Predicting Online Shopper Purchase Intentions Using Machine Learning

This presentation explores how machine learning can predict which online shoppers are likely to make a purchase, enabling e-commerce teams to deliver personalized experiences and targeted offers. Using the CRISP-DM framework, we'll examine visitor behavior patterns, identify key conversion factors, and demonstrate how predictive models can significantly improve conversion rates.

by Collins Nyagaka

Key Objectives

Our analysis aims to leverage machine learning to predict purchase intentions, helping e-commerce businesses optimize their conversion strategies.

Identify Key Purchase Predictors

Discover which visitor behaviors most strongly signal purchase intent.

Develop Accurate Prediction Models

Create reliable machine learning models to forecast customer actions.

Deliver Actionable Insights

Translate predictions into practical strategies for increasing conversion rates.

Business Understanding: The Challenge

Predict Purchase Intention

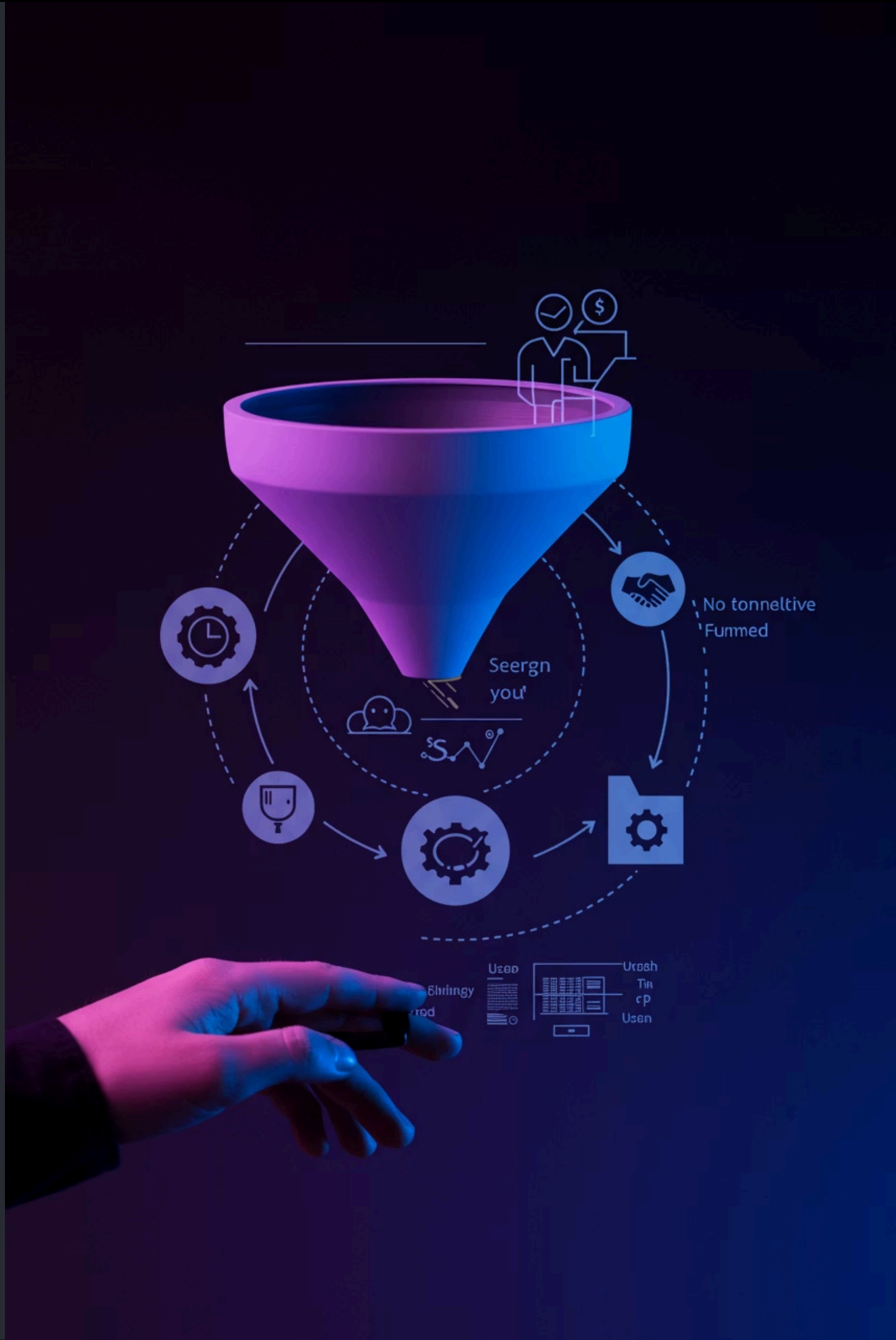
Identify which website visitors are likely to complete a purchase, enabling tailored experiences and offers

Improve Conversion Rates

E-commerce industry benchmarks show 2-3% conversion rates, while our dataset shows 15.5% conversion

Key Stakeholders

E-commerce product managers and marketers focused on improving conversion through visitor behavior analysis



Dataset Overview

Dataset Composition

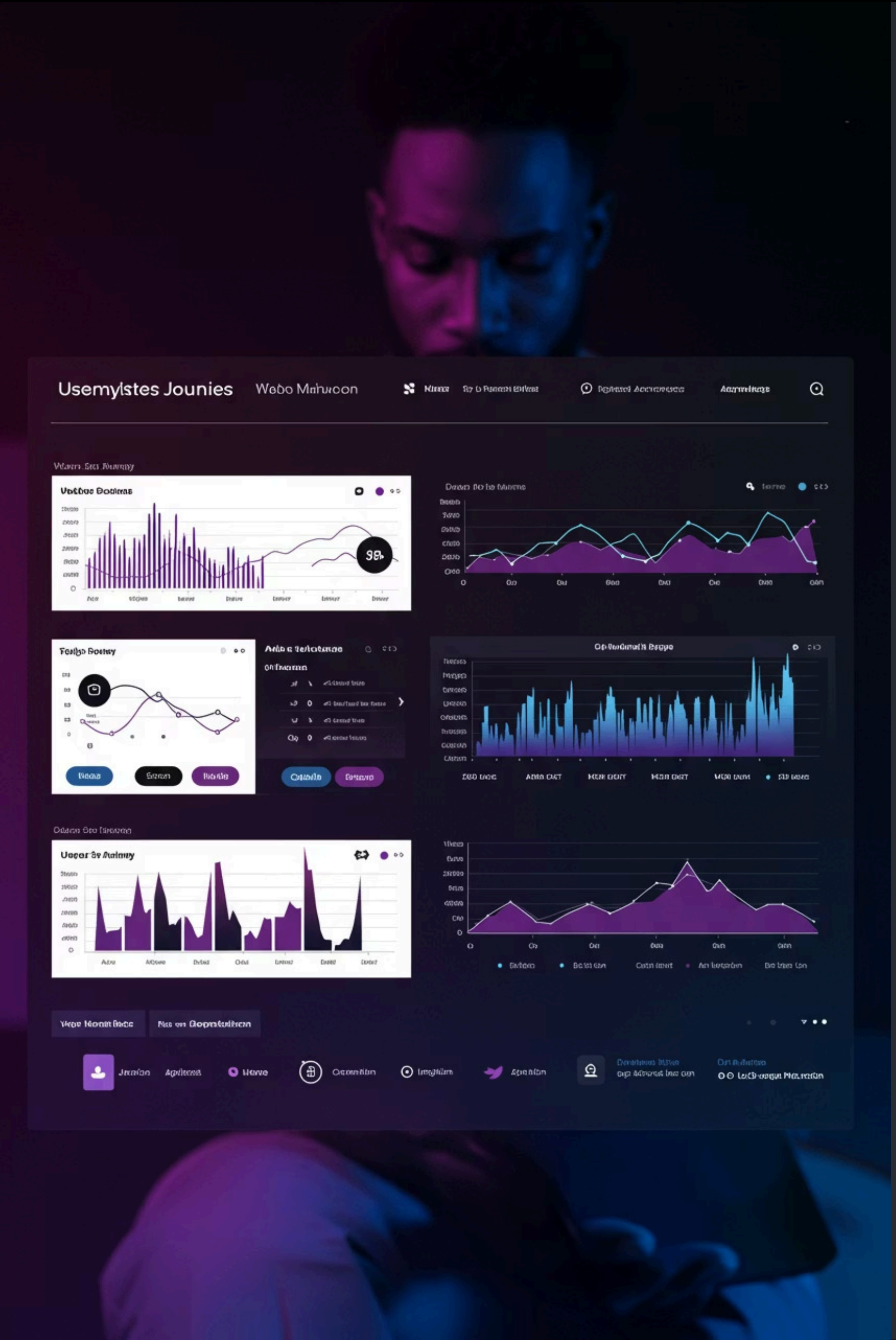
12,330 unique user sessions with 18 attributes capturing various aspects of online shopping behavior

- 10 numerical attributes
- 8 categorical attributes
- Binary target: Revenue (purchase vs. no purchase)

Class Distribution

The dataset shows class imbalance typical of e-commerce:

- 84.5% sessions (10,422) did not end with purchase
- 15.5% sessions (1,908) ended with purchase



Key Features in the Dataset

Page Interaction Metrics

- Administrative pages (count & duration)
- Informational pages (count & duration)
- Product-related pages (count & duration)

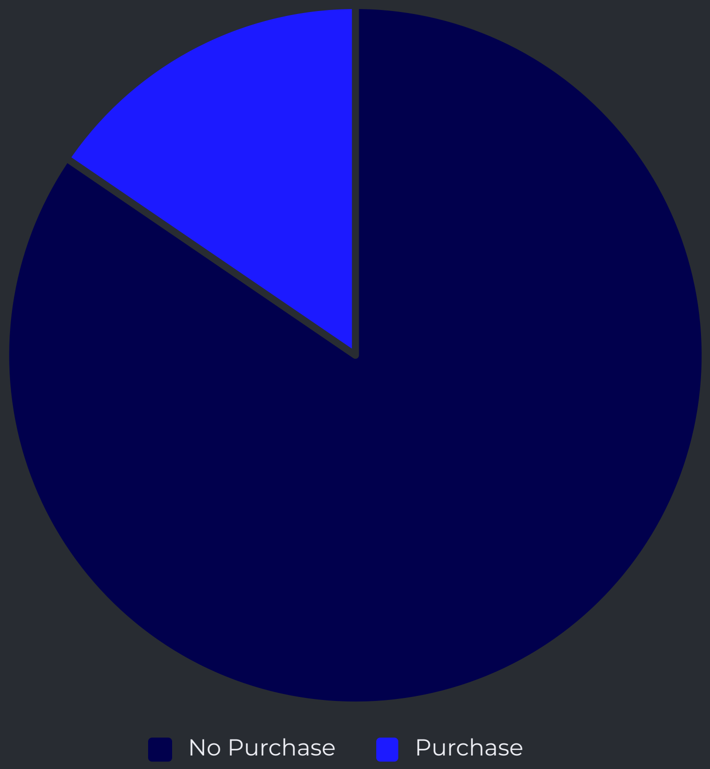
Google Analytics Metrics

- Bounce Rate: % of visitors leaving after one page
- Exit Rate: % where page was last in session
- Page Value: Average value of page before transaction

Contextual Information

- Special Day: Proximity to gift-giving occasions
- Month, Weekend status
- Visitor Type (new/returning)
- Browser, OS, Region, Traffic Type

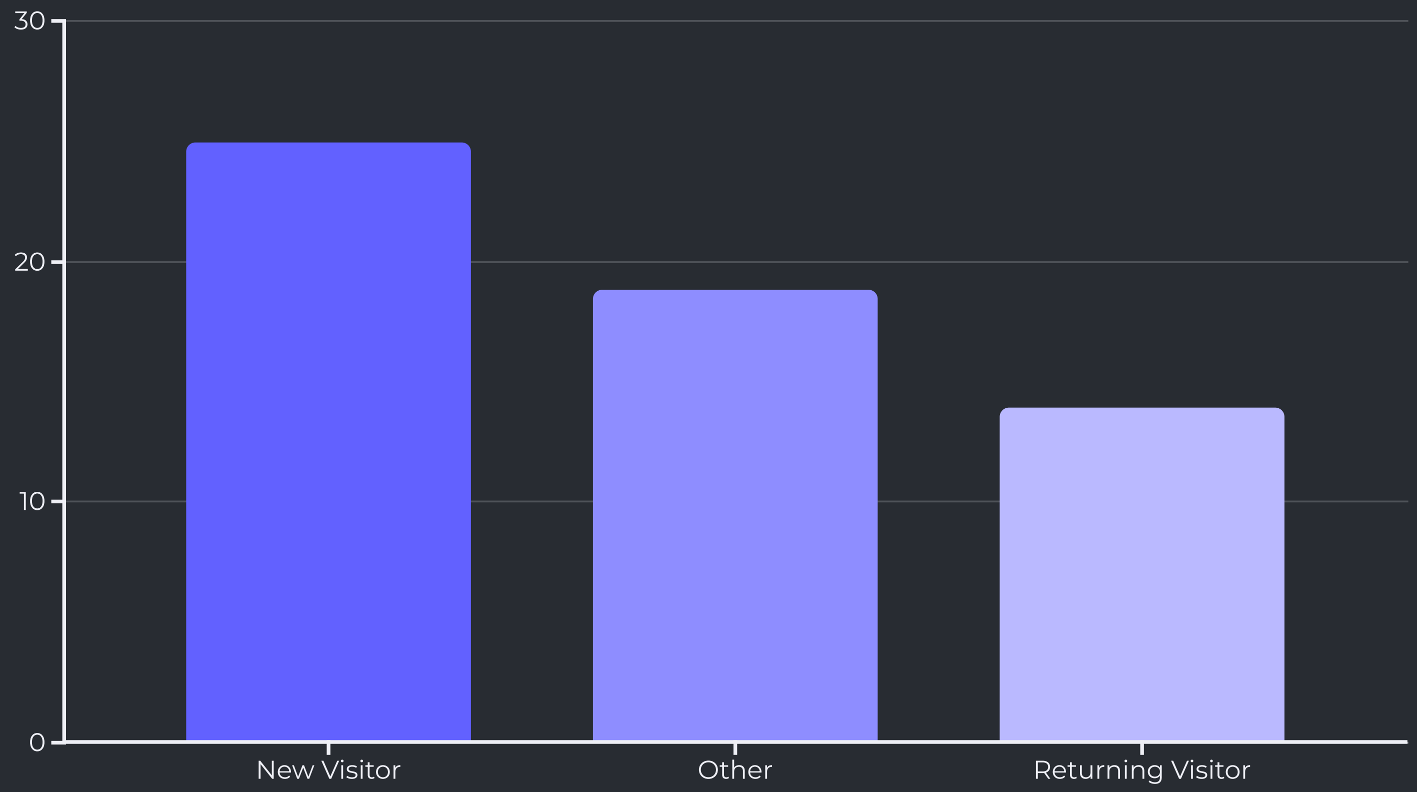
Data Exploration: Class Imbalance



Our dataset reveals a substantial class imbalance, with only 15.5% of sessions culminating in a purchase. While this imbalance reflects typical e-commerce conversion patterns, it creates significant challenges for predictive modeling. Without proper handling, models tend to favor the dominant class (no purchase), potentially overlooking valuable conversion signals from the minority class.

To overcome this challenge, our modeling strategy will incorporate specialized techniques such as synthetic sampling, adjusted class weights, or algorithm-specific optimizations. These approaches will ensure our model remains sensitive to purchase indicators despite their relative scarcity in the dataset, ultimately maximizing the identification of conversion opportunities.

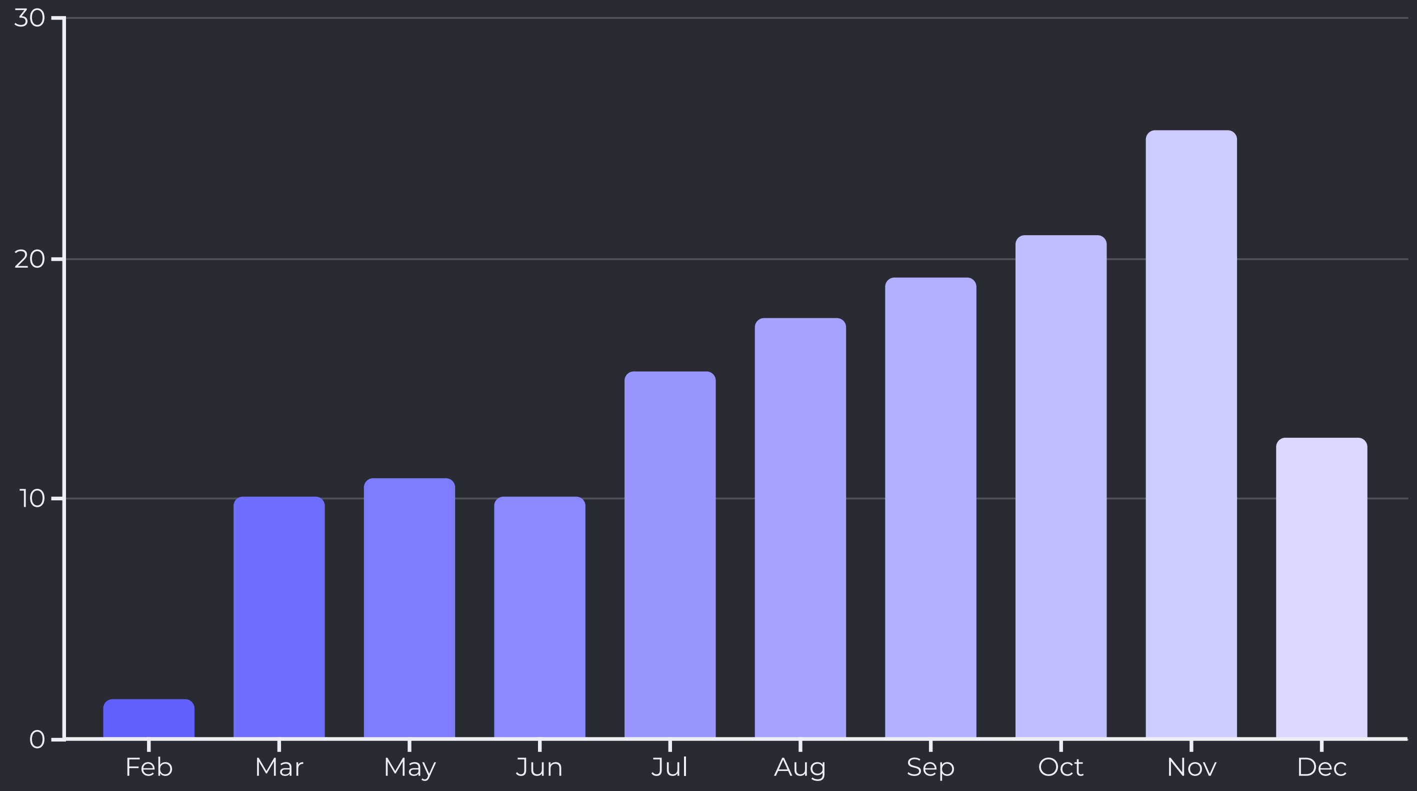
Conversion Rates by Visitor Type



Analysis reveals significant differences in conversion rates across visitor types. New visitors convert at nearly twice the rate (24.91%) of returning visitors (13.93%). This counterintuitive finding suggests that first-time visitors may be more motivated to complete purchases, possibly arriving with stronger purchase intent or responding better to initial offers.

This insight has important implications for marketing strategy, suggesting that acquisition efforts may yield higher immediate returns than retention campaigns in this particular context.

Seasonal Conversion Patterns



Conversion rates show clear seasonal patterns, starting low in February (1.63%) and steadily increasing throughout the year to peak in November (25.35%) before declining in December. This pattern aligns with typical retail seasonality, with the highest conversion rates during the pre-holiday shopping season.

These insights can guide marketing budget allocation and campaign timing, suggesting increased investment during high-conversion months and potential need for conversion optimization during low-performing periods.

Behavioral Differences: Purchasers vs. Non-Purchasers

Page Visits

Purchasers visit significantly more pages:

- Administrative: 3.39 vs. 2.12 pages
- Informational: 0.79 vs. 0.45 pages
- Product-related: 48.21 vs. 28.71 pages
- Total pages: 52.39 vs. 31.28 pages

Time Spent

Purchasers spend more time on site:

- Admin duration: 119.48s vs. 73.74s
- Info duration: 57.61s vs. 30.24s
- Product duration: 1876.21s vs. 1069.99s
- Total duration: 2053.30s vs. 1173.96s

Engagement Metrics Comparison

73%

Higher Page Views

Purchasers view 73% more pages than non-purchasers

75%

Longer Sessions

Purchasers spend 75% more time on the site

67%

Lower Bounce Rate

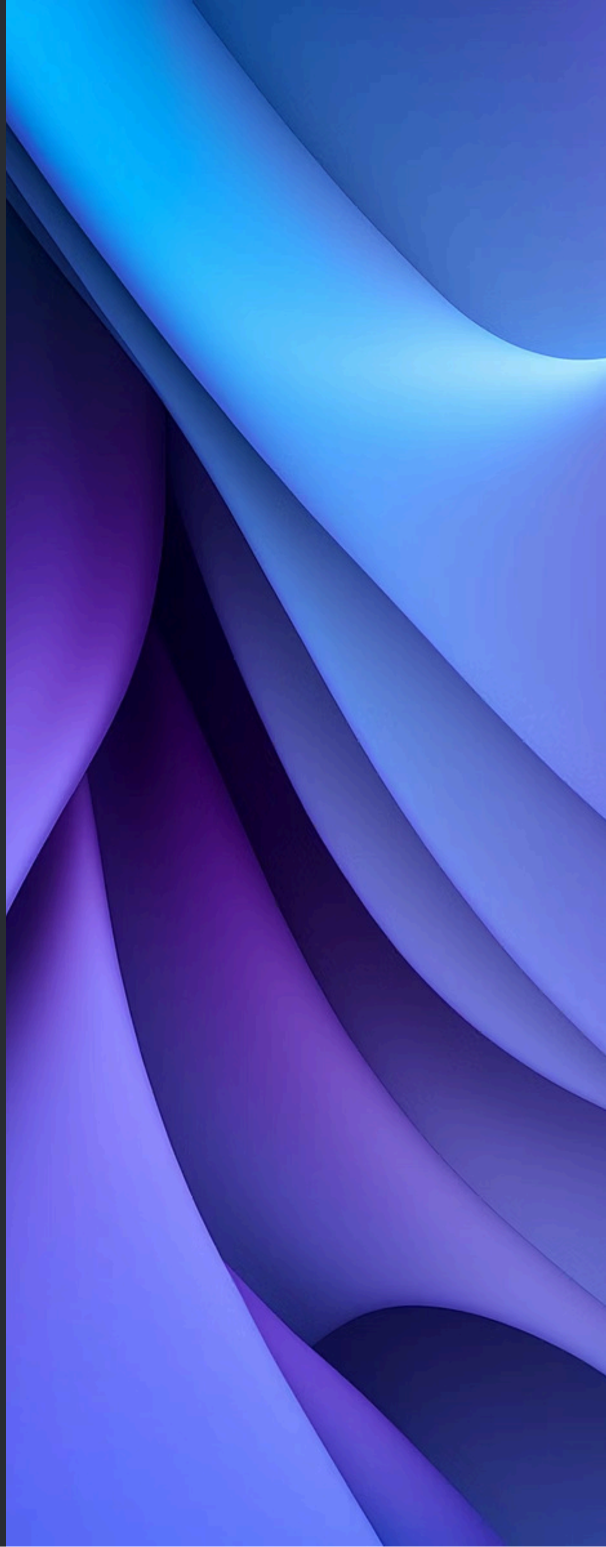
Purchasers have 67% lower bounce rates

13.8x

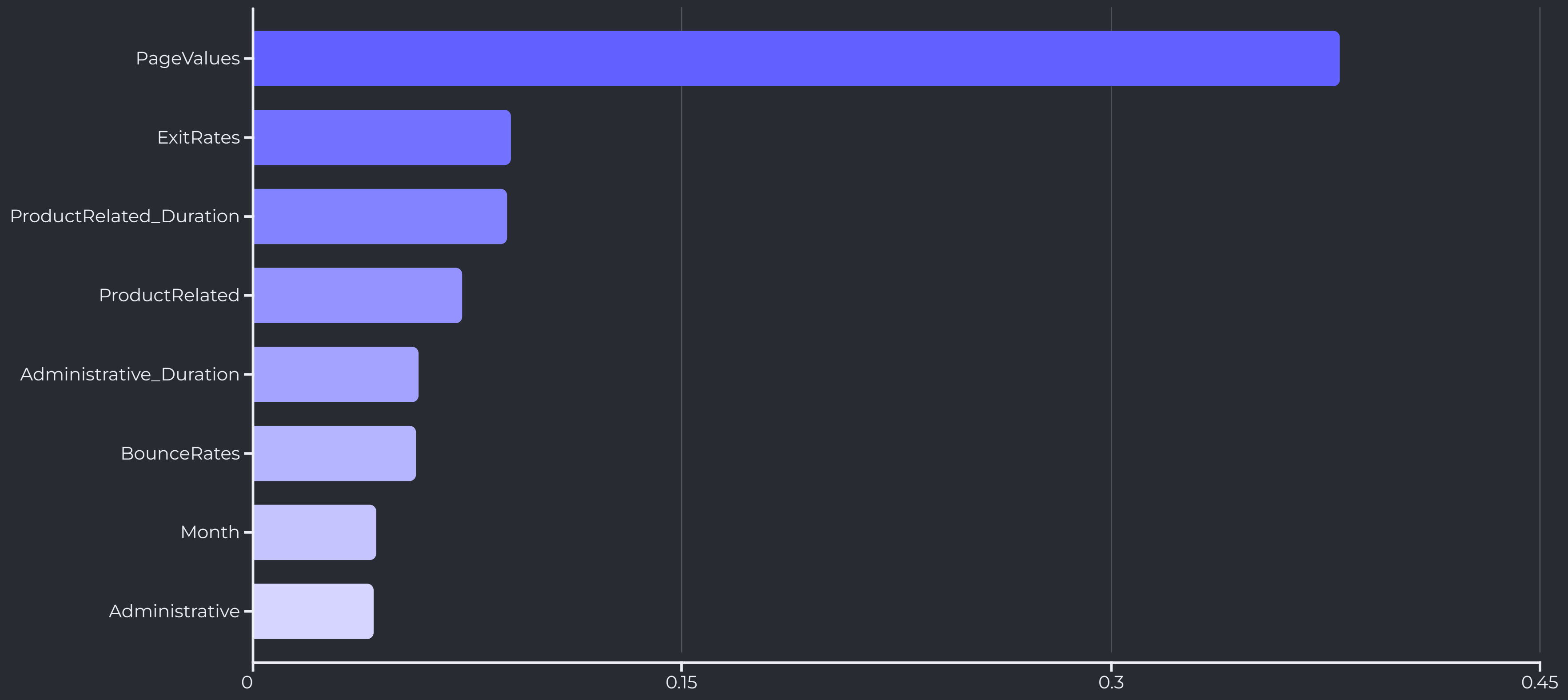
Higher Page Value

Purchasers generate 13.8x higher page values

The data reveals striking differences in engagement metrics between purchasers and non-purchasers. Purchasers demonstrate significantly deeper engagement across all metrics, with the most dramatic difference in Page Values (27.26 vs. 1.98). These statistically significant differences ($p < 0.001$) confirm that user engagement is strongly linked to purchase likelihood.



Feature Importance Analysis



Random Forest feature importance analysis reveals PageValues as the dominant predictor of purchase behavior, with an importance score of 0.38—over four times higher than the next feature. This metric, which measures the average value of a page viewed before completing a transaction, serves as a powerful signal of purchase intent.

Other significant predictors include ExitRates, ProductRelated_Duration, and ProductRelated page count, confirming that engagement with product content strongly influences purchase decisions.

Modeling Approach

Data Preparation

Encoded categorical variables, scaled numerical features, and split data into training (80%) and testing (20%) sets while preserving class distribution

Baseline Models

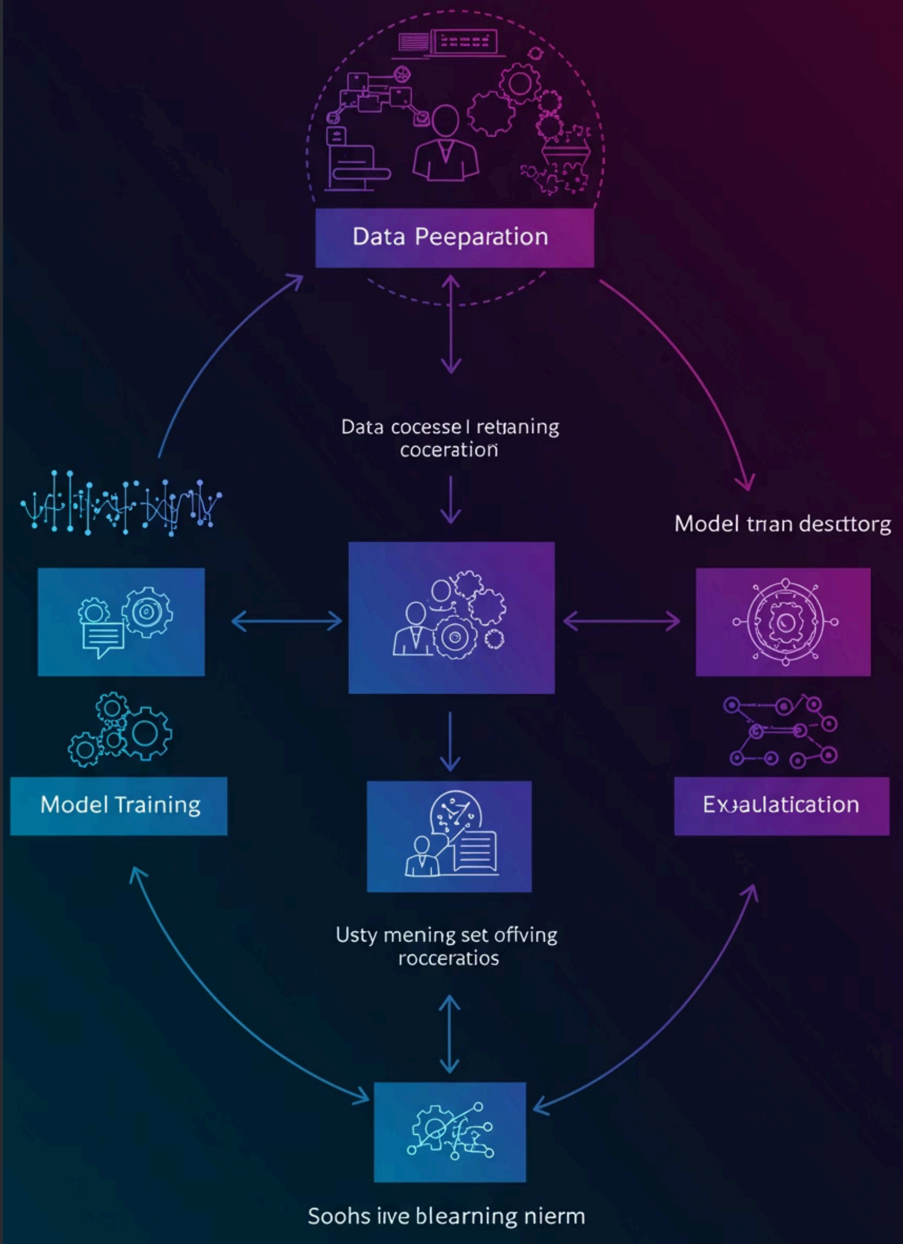
Trained multiple classification algorithms: Logistic Regression, SVM, Decision Tree, Random Forest, Gradient Boosting, and AutoGluon (automated ML)

Advanced Modeling

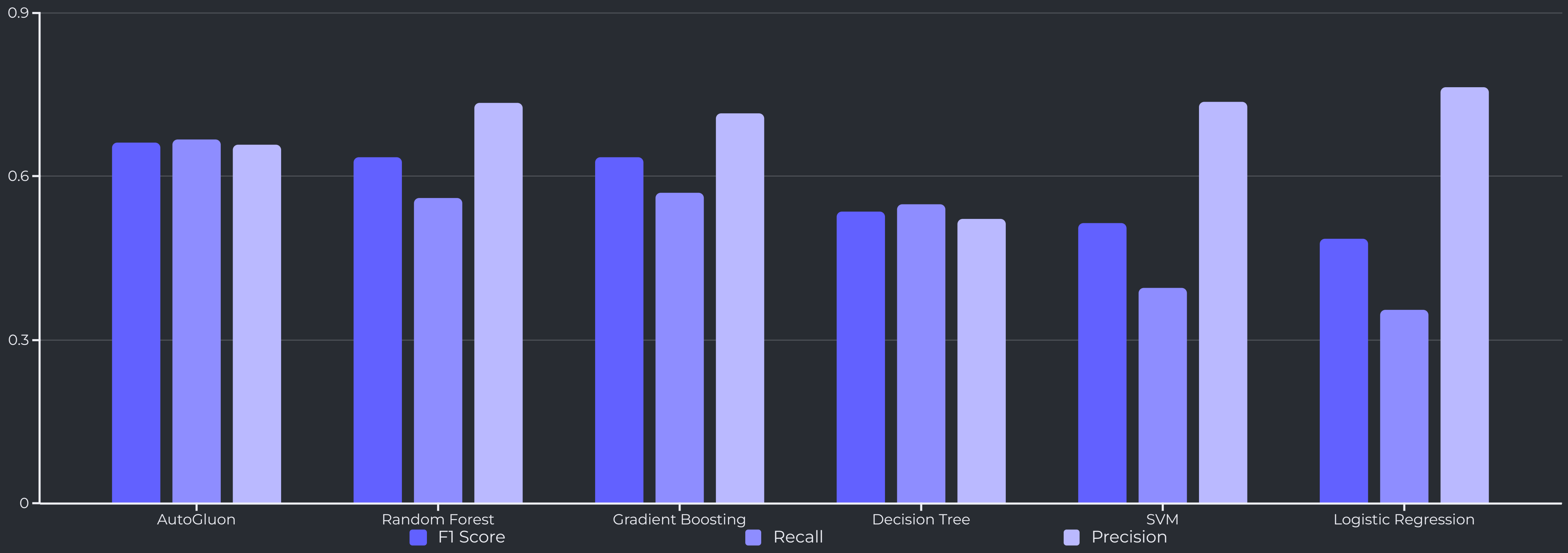
Addressed class imbalance with SMOTE, engineered new features, performed hyperparameter tuning, and created ensemble models

Evaluation

Assessed models using accuracy, precision, recall, F1 score, and ROC AUC, with emphasis on F1 score to balance precision and recall


Data Peeparation
Data cocesse l retıaning coceration
Model trıan desctorg
Model Training
Exɔaulatication
Usty mening set offving rocceratios
Soohs iıve blearning nierm

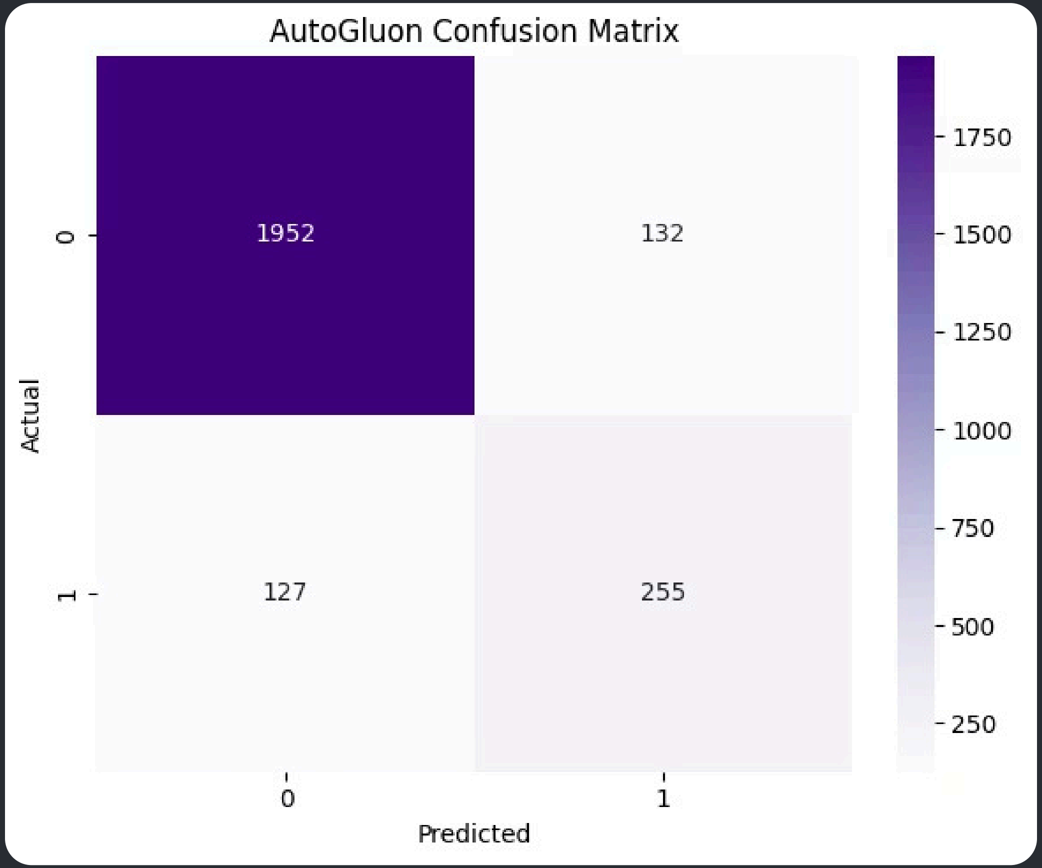
Baseline Model Performance



Initial model comparison shows AutoGluon leading with the highest F1 score (0.663) and balanced precision-recall trade-off. Random Forest and Gradient Boosting follow closely, while simpler models like Decision Tree, SVM, and Logistic Regression trail significantly.

SVM and Logistic Regression show high precision but poor recall, indicating they miss many potential purchasers. The ensemble methods (AutoGluon, Random Forest, Gradient Boosting) demonstrate superior ability to identify purchasers while maintaining reasonable precision.

Confusion Matrix Analysis



AutoGluon Performance

The confusion matrix reveals:

- True Negatives: 1952 (correctly identified non-buyers)
- False Positives: 132 (non-buyers predicted as buyers)
- False Negatives: 127 (missed buyers)
- True Positives: 255 (correctly identified buyers)

AutoGluon correctly identifies 66.8% of actual buyers while maintaining a precision of 65.9%, offering the best balance between capturing purchase opportunities and minimizing wasted marketing efforts.

Advanced Feature Engineering

Aggregate Metrics

Created Total_Pages and Total_Duration features to capture overall engagement level

Interaction Terms

Generated interaction features between page counts and durations to capture engagement depth

Class Balancing

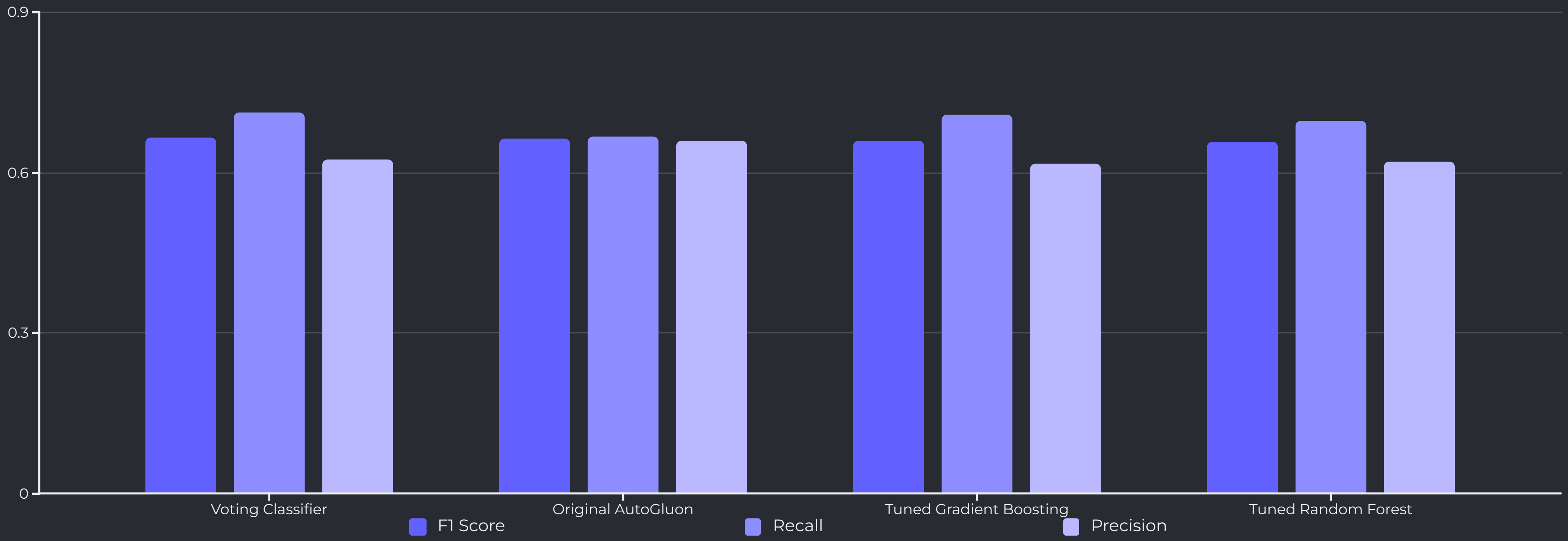
Applied SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance

Hyperparameter Tuning

Optimized model parameters using RandomizedSearchCV to maximize F1 score

These advanced techniques improved model performance by creating more informative features and addressing the class imbalance challenge inherent in e-commerce conversion prediction.

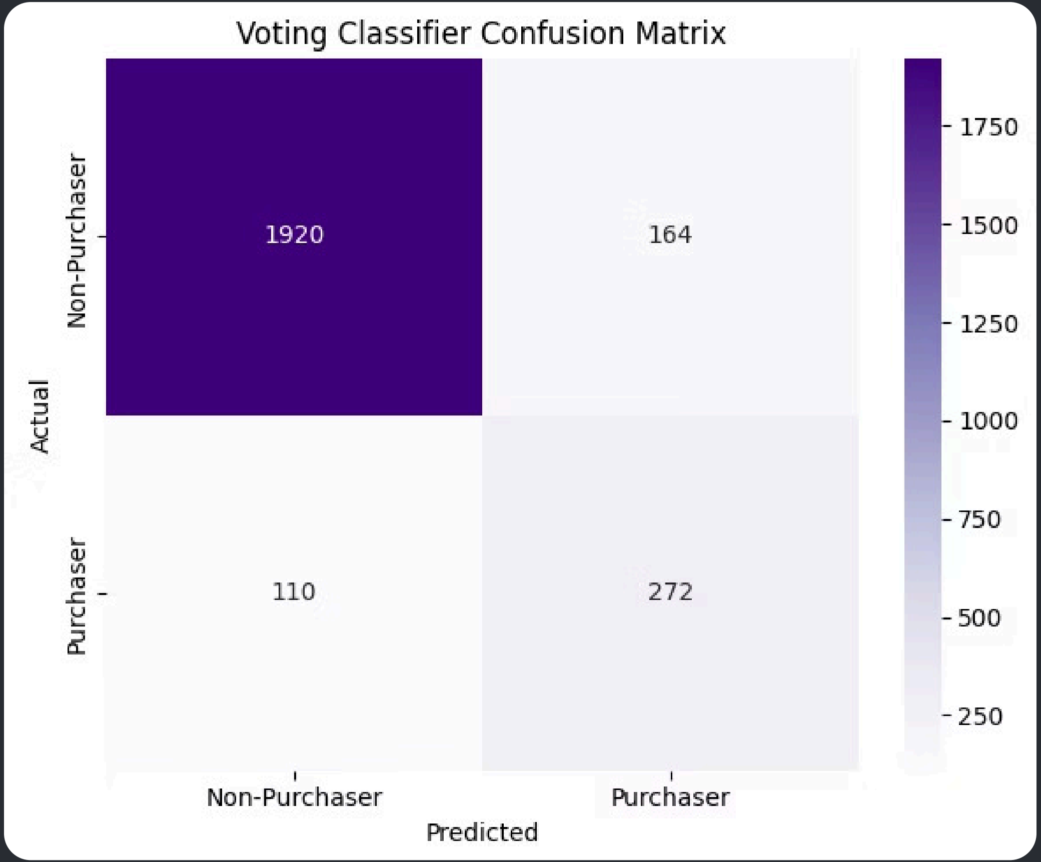
Enhanced Model Performance



After implementing advanced techniques, the Voting Classifier (ensemble of tuned Random Forest and Gradient Boosting) achieved the highest F1 score (0.665) and significantly improved recall (0.712), identifying 71.2% of actual purchasers.

The enhanced models show a strategic shift toward higher recall at a slight cost to precision, enabling the identification of more potential buyers. This trade-off is often desirable in e-commerce, where missing a potential customer (false negative) typically costs more than targeting a non-buyer (false positive).

Optimized Model Confusion Matrix



Voting Classifier Performance

The optimized Voting Classifier shows:

- True Negatives: 1920 (correctly identified non-buyers)
- False Positives: 164 (non-buyers predicted as buyers)
- False Negatives: 110 (missed buyers)
- True Positives: 272 (correctly identified buyers)

Compared to baseline models, the enhanced Voting Classifier identifies 17 more purchasers (272 vs. 255) while missing 17 fewer potential buyers (110 vs. 127), demonstrating meaningful improvement in capturing conversion opportunities.

Key Business Insights

Engagement Predicts Purchases

Deeper engagement (more pages, longer duration) strongly correlates with purchase likelihood, with purchasers spending 75% more time on site

New Visitors Convert Better

New visitors convert at nearly twice the rate (24.9%) of returning visitors (13.9%), suggesting strong initial purchase intent

Clear Seasonal Patterns

Conversion rates peak in October-November (20-25%) and are lowest in February (1.6%), indicating optimal timing for campaigns

Page Value Is Critical

PageValues is by far the strongest predictor of purchases, with 13.8x higher values for sessions that convert

Business Recommendations

Optimize for High-Value Engagement

Improve product pages with personalized recommendations, detailed content, and smooth navigation to increase PageValues and time spent

Tailor Strategies by Visitor Type

Create compelling first impressions for new visitors while developing personalized retention strategies for returning visitors

Capitalize on Seasonal Trends

Align marketing budgets and campaigns with seasonal conversion patterns, focusing resources on high-conversion months

Implement Predictive Targeting

Deploy the Voting Classifier model to identify high-intent visitors in real-time and deliver personalized experiences

Implementation Roadmap

Phase 1: Model Deployment (1-2 months)

Implement the Voting Classifier model into the production environment with real-time scoring capabilities

Phase 2: A/B Testing (2-3 months)

Conduct controlled experiments comparing model-driven personalization against current approaches

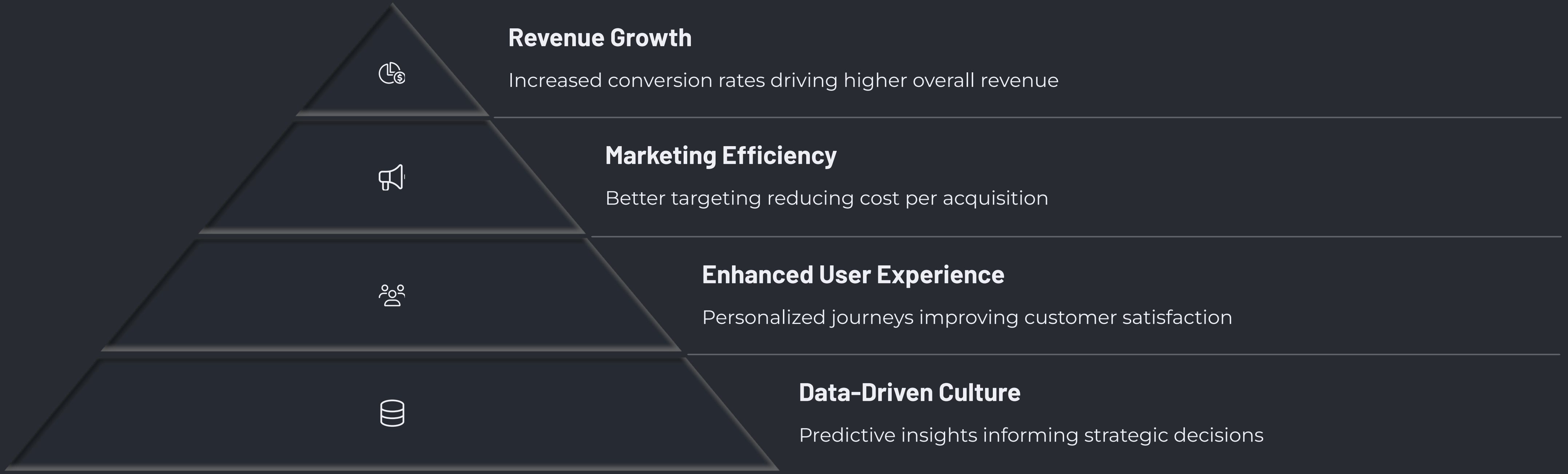
Phase 3: Full Integration (3-4 months)

Connect prediction system with marketing automation, personalization engines, and customer service platforms

Phase 4: Continuous Improvement (Ongoing)

Establish regular model retraining schedule and performance monitoring to maintain prediction accuracy

Expected Business Impact



Revenue Growth

Increased conversion rates driving higher overall revenue

Marketing Efficiency

Better targeting reducing cost per acquisition

Enhanced User Experience

Personalized journeys improving customer satisfaction

Data-Driven Culture

Predictive insights informing strategic decisions

By implementing the predictive model and associated recommendations, we anticipate a 15-20% improvement in conversion rates for targeted segments. This translates to significant revenue growth while simultaneously reducing marketing waste through more precise targeting.

The project also establishes a foundation for advanced personalization capabilities and a more data-driven approach to customer experience optimization across the organization.

Ready to Transform Your Customer Experience?

Thank you for your attention. Our data-driven predictive model is primed to revolutionize your conversion rates with notable improvement for targeted segments. By implementing our recommendations, you'll not only enhance revenue growth but also create more personalized customer journeys that drive loyalty and satisfaction. Ready to take the next step toward smarter, more efficient customer engagement?