Report: Design Decisions and Methodology

1. Exploratory Data Analysis (EDA)

The first step involved performing EDA to understand the dataset structure, distributions, and relationships between variables. Since the dataset contains numerical and categorical data, we analyzed the distribution of the target variable, examined potential correlations among features, and explored the relationships between key variables, such as BounceRates and ExitRates. Visualizations, including histograms, pair plots, and a correlation heatmap, provided a clearer understanding of the data's trends and patterns. This EDA phase allowed us to identify important features and decide on preprocessing steps required to handle categorical data effectively.

2. Data Preprocessing

Categorical variables were encoded using one-hot encoding, as this method ensures all data is numeric and ready for model input. We also applied feature scaling through StandardScaler to standardize the numerical values, which is essential for models sensitive to feature magnitudes, such as Logistic Regression and SVM.

3. Algorithm Selection

Three algorithms were selected for comparison: **Logistic Regression**, **Random Forest**, and **Support Vector Machine (SVM)**.

- **Logistic Regression** was chosen as a simple baseline due to its effectiveness in binary classification and interpretability.
- **Random Forest** was selected for its robustness in handling a mix of feature types and for its ability to provide feature importance, which helped us understand which variables contribute most to the target prediction. *Based on its performance, Random Forest was identified as the best-suited model for this problem, achieving a high accuracy score that outperformed other models.*
- **SVM** was included as it performs well in high-dimensional spaces and provides flexibility through kernel functions. Each algorithm was evaluated on accuracy, confusion matrix, and ROC AUC to ensure comprehensive performance insights.

4. Evaluation and Comparison

Model performance was assessed using metrics like accuracy, classification report, and ROC AUC score. The ROC curve analysis was particularly valuable in identifying the best-performing model across all classification thresholds. With an accuracy of 92%, Random Forest demonstrated superior predictive capabilities on this dataset. *While this accuracy is satisfactory, potential improvements include hyperparameter tuning, which could further boost performance.*

Conclusion

Random Forest emerged as the most suitable model for this problem, outperforming other algorithms in terms of accuracy. Its ability to rank feature importance offered insights into which variables most impact the target prediction, aiding in focused optimization efforts.

Business Impact of Results

Implementing a predictive model for customer behavior based on browsing patterns has significant business implications, particularly for enhancing user experience and optimizing marketing strategies. By accurately predicting the Target variable (likely representing a binary outcome like a conversion or customer engagement), businesses can personalize interactions and improve customer satisfaction. For instance, identifying high bounce rates in combination with other indicators allows a business to target specific users with tailored content or offers, potentially increasing conversion rates and reducing abandonment.

Additionally, understanding feature importance through models like Random Forest helps businesses prioritize their marketing and website optimization efforts. Insights into which attributes (e.g., ProductRelated_Duration or PageValues) most strongly influence customer behavior enable data-driven decisions about resource allocation for maximum impact. For instance, if ExitRates significantly correlate with lower conversions, efforts can be focused on optimizing these areas, such as by enhancing product page engagement or improving navigability.

In summary, using predictive analytics in customer behavior enables businesses to improve engagement, optimize conversion rates, and tailor marketing strategies effectively. These targeted interventions not only increase revenue potential but also foster long-term customer loyalty through personalized and responsive user experiences.