PROJECT PRESENTATION

PREDICTING ONLINE SHOPPERS' PURCHASING INTENTIONS

BUSINESS PROBLEM

- E-commerce sites face challenges in predicting customer purchase intentions. Optimizing conversion rates require accurate prediction of customer behaviour. The goal is to enhance user experience while also increasing revenue.
- **Low Conversion Rates** : Many online shopping sessions do not result in purchases. This results in inefficient resource utilization and missed revenue opportunities.
- **Customer Abandonment** : Customers often leave without completing transactions, leading to shopping cart abandonment.
- **Resource Allocation** : Marketing campaigns and promotional resources are often directed to users with low likelihood of purchase, leading to wasted effort and budget.

MOTIVATION

The objective is to develop and identify the most effective predictive models for detecting purchase intentions, empowering businesses to refine their strategies for enhanced customer engagement and increased sales. The anticipated outcomes are as follows:

- **Revenue Growth**: Accurate prediction of purchase intent allows businesses to target potential buyers effectively, improving conversion rates and maximizing revenue.
- **Enhanced Customer Experience**: Personalized interventions can reduce abandonment by addressing user hesitations in real time.
- **Resource Optimization**: By minimizing wasted resources on non-converting customers, businesses can redirect their efforts to users with higher conversion potential.
- **Competitor Edge**: Advanced analytics can provide an edge in a competitive e-commerce landscape by enabling smarter decision-making and customer engagement strategies.

PREVIOUS WORK ON THE DATASET

- Prior work by **Sakar et al., 2018** used **MLP** and **LSTM** for real-time prediction of online shoppers' purchasing intentions.
- **Objective :**
- Developed a two-module system for real-time e-commerce analytics:
  - Purchasing Intention Prediction
  - Abandonment Likelihood Prediction
- **Approaches :**
- **Module 1** : Machine Learning Models (MLP, SVM, Random Forest) for purchase prediction.
  - **Best Model: Multilayer Perceptron (MLP)**
    - Achieved 87.24% accuracy and 0.86 F1-score after addressing class imbalance with oversampling.
  - Features like **Page Value**, **Exit Rate**, and **Bounce Rate** were identified as the most critical predictors.
- **Module 2** : Deep Learning (LSTM-RNN) for sequential clickstream analysis.
  - Accurately predicted abandonment likelihood based on navigation patterns.
  - Outperformed traditional Hidden Markov Models (HMM) in scalability and prediction accuracy.

PREVIOUS WORK ON THE DATASET

Innovations

- Introduced a dual-module system for simultaneous purchase intention and abandonment prediction.
- LSTM-RNN for clickstream analysis demonstrated superior performance in real-time applications.
- Feature selection using the mRMR method optimized model performance and reduced computational overhead.

Impact

- Enabled personalized content delivery to users with purchase intent but high abandonment likelihood.
- Increased e-commerce conversion rates by improving real-time decision-making.
- Provided a benchmark for using MLP, achieving higher accuracy than traditional classifiers like SVM and Random Forest.
- Highlighted the importance of combining clickstream data with session information for better predictions.

OPPORTUNITIES IN USING BUSINESS ANALYTICS FOR E-COMMERCE

The power of business analytics lies in its ability to turn complexity into clarity. By enabling smarter decisions, driving efficiency, and delivering personalized experiences, analytics becomes an indispensable tool for modern businesses to thrive in a competitive and ever-changing environment.

- **Tailoring Marketing Efforts** : Business analytics allows segmentation of users based on factors like browsing behaviour, session timing, or traffic source. Targeted promotions and advertisements can be designed to resonate with specific user groups, enhancing marketing ROI.
- **Strategic Decision-Making** : Data from e-commerce sessions provides insights into customer preferences, product performance, and traffic sources. These insights guide decisions on inventory management, product placement, and resource allocation, contributing to long-term growth.
- **Reducing Cart Abandonment** : Identifying users likely to abandon their sessions enables real-time interventions, such as offering discounts or personalized recommendations to retain potential buyers.
- **Competitor Edge** : Advanced analytics can provide an edge in a competitive e-commerce landscape by enabling smarter decision-making and customer engagement strategies.

APPROACH TO ADDRESSING THE PROBLEM

We followed a systematic approach of data cleaning, feature engineering, and class balancing to prepare the dataset for predictive modelling. A mix of interpretable (Logistic Regression, Decision Tree) and advanced models (MLP, XGBoost) was applied, with **XGBoost** demonstrating the best performance. The insights from these models, such as feature importance and performance metrics, drive actionable strategies to address the business problem effectively.

- **Data Used** : **UCI Online Shoppers Intention Dataset.**
- **Key Characteristics**:
  - Size: 12,330 sessions, each associated with a unique user, to ensure unbiased results across campaigns, seasons, or user profiles.
  - **Features**:
    - **Numerical features** like Bounce Rate, Exit Rate, Page Value, Duration on Product Pages.
    - **Categorical features** like Operating System, Browser, Region, Visitor Type, and Traffic Source.
    - **Target Variable**: Revenue (Binary - Purchase or No Purchase).

APPROACH TO ADDRESSING THE PROBLEM

- **PREPROCESSING STEPS**
- **Handling Missing Data** : Checked for null values; dataset was cleaned accordingly.
- **Encoding** : One-hot encoding was applied to categorical variables for compatibility with machine learning models.
- **Balancing the Target Variable** :
  - The dataset exhibited class imbalance (84.5% sessions did not lead to purchases).
  - Techniques like Stratified Split and SMOTE were applied to balance the classes effectively.
- **Scaling** : Numerical features were standardized using StandardScaler to ensure consistency across models.

APPROACH TO ADDRESSING THE PROBLEM

- **Models Chosen to Apply and Test**

1. **Logistic Regression**: A simple and interpretable baseline model, Used to establish benchmark performance and evaluate linear relationships.
2. **Decision Tree**: Captures non-linear relationships and provides interpretable decision rules. Allows visualization of how features impact predictions.
3. **Random Forest**: An ensemble method that improves robustness and accuracy by combining multiple decision trees. Provides feature importance scores, aiding in insights about influential predictors.
4. **Support Vector Machine (SVM):** Applied to create complex decision boundaries, especially useful for smaller feature sets. Radial Basis Function (RBF) kernel was tested to capture non-linear patterns.
5. **Multilayer Perceptron (MLP):** A feedforward neural network to capture complex relationships in the data. Tuned hyperparameters like hidden layers, learning rates, and activation functions.
6. **XGBoost**: Gradient boosting method known for handling structured data effectively. Used to enhance accuracy by capturing subtle feature interactions.

RESULTS

- Our analysis revealed that the **XGBoost model without SMOTE** emerged as the most effective for predicting online shoppers' purchasing intentions. This model achieved the highest F1-score of 0.64, balancing precision and recall effectively. It accurately identified potential buyers while minimizing false positives, ensuring that marketing efforts are targeted and resources are allocated efficiently.
- **Performance Metrics :**

- **F1-Score** for Purchasing Class (Revenue = TRUE): **0.64** - Indicates a strong balance between precision and recall, ensuring the model effectively identifies potential buyers while minimizing false positives.
- **Precision 0.71** : Reflects the model's ability to accurately predict actual buyers without over-targeting non-buyers.
- **Recall: 0.58 -** Demonstrates its effectiveness in capturing a significant portion of potential buyers, reducing the risk of missed opportunities.
- **Accuracy: 0.85 -** Highlights the model's overall robustness in classifying both purchasing and non-purchasing sessions.

- **Feature Insights** : The XGBoost model identified Page Value, Bounce Rate, and Exit Rate as the most critical predictors. These insights are actionable, allowing businesses to optimize website content and marketing strategies.

RESULTS

- Why **XGBoost Without SMOTE** is the Best Model

1. **Balanced Performance** : Unlike models with SMOTE, XGBoost without oversampling achieved a superior balance between precision and recall. This balance is crucial for e-commerce businesses where false positives (wasting resources on non-buyers) and false negatives (missing potential buyers) have significant impacts.
2. **Robustness and Efficiency** : XGBoost effectively handled the dataset's imbalanced nature without introducing noise from synthetic data, maintaining high predictive accuracy and reliability. It avoids overfitting to oversampled data, ensuring better generalization to unseen user sessions.
3. **Actionable Insights** : The model provided clear insights into feature importance, enabling targeted interventions and website optimizations that align with user behaviour.
4. **Business Alignment** : The model's strong precision ensures that marketing resources are utilized efficiently, reducing unnecessary spend on unlikely buyers. Its relatively high recall ensures that potential buyers are captured, maximizing revenue opportunities.

- In summary, XGBoost without SMOTE excelled as the best model due to its ability to balance precision and recall effectively, provide actionable insights, and align with business objectives for efficient resource allocation and revenue growth.

SUMMARY OF ALL MODELS

Models		Precision	Recall	F1-Score	Accuracy
Logistic Regression	FALSE	0.95	0.89	0.92	0.87
	TRUE	0.56	0.73	0.63	
Logistic Regression with Smote	FALSE	0.95	0.89	0.92	0.86
	TRUE	0.54	0.74	0.63	
Random Forest	FALSE	0.92	0.96	0.94	0.9
	TRUE	0.74	0.55	0.63	
Random Forest with SMOTE	FALSE	0.94	0.92	0.93	0.89
	TRUE	0.62	0.7	0.66	
Decision Tree	FALSE	0.94	0.9	0.92	0.87
	TRUE	0.55	0.69	0.61	
Decision Tree with SMOTE	FALSE	0.94	0.9	0.92	0.87
	TRUE	0.55	0.7	0.62	
Decision Tree with ADASYN	FALSE	0.94	0.9	0.92	0.87
	TRUE	0.55	0.7	0.62	
MLP	FALSE	0.92	0.95	0.94	0.89
	TRUE	0.69	0.56	0.62	
MLP with SMOTE	FALSE	0.94	0.89	0.92	0.86
	TRUE	0.55	0.71	0.62	
SVM	FALSE	0.95	0.9	0.92	0.87
	TRUE	0.57	0.74	0.64	
SVM with SMOTE	FALSE	0.87	0.96	0.91	0.84
	TRUE	0.45	0.19	0.26	
XGBoost	FALSE	0.93	0.96	0.94	0.9
	TRUE	0.71	0.58	0.64	
XGB with SMOTE	FALSE	0.94	0.93	0.93	0.89
	TRUE	0.63	0.69	0.66	

CONCLUSION

- Our project demonstrates the significant value machine learning models bring to predicting whether online shoppers are likely to complete a purchase. By employing careful data balancing techniques and thoughtful feature engineering, we enhanced the models' ability to capture meaningful patterns and generate actionable insights. Among all the tested models, the **XGBoost model without SMOTE** emerged as the top performer, striking an effective balance between precision and recall. This balance ensures that potential buyers are accurately identified without an excessive number of false positives, making it an ideal choice for e-commerce businesses looking to optimize their marketing efforts.
- The XGBoost model's reliability underscores its ability to capture valuable opportunities while minimizing wasted resources, leading to more efficient marketing campaigns. Additionally, the results align closely with findings in previous research, which highlighted the effectiveness of machine learning techniques, including neural networks and ensemble methods, for addressing imbalanced datasets. This alignment reinforces the potential of these models for personalized marketing and customer engagement. By integrating machine learning, businesses can improve customer targeting, reduce unnecessary advertising expenses, and drive revenue growth, establishing a competitive edge in the dynamic e-commerce landscape.

RECOMMENDATIONS

- **Targeted Marketing Campaigns** :

Leveraging the model’s ability to accurately identify potential buyers, businesses can develop targeted marketing strategies focused on high-probability customers. By concentrating promotional efforts and incentives on users predicted to make a purchase, businesses can enhance conversion rates while minimizing unnecessary advertising costs on users less likely to buy.

- **Personalized Customer Experience** :

Key features highlighted by the model, such as Product Related Duration and Page Value, offer opportunities to create a more tailored shopping experience. By utilizing these insights, businesses can customize product recommendations, optimize website layouts, and design real-time promotions, fostering an engaging shopping environment that boosts sales and improves customer satisfaction.

- **Optimize Website Design Based on Insights**:

Key predictors such as Bounce Rate and Exit Rate suggest areas where website usability and navigation can be improved to reduce session abandonment and guide users toward completing purchases.

THANK YOU