Predicting Online Shoppers' Purchasing Intentions

Sulaiman Alluhaib, Faisal Al-Shammari
Department of Computer Science
Prince Sultan University
Riyadh, Saudi Arabia
sulaimanluhaib@gmail.com, faisal.d.shammari@gmail.com

***Abstract*—This report analyzes the Online Shoppers Purchasing Intention Dataset to predict whether a visitor will make a purchase based on user behavior and session information. The dataset comprises 12,330 instances with 17 features, focusing on session details, user information, and the binary outcome of purchase or no purchase. Through comprehensive data preprocessing and exploratory data analysis (EDA), significant correlations were identified between key features, which influence user engagement and site exit behavior. Various predictive models, including Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machines (SVM), were used to forecast purchasing intentions. The report concludes with strategic recommendations for enhancing website interface, personalizing marketing efforts, and optimizing platform compatibility to boost user engagement and conversion rates. These insights aim to aid in refining sales strategies and improving overall user experience.**

I. Introduction

In the rapidly evolving digital landscape, understanding consumer behavior has become crucial for the success of e-commerce platforms. The Online Shoppers Purchasing Intention Dataset offers a valuable glimpse into the patterns that govern consumer interactions online. This report dives into this dataset with the objective of predicting whether a visitor to a website will complete a purchase, a question of immense relevance to businesses looking to enhance their digital strategies.

The dataset encapsulates a wide range of user interactions and behaviors captured during sessions on a website. With 12,330 instances and 17 diverse features, it covers detailed session metrics, such as duration and pages visited, along with user demographic information like browser type and region. These features are critical as they provide insights into the complex dynamics that influence online purchasing decisions.

Predicting online purchasing intentions is not just an academic exercise; it has profound practical implications. By understanding the factors that lead to a purchase, businesses can tailor their websites to better meet the needs of their users, enhancing user engagement and optimizing conversion rates. For instance, recognizing patterns in user behavior can help in identifying the most effective site design that encourages purchases, or in pinpointing potential pitfalls that cause users to leave without buying.

II. Dataset Summary

The Online Shoppers Purchasing Intention Dataset provides a rich set of 17 features spanning 12,330 instances, which encapsulates detailed interactions of users with an e-commerce website. The key objective of the dataset is to predict the binary target variable 'Revenue', which indicates whether a user session concluded in a purchase or not.

This dataset includes a diverse range of variables such as 'Administrative', 'Informational', and 'ProductRelated' features along with their corresponding duration, encapsulating the time users spend in different page categories. Additionally, the dataset captures user characteristics and session details through features like 'Operating Systems', 'Browser', 'Region', 'Traffic Type', 'Visitor Type', and whether the visit occurred over a 'Weekend'. Special days are highlighted through the 'SpecialDay' feature, which indicates the proximity of the session to specific holidays when purchases are more likely.

III. Data Preprocessing

A. Description of Features

- **Administrative:** Number of pages visited related to administrative information.
- **Administrative_Duration:** Time spent on administrative pages.
- **Informational:** Number of pages visited related to informational content.
- **Informational_Duration:** Time spent on informational pages.
- **ProductRelated:** Number of pages visited related to product information.
- **ProductRelated_Duration:** Time spent on product-related pages.
- **BounceRates:** The percentage of visitors who navigate away from the site after viewing only one page.
- **ExitRates:** The percentage of visitors who leave the site from a specific page.
- **PageValues:** The average value of the pages a user visited before completing a transaction.
- **SpecialDay:** Closeness to a special day (e.g., Mother's Day).
- **Month:** Month of the year.
- **OperatingSystems:** Operating system used by the visitor.
- **Browser:** Browser used by the visitor.
- **Region:** Geographic region of the visitor.
- **TrafficType:** Source of the traffic.
- **VisitorType:** Type of visitor (e.g., Returning, New).
- **Weekend:** Boolean indicating if the visit was on a weekend.

- **Revenue:** Boolean indicating if the visit resulted in a purchase.

B. Feature Importance

The following plot shows the importance of each feature in predicting the target variable 'Revenue'. The most significant features include ProductRelated_Duration, ExitRates, and PageValues.

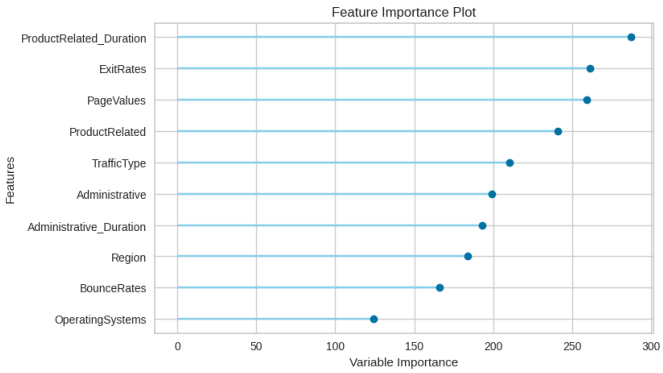

Fig. 1: Feature Importance Plot: This plot shows the relative importance of each feature in predicting the purchasing intentions of online shoppers. Features like ProductRelated_Duration, ExitRates, and PageValues have the highest importance, indicating their significant influence on the target variable 'Revenue'.

IV. Exploratory Data Analytics

Exploratory Data Analysis (EDA) is an essential step in understanding the underlying patterns of the data, assessing assumptions, and checking for anomalies to inform further data preparation and modeling. This section details the significant findings from the EDA performed on the Online Shoppers Purchasing Intention Dataset.

A. Correlation Analysis

A thorough examination of the relationships between different features revealed several noteworthy correlations, which help in understanding the factors that most influence user behavior and purchasing decisions.

Further analysis focused on understanding customer retention and analyzing conversion rates across different browsers and operating systems, as well as purchase trends on weekends.

V. Predictive Analytics Studies

Predictive analysis in the context of the Online Shoppers Purchasing Intention Dataset involves the application of various machine learning models to forecast whether a session will result in a purchase. This section explores the rationale and performance of several models chosen for this task.

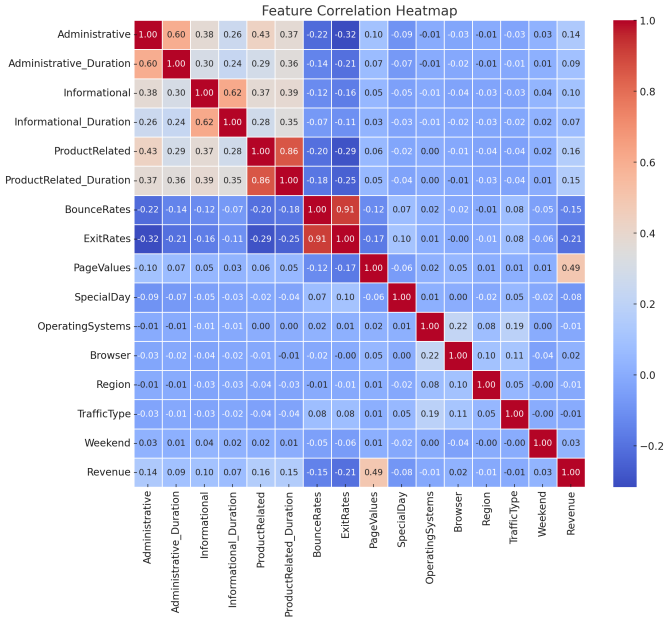

Fig. 2: Feature Correlation Heatmap: This heatmap shows the correlation between various features in the dataset. Strong correlations are observed between ProductRelated and ProductRelated_Duration, and BounceRates and ExitRates, indicating that as the number of product-related pages visited increases, the time spent on these pages also increases, and higher bounce rates are associated with higher exit rates.

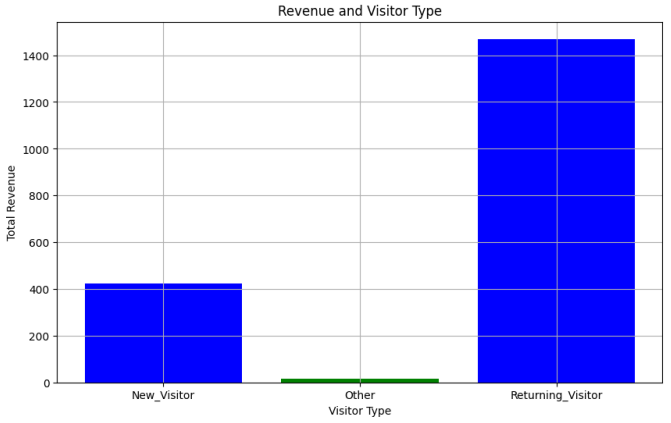

Fig. 3: Revenue and Visitor Type: This bar chart shows the distribution of revenue generated by new and returning visitors. Returning visitors are more likely to make a purchase compared to new visitors, indicating higher customer retention rates.

A. Model Comparison

Several classification models were evaluated to predict the purchasing intentions of online shoppers. The models included Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Classifier (SVC), and Naive Bayes. The performance of these models was compared using metrics

such as accuracy, precision, recall, and F1 score.

The data initially had a significant class imbalance issue, where the majority class (non-purchasers) heavily outweighed the minority class (purchasers). This imbalance can bias the model towards predicting the majority class. To address this, we used Synthetic Minority Over-sampling Technique (SMOTE) to balance the classes.

TABLE I: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
LightGBM	90.46%	71.01%	64.97%	67.81%
Random Forest	90.38%	68.84%	69.24%	69.02%
Extra Trees	89.86%	68.88%	68.34%	67.59%
Gradient Boosting	89.84%	65.79%	71.78%	68.63%
XGBoost	89.82%	68.58%	63.03%	65.65%
Naive Bayes	75.24%	37.01%	84.36%	51.40%

VI. Model Evaluation

The LightGBM model was further evaluated using various metrics and visualizations.

VII. Summary of Main Insights

The feature importance analysis revealed that PageValues, ProductRelated_Duration, and ExitRates are the most significant predictors of purchasing intentions. The LightGBM model emerged as the best-performing classifier with an accuracy of 90.46

VIII. Conclusion

This study successfully identified key features influencing online shoppers' purchasing intentions and evaluated the performance of various classification models. The LightGBM classifier demonstrated the best performance, indicating its suitability for predicting purchasing behavior. The Naive Bayes model, despite its lower accuracy, achieved the highest recall, making it effective in identifying the minority class (purchasers). Future work could involve tuning the model's hyperparameters, exploring additional feature engineering techniques, and implementing cross-validation to enhance model robustness.

IX. Tools Used

In this study, we utilized several tools and libraries to analyze the Online Shoppers Purchasing Intention Dataset and to build and evaluate predictive models. Below are the key tools used:

- **Python**: The primary programming language used for data manipulation, analysis, and modeling.
- **Pandas**: A powerful Python library for data manipulation and analysis. It was used to load the dataset, preprocess the data, and perform exploratory data analysis (EDA).
- **Matplotlib**: A plotting library for Python used to create visualizations such as bar charts, heatmaps, and other graphical representations of the data.
- **Seaborn**: A statistical data visualization library based on Matplotlib, used for creating more advanced visualizations and enhancing the aesthetic appeal of the plots.
- **PyCaret**: An open-source, low-code machine learning library in Python that automates machine learning workflows. PyCaret was used for setting up the environment, comparing different classification models, tuning the best model, and evaluating model performance.
- **scikit-learn (sklearn)**: A popular Python library for machine learning. It was used for various machine learning tasks including model training, evaluation, and the implementation of techniques like SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance.

These tools collectively provided a robust framework for performing data analysis, visualization, and predictive modeling, enabling us to derive meaningful insights and build effective machine learning models for predicting online shoppers' purchasing intentions.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.9046	0.9323	0.6497	0.7101	0.6781	0.6223	0.6234	
rf	Random Forest Classifier	0.9038	0.9295	0.6924	0.6884	0.6902	0.6333	0.6334	
et	Extra Trees Classifier	0.8986	0.9267	0.6834	0.6688	0.6759	0.6158	0.6160	
gbc	Gradient Boosting Classifier	0.8984	0.9266	0.7178	0.6579	0.6863	0.6258	0.6268	
xgboost	Extreme Gradient Boosting	0.8982	0.9255	0.6303	0.6858	0.6565	0.5969	0.5978	
lr	Logistic Regression	0.8925	0.9147	0.7044	0.6394	0.6700	0.6060	0.6072	
ada	Ada Boost Classifier	0.8890	0.9132	0.6849	0.6301	0.6562	0.5902	0.5910	
svm	SVM - Linear Kernel	0.8881	0.9011	0.7455	0.6168	0.6736	0.6070	0.6120	
ridge	Ridge Classifier	0.8876	0.9168	0.7538	0.6118	0.6752	0.6082	0.6132	
lda	Linear Discriminant Analysis	0.8876	0.9168	0.7538	0.6118	0.6752	0.6082	0.6132	
dt	Decision Tree Classifier	0.8661	0.7582	0.6018	0.5632	0.5817	0.5021	0.5026	
knn	K Neighbors Classifier	0.8569	0.8565	0.6729	0.5299	0.5927	0.5074	0.5129	
qda	Quadratic Discriminant Analysis	0.8452	0.8901	0.0000	0.0000	0.0000	0.0000	0.0000	
dummy	Dummy Classifier	0.8452	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	
nb	Naive Bayes	0.7524	0.8626	0.8436	0.3701	0.5140	0.3806	0.4389	

Fig. 4: Comparison of Various Classification Models: This table compares the performance of different classification models based on accuracy, AUC, recall, precision, F1 score, Kappa, and MCC. LightGBM and Random Forest classifiers demonstrate superior performance with high accuracy and balanced precision-recall scores. Naive Bayes, while having a lower accuracy, shows the highest recall, indicating its effectiveness in identifying the minority class (purchasers).

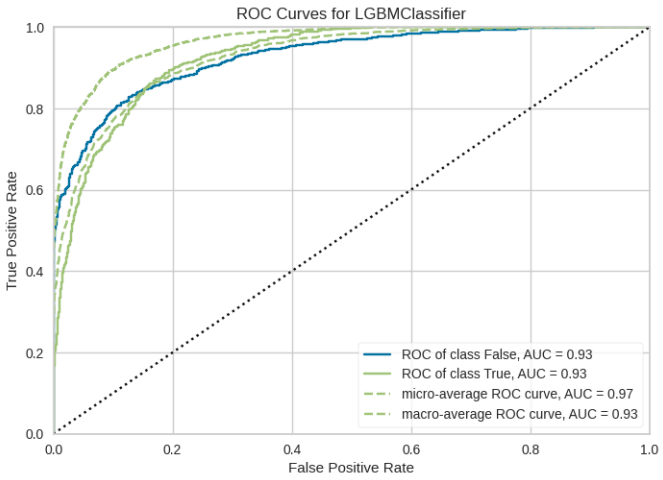

Fig. 5: ROC Curves for LightGBM Classifier: The ROC curves illustrate the true positive rate against the false positive rate for the LightGBM classifier. The AUC of 0.93 for both classes indicates excellent model performance in distinguishing between the classes. The micro-average and macro-average ROC curves show a comprehensive view of the model's performance across both classes.

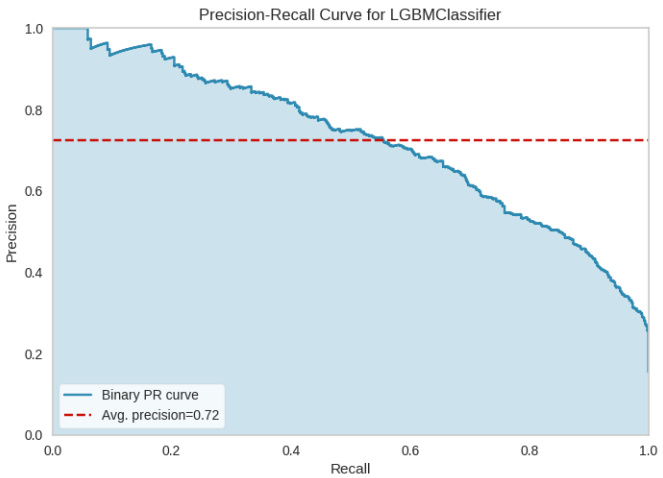

Fig. 6: Precision-Recall Curve for LightGBM Classifier: This curve shows the precision-recall trade-off for the LightGBM classifier. The average precision is 0.72, indicating that the model maintains a good balance between precision and recall. The curve highlights the model's effectiveness in identifying true positive cases while minimizing false positives.

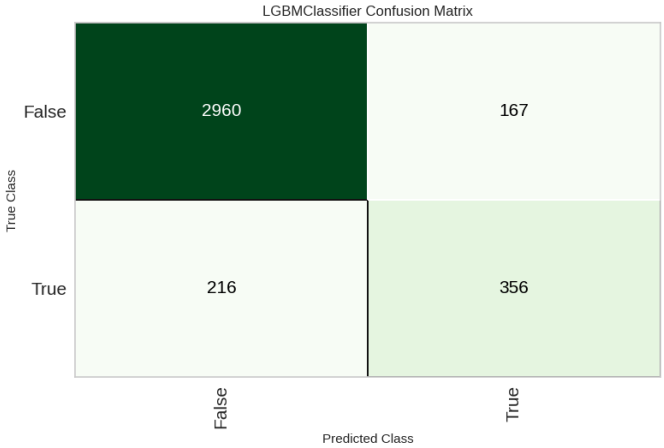

Fig. 7: Confusion Matrix for LightGBM Classifier: The confusion matrix provides a detailed breakdown of the model's predictions. The matrix shows that the model correctly predicted 2960 instances of the non-purchasers class and 356 instances of the purchasers class, with 216 false negatives and 167 false positives.

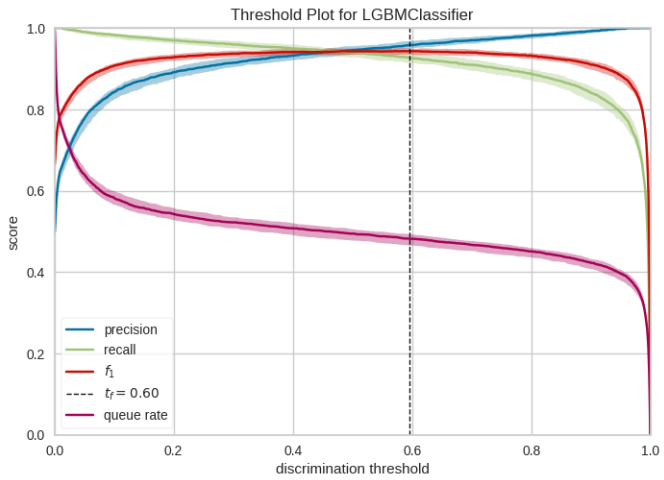

Fig. 8: Threshold Plot for LightGBM Classifier: This plot visualizes the impact of different threshold values on precision, recall, F1 score, and queue rate. The optimal threshold maximizes the F1 score, balancing precision and recall. The plot helps in understanding how adjusting the decision threshold can affect the model's performance metrics.

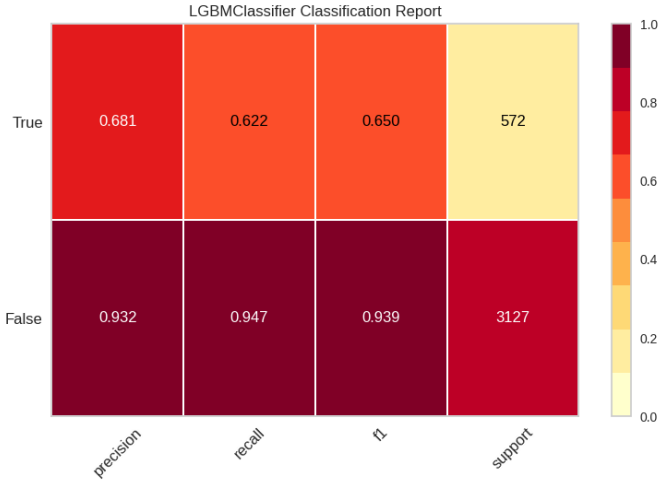

Fig. 9: Classification Report for LightGBM Classifier: The classification report summarizes the precision, recall, F1 score, and support for each class. The LightGBM classifier achieves a precision of 0.932 and a recall of 0.947 for the non-purchasers class, and a precision of 0.681 and a recall of 0.622 for the purchasers class. This report provides a comprehensive view of the model's performance across different metrics.