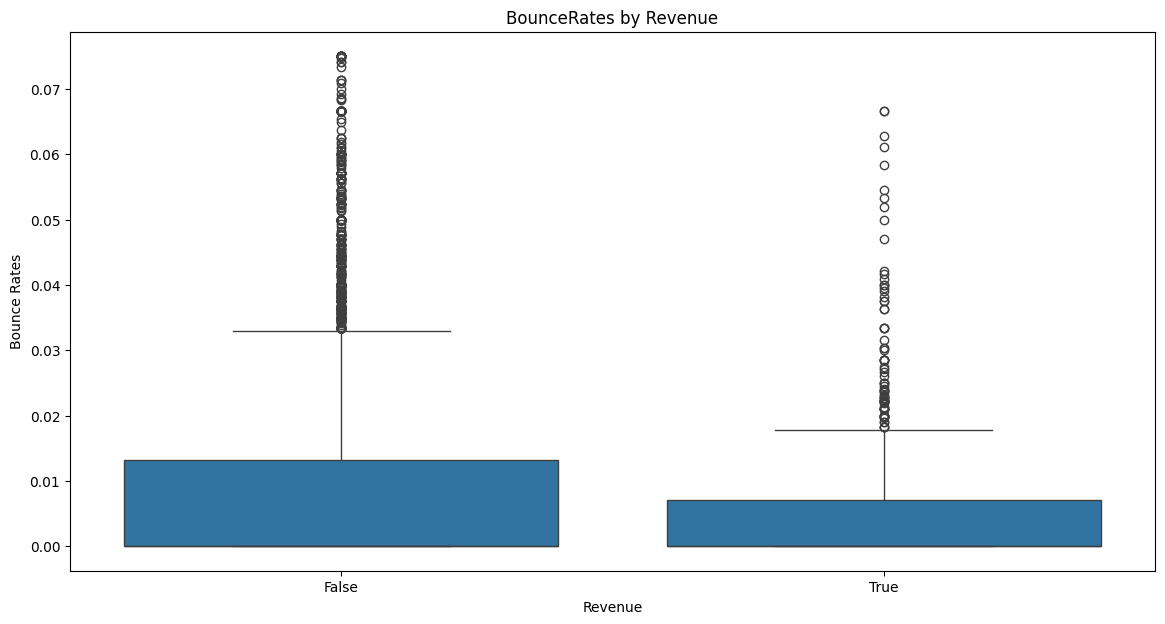
# Introduction (*Heading 1*)

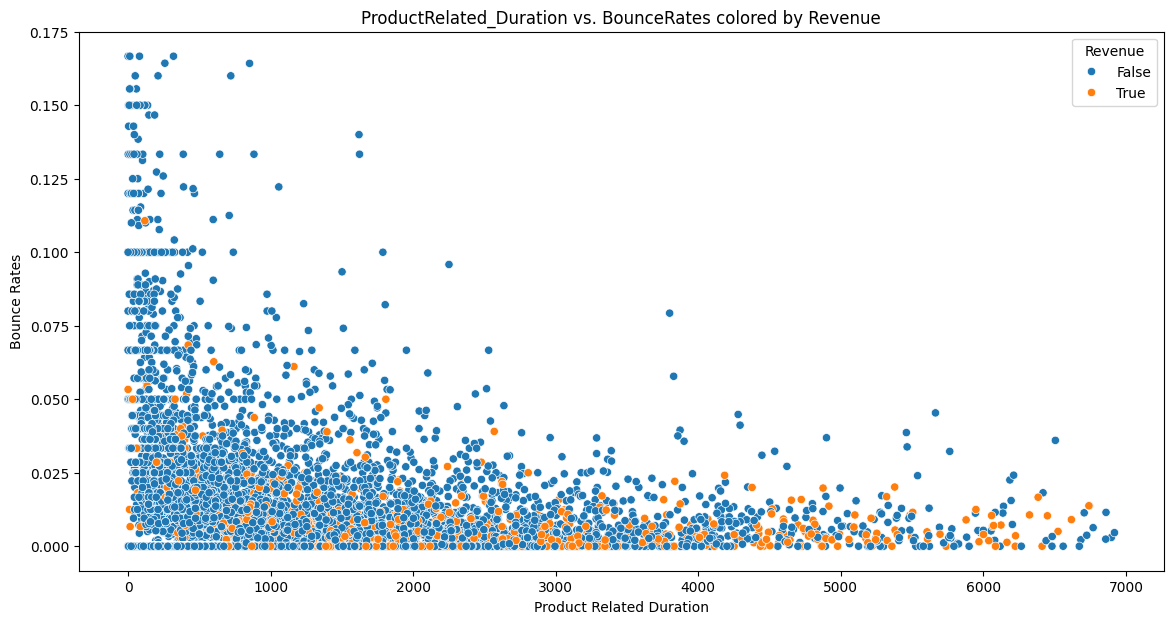
The rapid expansion of e-commerce has necessitated a deeper understanding of the factors influencing revenue generation. This study focuses on identifying and analyzing key user behavior features that impact revenue, aiming to provide data-driven insights for improving business strategies. The relevance of this research lies in its potential to guide e-commerce platforms in enhancing user engagement and optimizing conversion rates.

# Descriptive Analytics

*AUser Behavior Insights*

1 ) Distribution of Purchases by Visitor Type: Returning customers generate the bulk of revenue, emphasizing the importance of customer retention strategies.

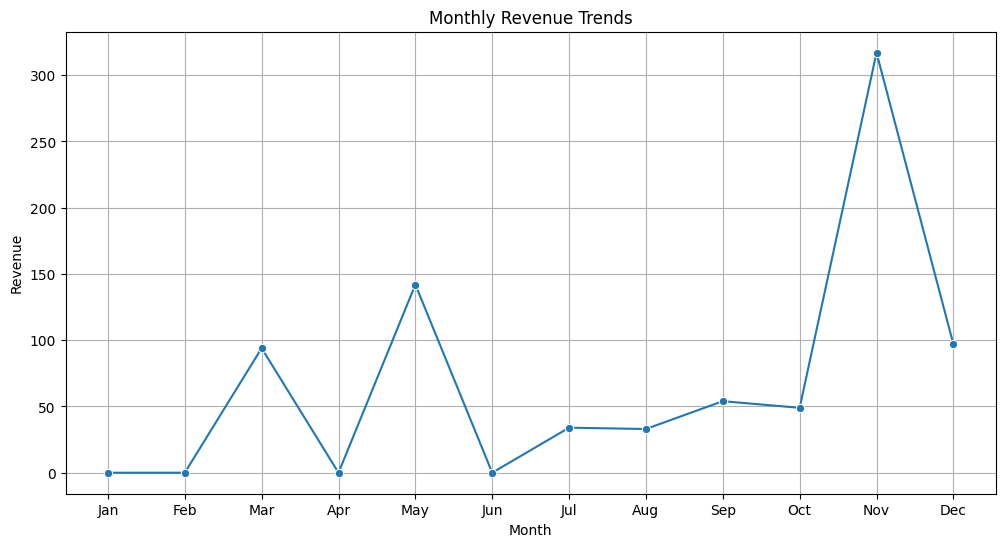
2) Bounce Rates by Revenue: Lower bounce rates correlate with higher revenue, suggesting the need for improved page load times and user navigation.

3) ProductRelated\_Duration vs. BounceRates:Successful sessions show lower bounce rates regardless of duration, highlighting the importance of quality engagement.

Monthly Revenue Trends

1) Seasonal Peaks: Significant revenue spikes occur in May and November, likely due to special shopping events.

2) Consistent Low Periods: Low revenue periods in January, February, April, and July suggest potential for targeted marketing efforts.



Predictive Analytics Studies

## Models Used

My analysis involves the training and the optimization of four different classification algorithms: Logistic Regression, Decision Tree Classifier, Support Vector Machine, and K-Nearest Neighbors. Those models are being evaluated based on these performance metrics: Accuracy, precision, recall, and F1 score.

1. Model Performance

* Logistic Regression:
  + Accuracy: 0.91
  + Precision: 0.65
  + Recall: 0.38
  + F1 Score: 0.48
* Decision Tree
  + Accuracy: 0.88
  + Precision: 0.44
  + Recall: 0.46
  + F1 Score: 0.45
* SVM
  + Accuracy: 0.92
  + Precision: 0.69
  + Recall: 0.46
  + F1 Score: 0.55
* KNN
  + Accuracy: 0.91
  + Precision: 0.63
  + Recall: 0.41
  + F1 Score: 0.50

1. Model Optimization:

Each of the models above underwent hyperparameter tuning using GridSearchCV to find the optimal parameters to improve the parameters. GridSearchCV works through multiple different combinations of values for parameters, and undergoes cross validation to determine which combination provides the best performance.

1. Model Performance

After performing GridSearchCV the results of the optimized models are as follows:

* Logistic Regression:
  + Accuracy: 0.91
  + Precision: 0.79
  + Recall: 0.68
  + F1 Score: 0.72
* Decision Tree:
  + Accuracy: 0.91
  + Precision: 0.75
  + Recall: 0.72
  + F1 Score: 0.73
* SVM:
  + Accuracy: 0.92
  + Precision: 0.79
  + Recall: 0.72
  + F1 Score: 0.75
* KNN:
  + Accuracy: 0.91
  + Precision: 0.79
  + Recall: 0.68
  + F1 Score: 0.72

Based on the metrics after optimization we can conclude that the model with the highest F1 score is SVM, therefor SVM is the best model based on the metrics

# Summary of Main Insights

## User Engagement

* Users tend to spend minimal time on administrative and informational pages, suggesting that these are not primary areas of interest or decision-making.
* Product-related pages, however, show higher engagement both in terms of visits and time spent, highlighting their importance in the user journey.

## Optimization Opportunities

* Administrative Pages: Streamlining administrative content to be more efficient could enhance user experience, as users tend to spend limited time here.
* Informational Pages: Improving the content or accessibility of informational pages could potentially increase engagement if these pages are deemed valuable.
* Product Pages: Since users spend most of their time on product-related pages, ensuring these pages are highly informative, visually appealing, and easy to navigate could further enhance user satisfaction and potentially drive conversions.

## Conversion Insights

* While the graphs do not directly show conversion rates, the high engagement on product-related pages suggests a strong correlation with revenue. Focusing on optimizing these pages could lead to better conversion outcomes.

## Seasonal Trends and Monthly Revenue

* Seasonal Peaks: Significant revenue spikes are observed during May and November, likely due to special shopping seasons and holidays. These months should be a focus for targeted marketing campaigns and inventory planning.
* Consistent Low Periods: Revenue is consistently lower in January, February, April, and July, indicating potential off-peak months. These periods could benefit from promotions and marketing efforts to boost sales.

# Conclusion

This study highlights critical user behavior features influencing revenue in e-commerce. By optimizing page engagement, leveraging seasonal trends, and enhancing new and returning visitor experiences, e-commerce platforms can significantly improve their revenue generation. Future work will focus on refining predictive models and exploring advanced machine learning techniques to further enhance accuracy and insights.