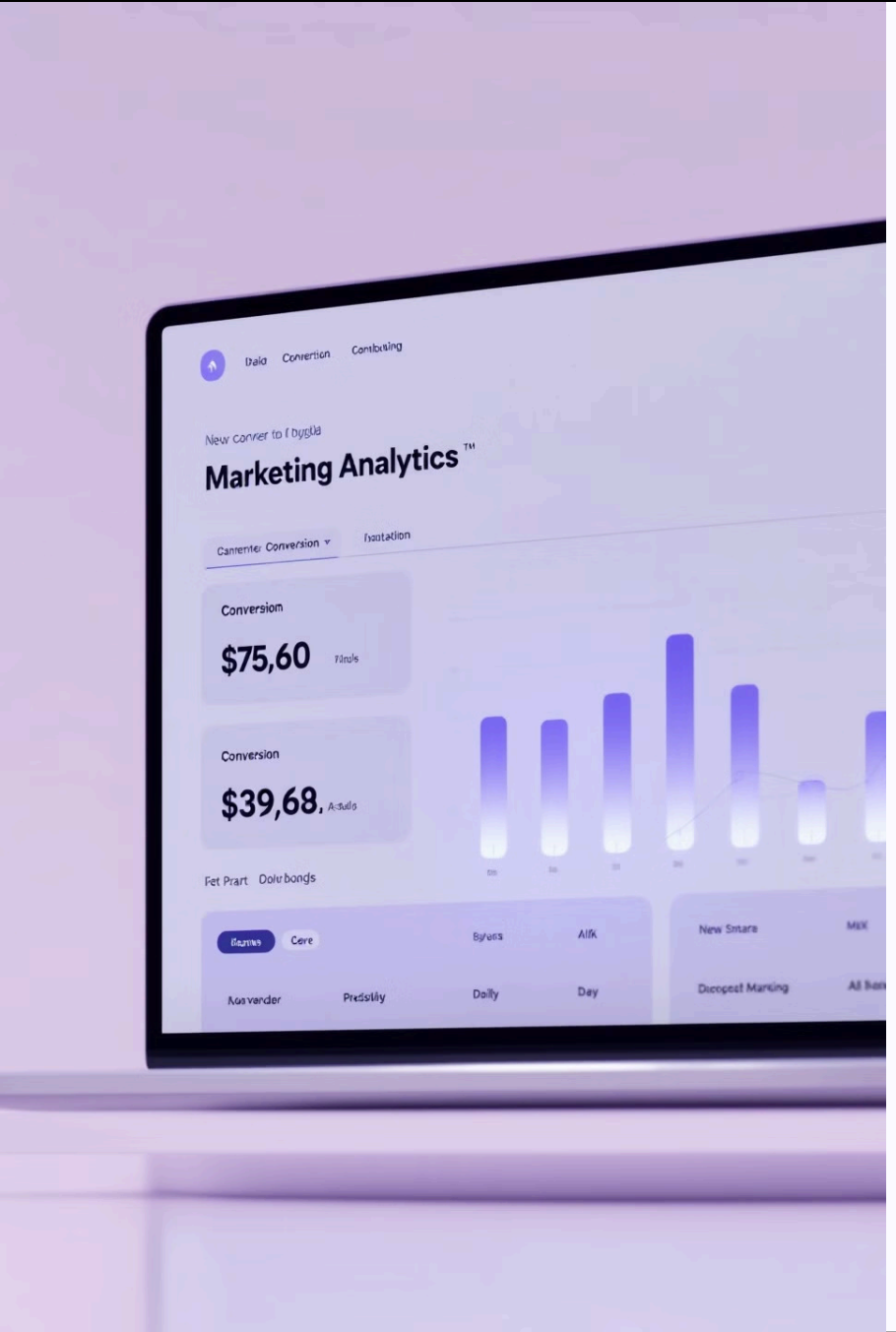


Optimizing Online Shopper Conversion

A predictive modeling approach using data mining techniques to identify online shoppers likely to convert.

Project Overview

1 Objective

Develop a predictive model to identify online shoppers likely to convert.

2 Dataset

Online Shoppers Purchasing Intention Dataset from UCI Machine Learning Repository.

3 Results

Final model achieved approximately 90% accuracy in predicting conversions.

E-commerce Challenges

Data Overload

Businesses struggle with vast amounts of heterogeneous digital interaction data.

Insight Extraction

Organizations find it difficult to extract actionable insights from user data.

Conversion Prediction

Accurately predicting which visitors will convert remains a significant challenge.

Research Foundation

1

Early Studies (2003)

Bucklin and Sismeiro demonstrated the value of clickstream data in distinguishing buyers from browsers.

2

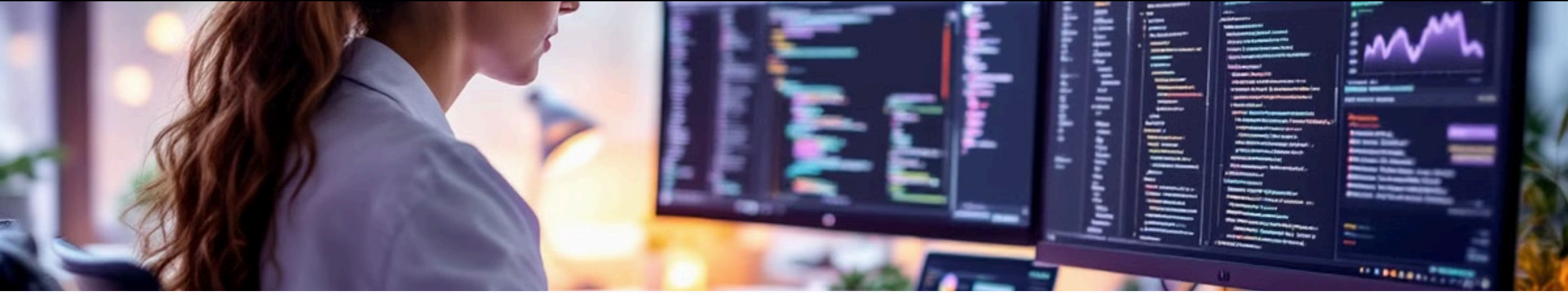
Mid-Stage Research (2016-2018)

Moro and Kumar applied various machine learning techniques to capture consumer behavior nuances.

3

Recent Advances

Deep learning and ensemble methods like random forests improved predictive accuracy.



Project Methodology

Data Collection & Preparation

Cleaning, transformation, and normalization of the dataset.

Exploratory Data Analysis

Statistical summaries and visualization to uncover trends and relationships.

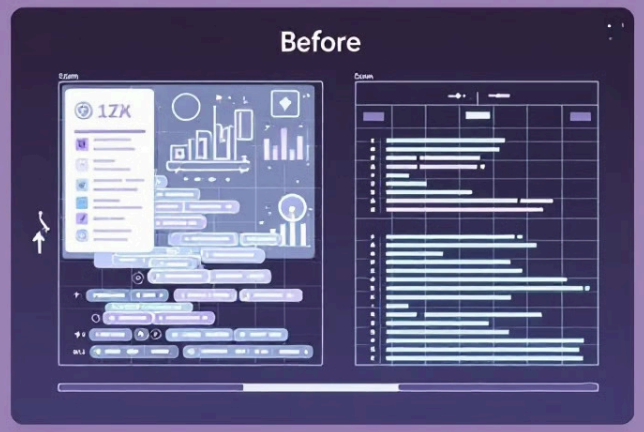
Feature Engineering & Selection

Creating interaction features and applying selection methods.

Model Training & Tuning

Training multiple classifiers with hyperparameter optimization.

Data cleaning



After

Data Preparation Process

1

Duplicate Removal

Eliminating redundant entries to ensure data integrity.

2

Missing Value Imputation

Filling gaps in the dataset with statistically appropriate values.

3

Categorical Encoding

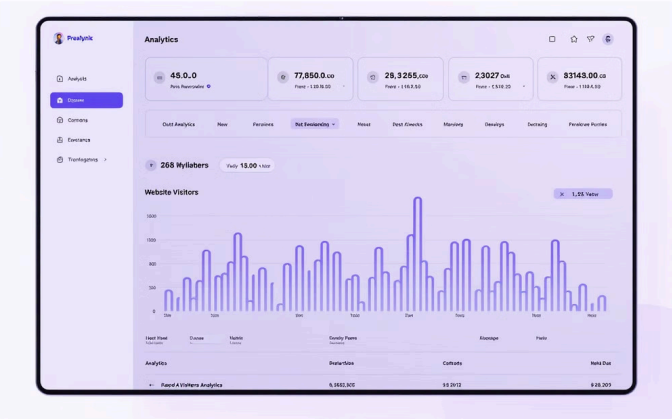
Converting categorical variables through one-hot encoding.

4

Numerical Normalization

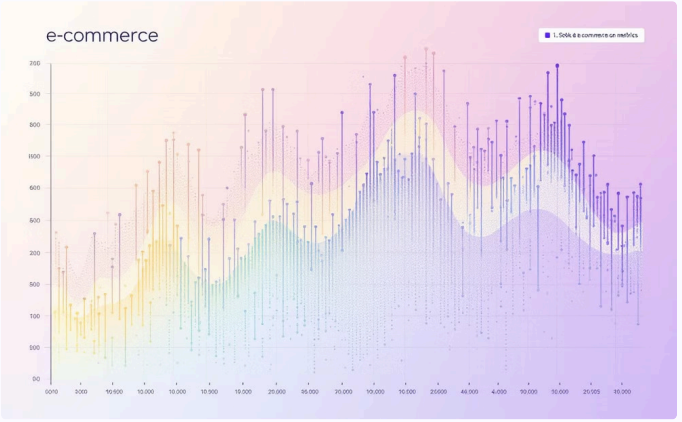
Scaling numerical features to ensure consistency across the dataset.

Exploratory Data Analysis



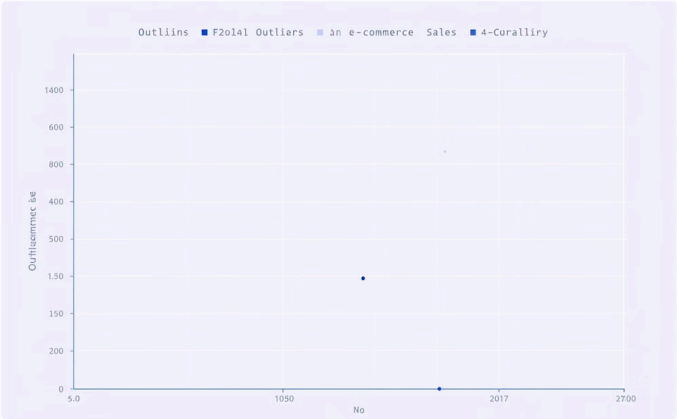
Statistical Summaries

Generated descriptive statistics to understand data distributions and outliers.



Correlation Analysis

Identified relationships between variables using heatmaps and correlation matrices.



Outlier Detection

Used boxplots to identify and address extreme values in the dataset.

Feature Engineering Innovations

Interaction Features

Created new features by combining PageValues and ExitRates to capture relationship effects.

Variable Binning

Binned continuous variables to capture non-linear relationships in the data.

Feature Selection

Applied correlation-based methods and model-based importance to retain top predictors.

Model Training Approach

Algorithm Selection

Trained multiple classifiers including logistic regression and random forests.

Each algorithm was evaluated for its strengths in handling the specific dataset characteristics.

Hyperparameter Tuning

Used grid search to fine-tune model parameters for optimal performance.

Systematically explored parameter combinations to maximize classification accuracy.

Validation Strategy

Implemented cross-validation to ensure model robustness and prevent overfitting.

Reserved 30% of data as a hold-out test set for final evaluation.

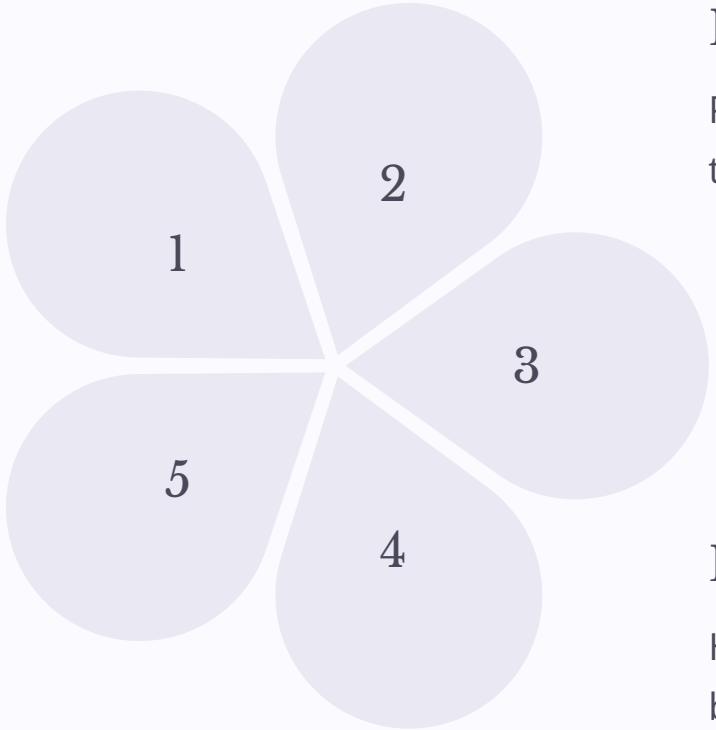
Evaluation Metrics

Accuracy

Overall correctness of predictions, reaching approximately 90%.

ROC-AUC

Area under the ROC curve, measuring discrimination capability.



Precision

Proportion of positive identifications that were actually correct.

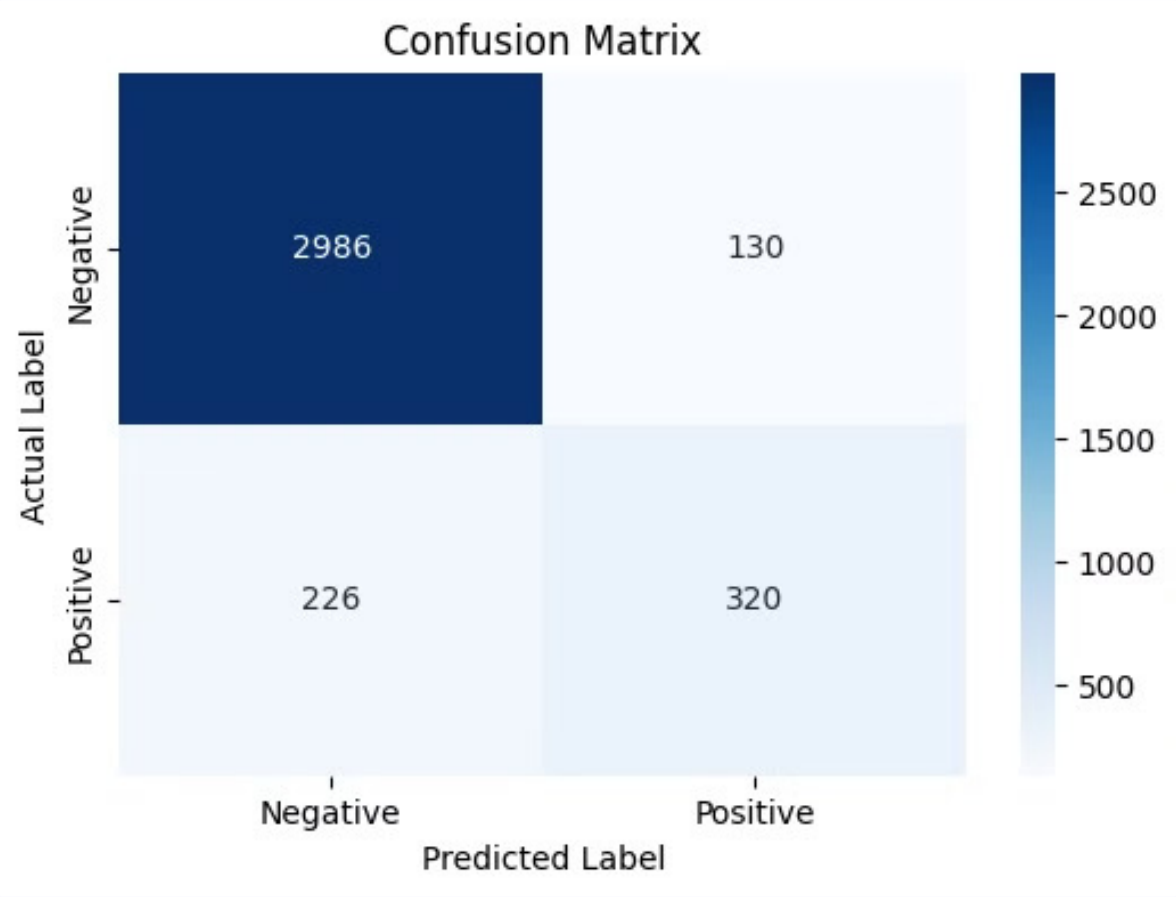
Recall

Proportion of actual positives that were correctly identified.

F1-Score

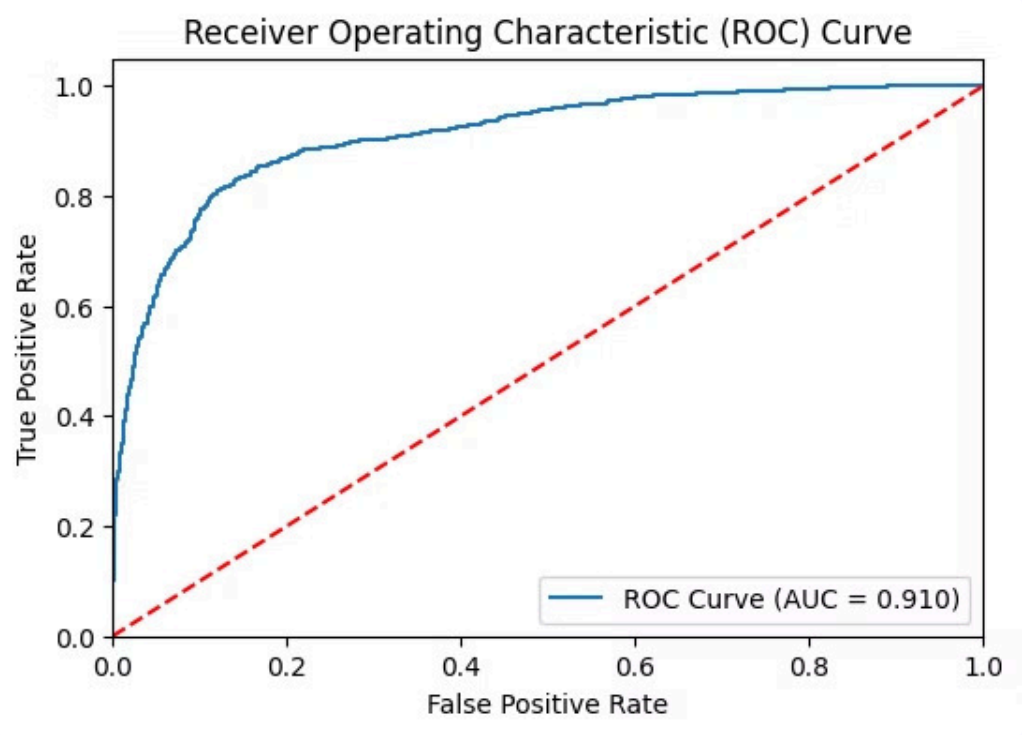
Harmonic mean of precision and recall, balancing both metrics.

Confusion Matrix Results



The confusion matrix shows the model effectively distinguishes between classes with minimal misclassifications.

ROC Curve Analysis

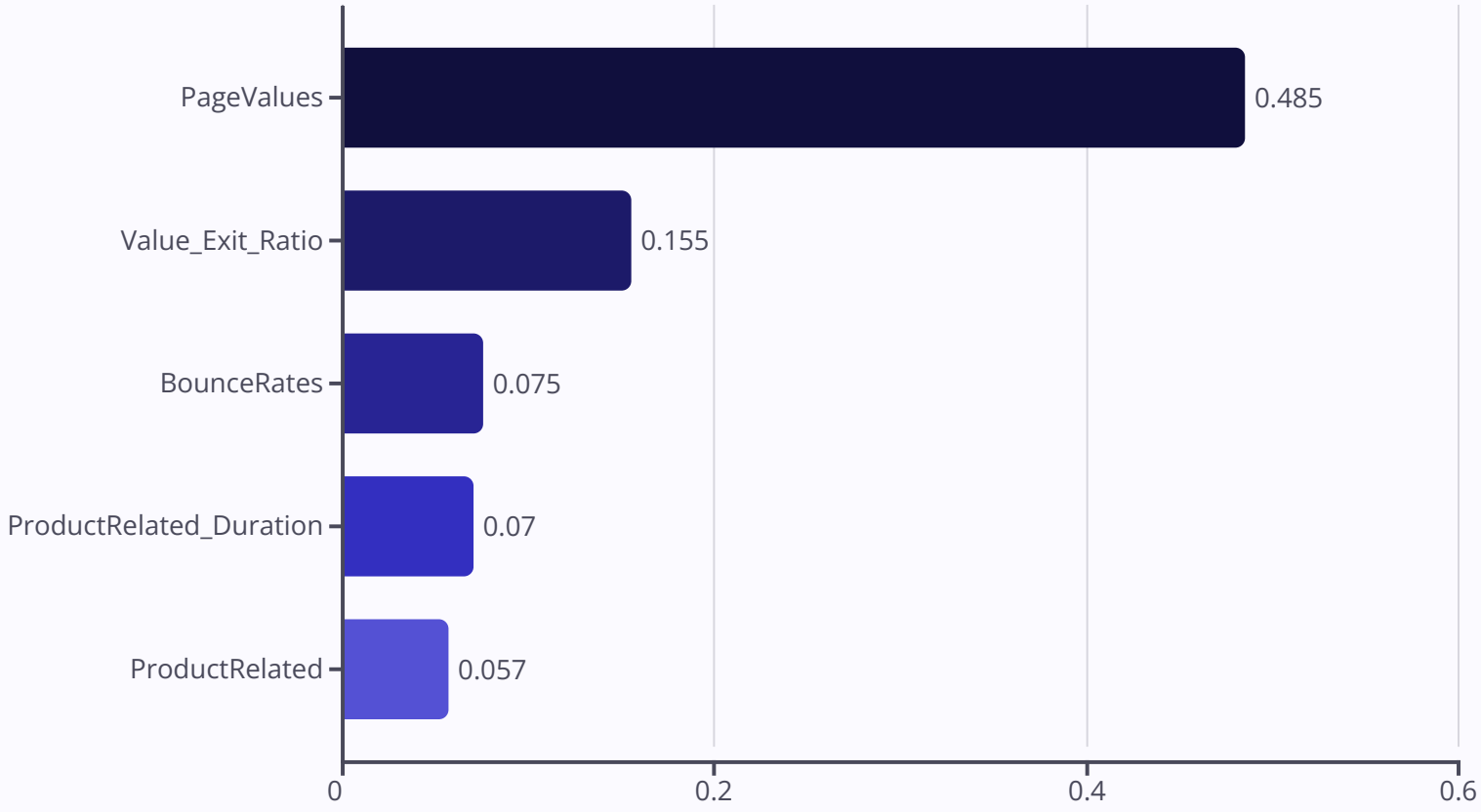


The ROC curve demonstrates the model's ability to distinguish between positive and negative classes.

A high area under the curve confirms the model's strong discrimination capability.

AUC = 0.91

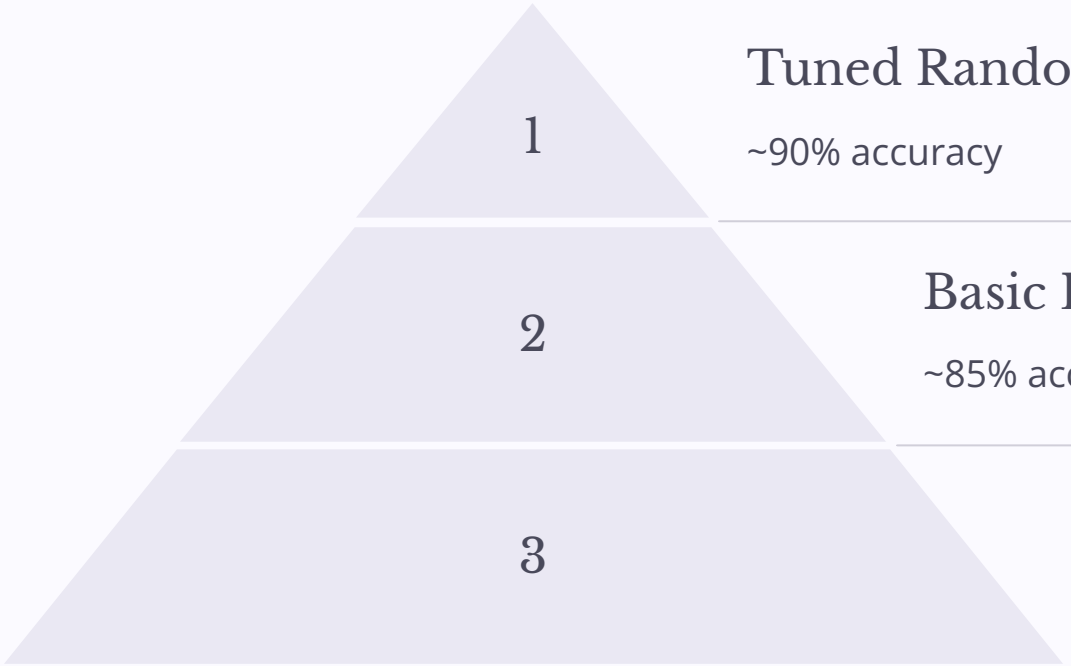
Key Predictors of Conversion



Feature importance analysis revealed PageValues as the most significant predictor of conversion.

Engineered features like Value_Exit_Ratio proved highly valuable in improving predictions.

Model Comparison



Tuned Random Forest

~90% accuracy

Basic Random Forest

~85% accuracy

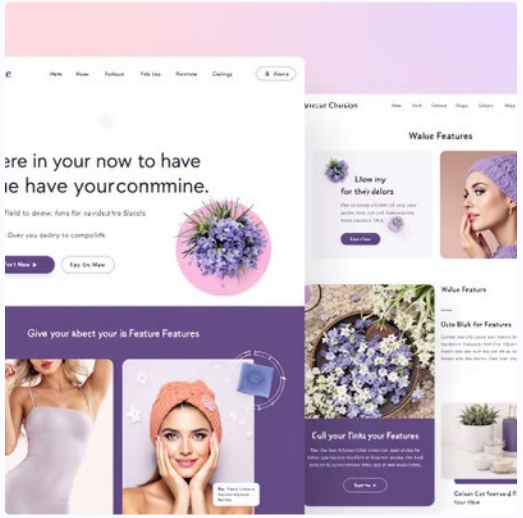
Logistic Regression

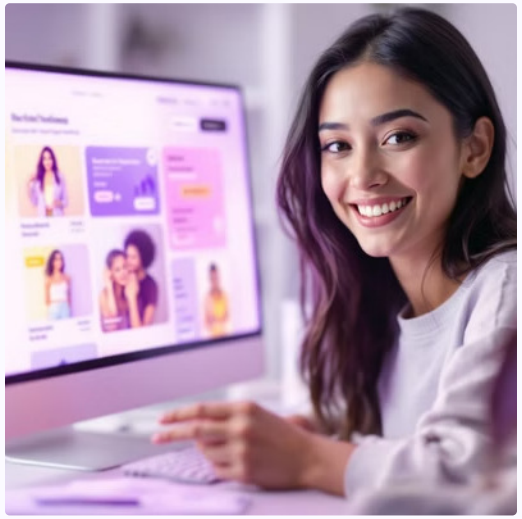
~80% accuracy

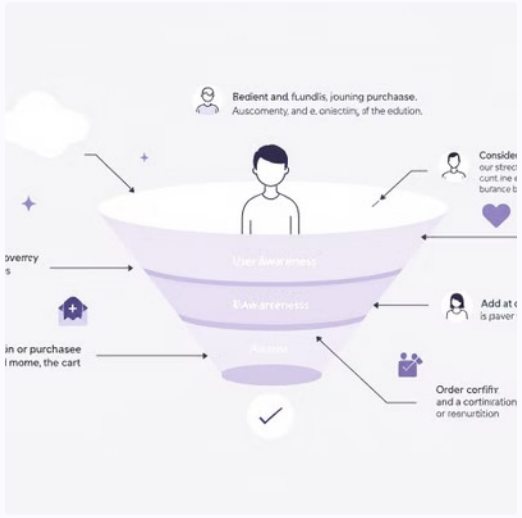
The tuned random forest outperformed other models, demonstrating the value of ensemble methods and hyperparameter optimization.

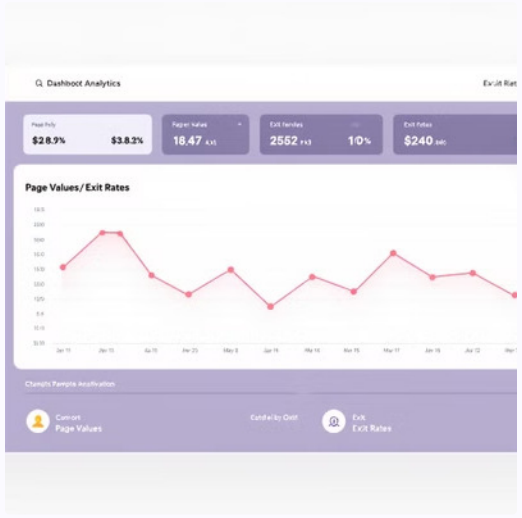
Each model showed progressively better performance with added complexity and tuning.

Key Insights









PageValues emerged as the critical predictor, suggesting content strategy should focus on enhancing high-value pages.

Interaction features significantly improved model performance, capturing complex relationships between metrics.

Model Trade-offs

1

Performance

90% accuracy with tuned random forest

2

Complexity

Ensemble methods increase computational requirements

3

Interpretability

Feature importance partially mitigates "black box" nature

4

Scalability

May require optimization for larger datasets

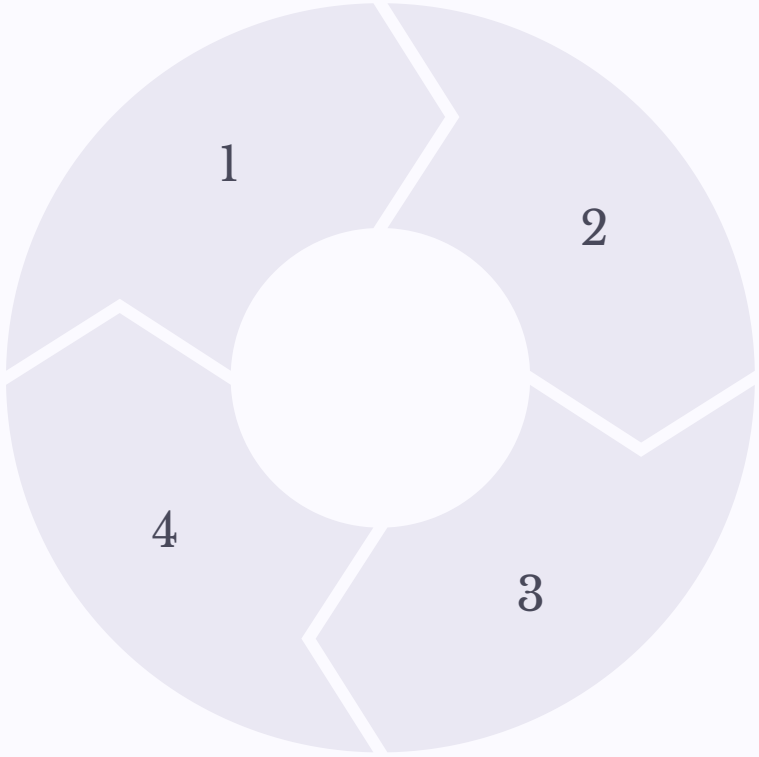
Actionable Recommendations

Optimize High-Value Pages

Focus on pages with strong conversion potential.

Personalize User Experience

Tailor content based on predictive insights.



Reduce Exit Rates

Improve engagement on pages with high bounce rates.

Enhance Product Pages

Increase time spent on product-related content.

Future Improvements

1. **Scale to Larger Datasets**
   Explore optimization techniques for handling bigger e-commerce datasets.
2. **Integrate Real-Time Data**
   Incorporate streaming analytics for dynamic conversion prediction.
3. **Explore Deep Learning**
   Investigate neural networks to capture more complex patterns.
4. **Expand Feature Engineering**
   Develop more sophisticated interaction features and transformations.

Conclusion

90%

Accuracy

Final model performance

#1

PageValues

Top conversion predictor

4+

Key Features

Critical predictive factors

The tuned random forest model achieved impressive accuracy, validating our approach to preprocessing and feature engineering.

PageValues emerged as the most critical predictor, offering clear direction for marketing optimization.

References

- Bucklin & Sismeiro (2003). A model of website browsing behavior estimated on clickstream data.
- Guo, Sun, & Xu (2021). Enhancing e-commerce conversion predictions with engineered user behavior features.
- Kumar, Gupta, & Li (2018). Factors influencing online purchase decisions: A machine learning perspective.
- Moro, Cortez, & Rita (2016). A data-driven approach to predict the success of bank telemarketing.
- Zhang, Zhang, & Yuan (2019). Predicting online shopping behavior using deep learning.