Understanding Customer Interactions: Predicting Revenue Generation

Executive Summary

Objective and Scope

This project was designed to explore the dynamics of customer interaction and their impact on revenue generation for a specific online shopping website through a featured dataset. Due to constraints in accessing detailed individual customer information, we constructed customer profiles relying on relevant columns within the datasets. Subsequently, we identified and employed a predictive classification model with optimal performance to forecast the whether a given web browsing session translating into revenue. The insights gleaned from this analysis offer valuable implications for tailoring future campaigns to distinct audience segments, thereby maximizing profitability.

Key Findings and Highlights

The dataset predominantly comprises returning customers, with Region 1 exhibiting the highest customer volume. Browsers 2 and 1 are the primary means of accessing the website, presumably suggesting a prevalent trend of customers favoring mobile devices for web interactions. Despite early promotions, customers seem inclined to defer purchases to special days or holidays rather than planning ahead. In the realm of predictive modeling, the Random Forest algorithm emerged as the optimal choice for forecasting revenue generation in this dataset, demonstrating the highest area under the Precision-Recall curve. The dataset's inherent imbalance, marked by over 80% of sessions not generating revenue, notably contributes to the model's heightened accuracy in predicting non-revenue-generating sessions.

Conclusion

In conclusion, the revenue forecasting model stands as a pivotal asset for strategic decision-making in e-commerce. Beyond its primary function of predicting revenue generation, the model has the potential to aid inventory planning and significantly enhance marketing strategies by offering precise insights during peak revenue periods. Its impact extends to resource allocation, guiding e-commerce businesses to efficiently allocate resources where customer interactions are most likely to yield revenue.

Action Plan

To leverage the revenue forecasting model effectively, prioritize implementing predictive inventory planning and refining marketing strategies during peak revenue periods. Allocate resources strategically based on forecasted high-revenue interactions and regularly update allocations to adapt to changing trends. By systematically adopting these actions, businesses can fully harness the model's potential, fostering adaptive and sustained growth in the dynamic e-commerce landscape.

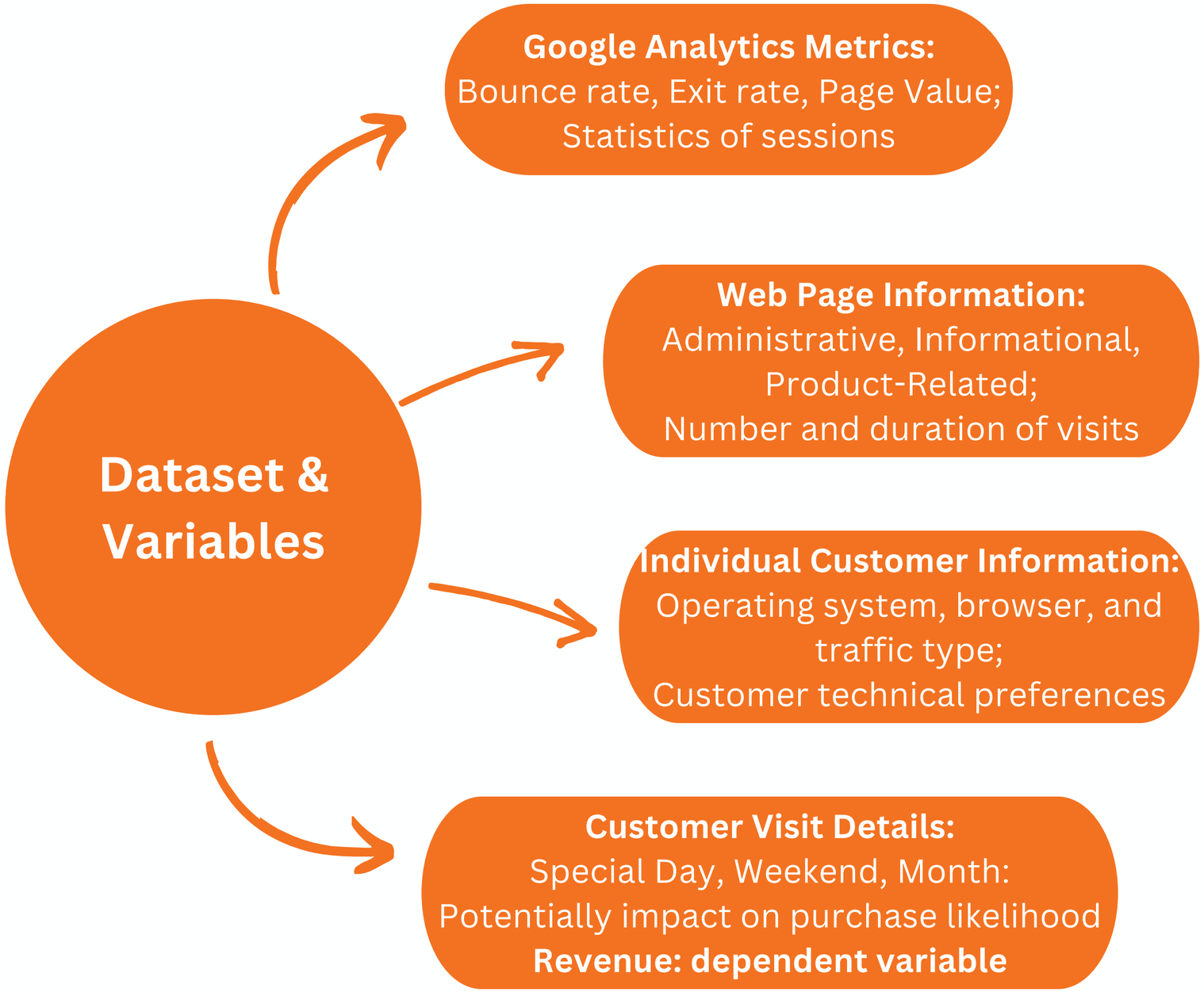
Problem Description

Business Context
Growing importance of interpreting online customer interactions in e-commerce.
Impact on marketing strategies, website design, and sales enhancement.

Project Aim
To harness data on customer interactions, visit timings, and types.
To develop supervised machine learning models for predicting session revenue outcomes.
Aiming to offer insights into customer behavior to refine online marketing strategies.

Dataset & Variables

10 numeric and 8 categorical divided into 4 categories

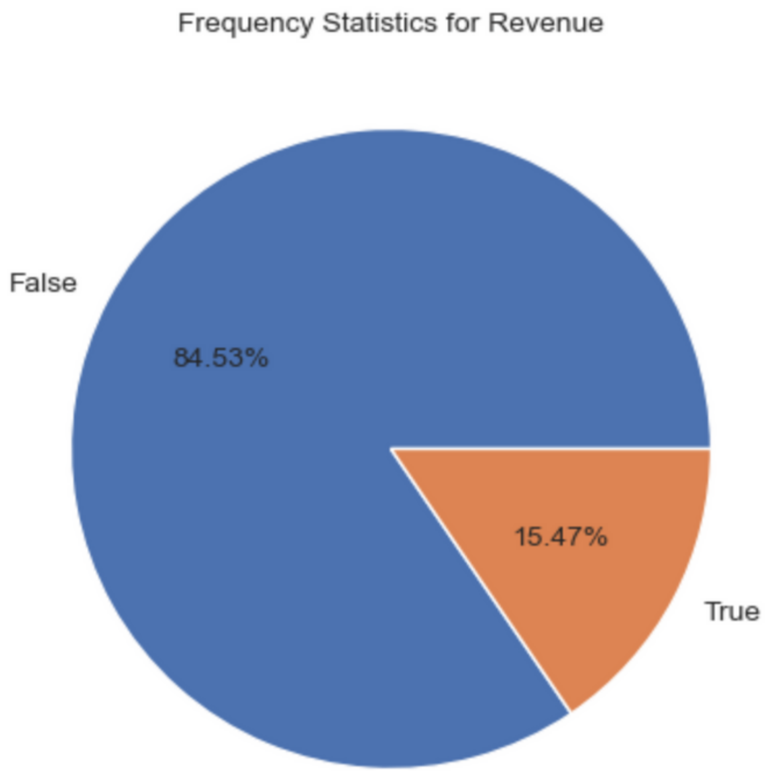


Data Exploratory Analysis

As part of our analysis, we conducted exploratory research, which involved examining summary statistics, visualizing data, creating frequency tables, and performing correlation analyses (Appendix). Here are some key insights we gained:

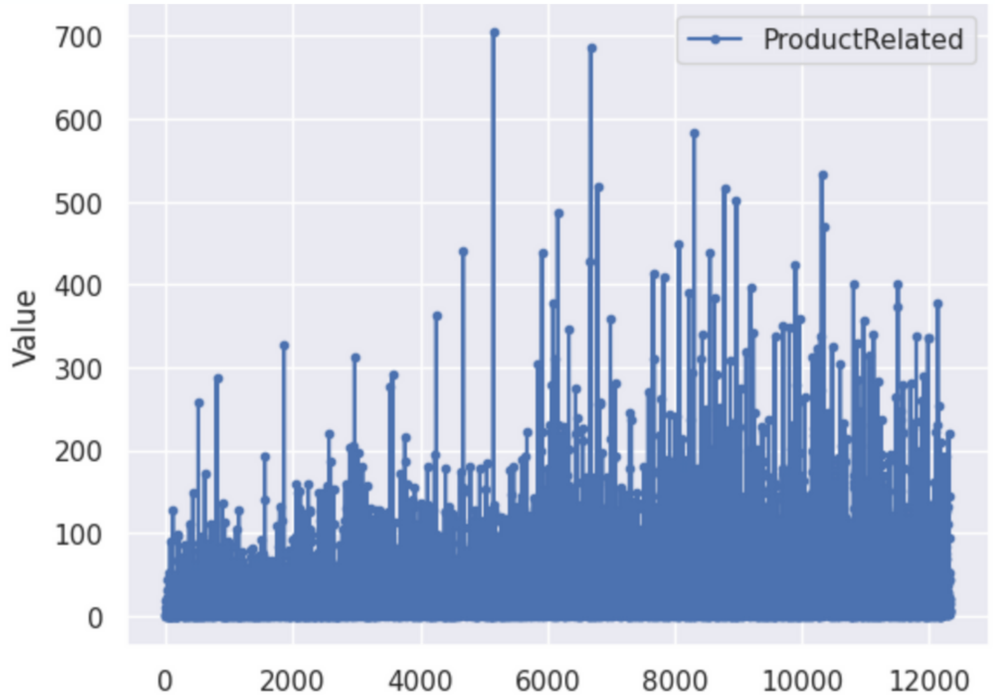
Data Exploratory Analysis

Revenue as an Indicator: The presence of a 'Revenue' boolean variable in the dataset suggests that understanding which factors contribute to successful transactions is a key focus.



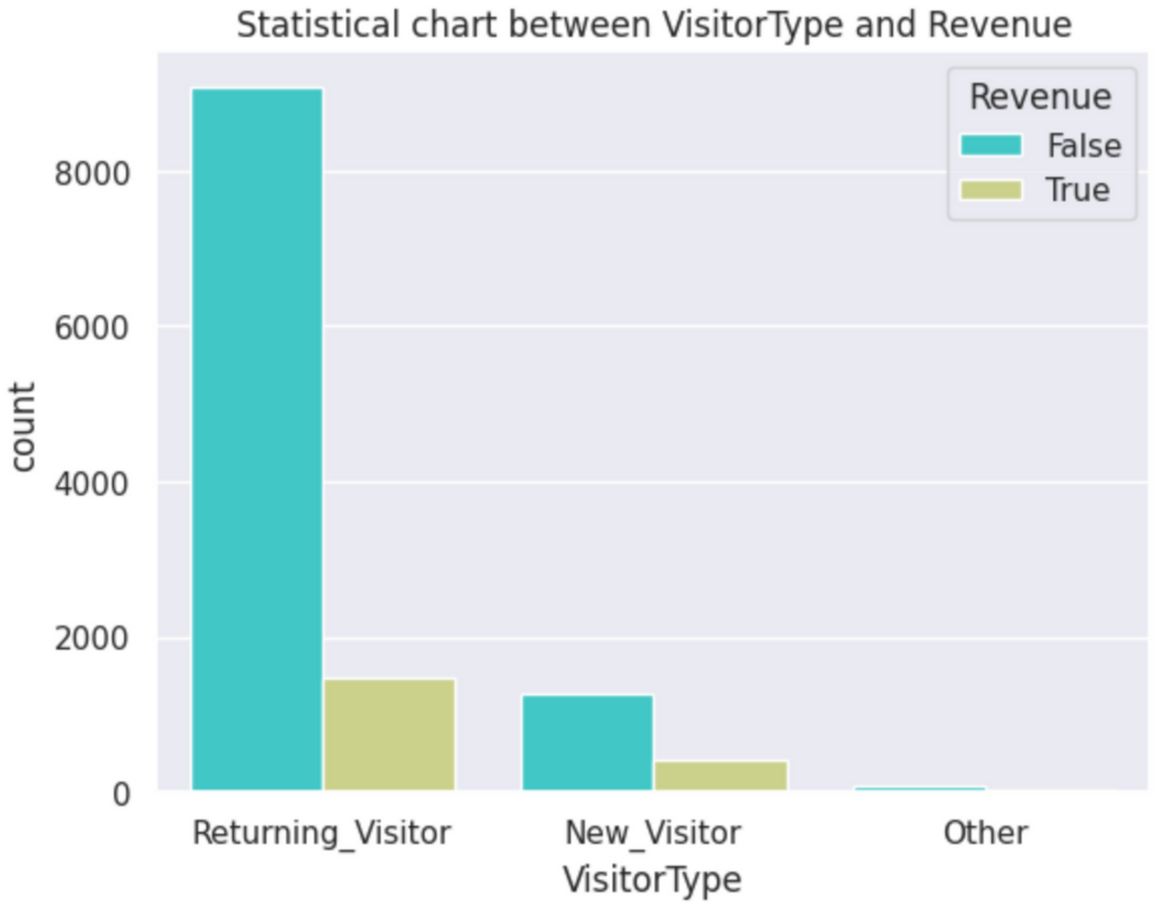
Data Exploratory Analysis

Product Pages are Key: There's significant engagement with product-related pages, mostly concentrated around 200-300 types of pages, highlighting interest in specific products or categories.



Data Exploratory Analysis

Predominance of Returning Visitors: The majority of visitors are returning, indicating strong customer retention or repeated interest.



Data Exploratory Analysis

Diverse Device and Browser Usage: Data shows a variety of operating systems and browsers used to access the site, reflecting diverse technology preferences.

Browser	
1	20.965126
2	53.536091
3	20.721817
4	3.876723
5	0.048662
6	0.154096
7	0.056772
8	0.640714

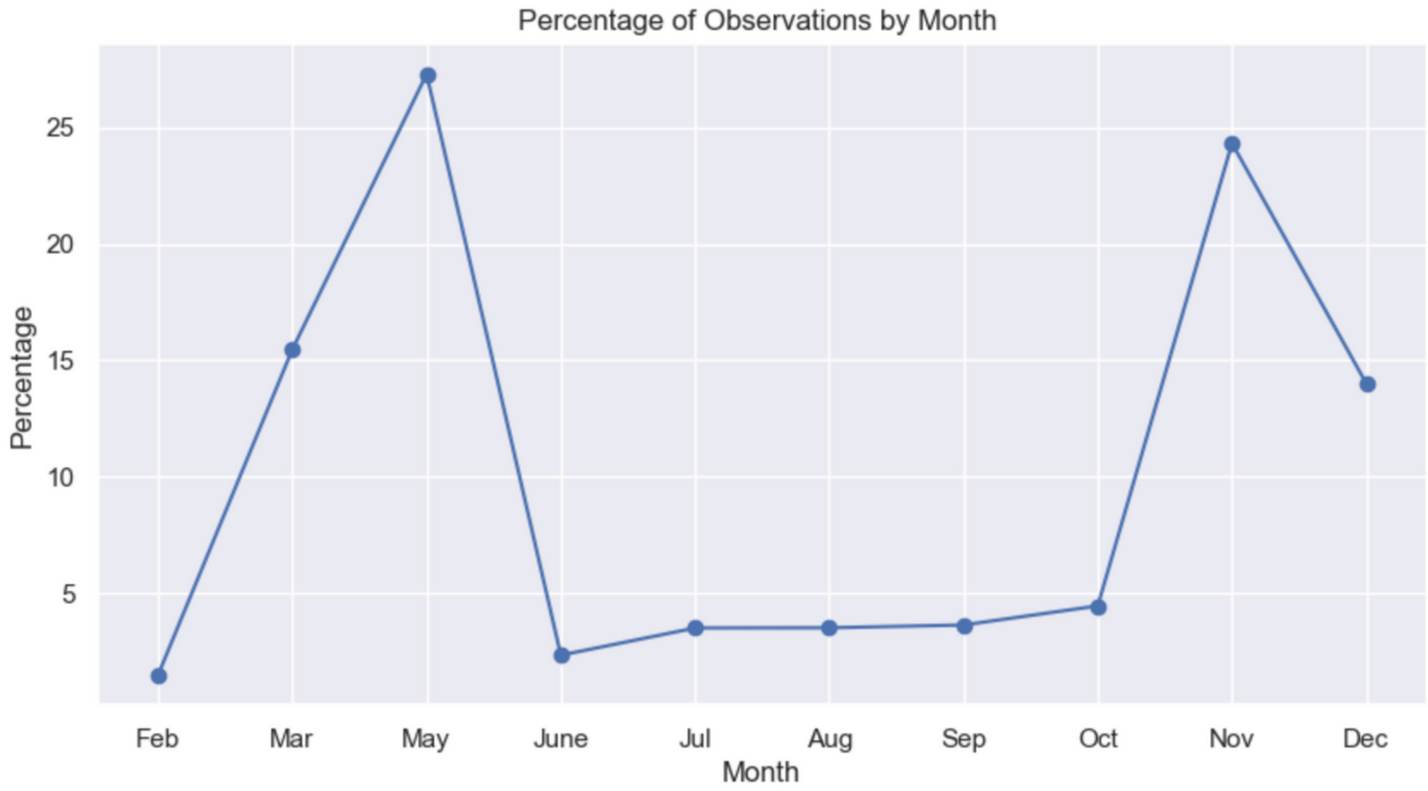
col_0 % observations

OperatingSystems	
1	20.965126
2	53.536091
3	20.721817
4	3.876723
5	0.048662
6	0.154096
7	0.056772
8	0.640714

col_0 % observations

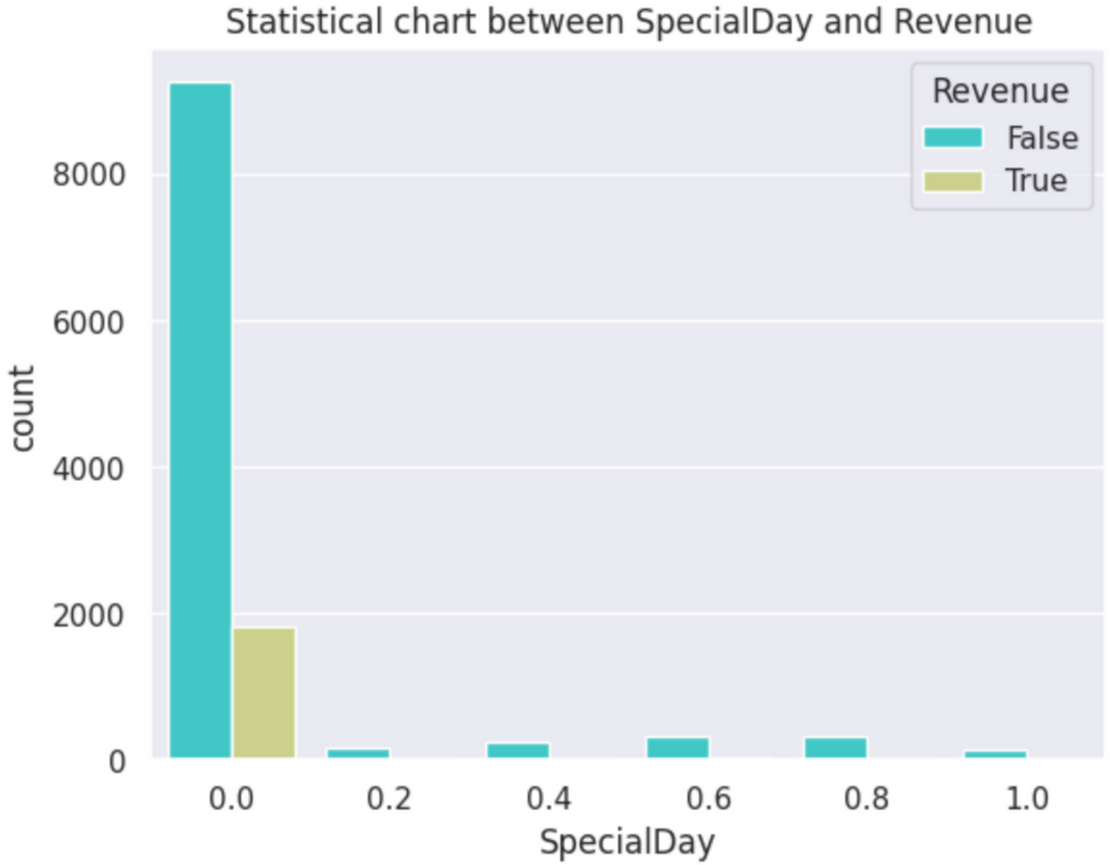
Data Exploratory Analysis

Monthly Traffic Variations: Certain months (like May, November, and March) show notably higher traffic, possibly linked to seasonal trends



Data Exploratory Analysis

Impact of Special Days: The 'Special Day' feature indicates how close session occurrences are to holidays in a scale of 0 to 1, crucial for understanding buying behavior during specific periods.



Data Exploratory Analysis

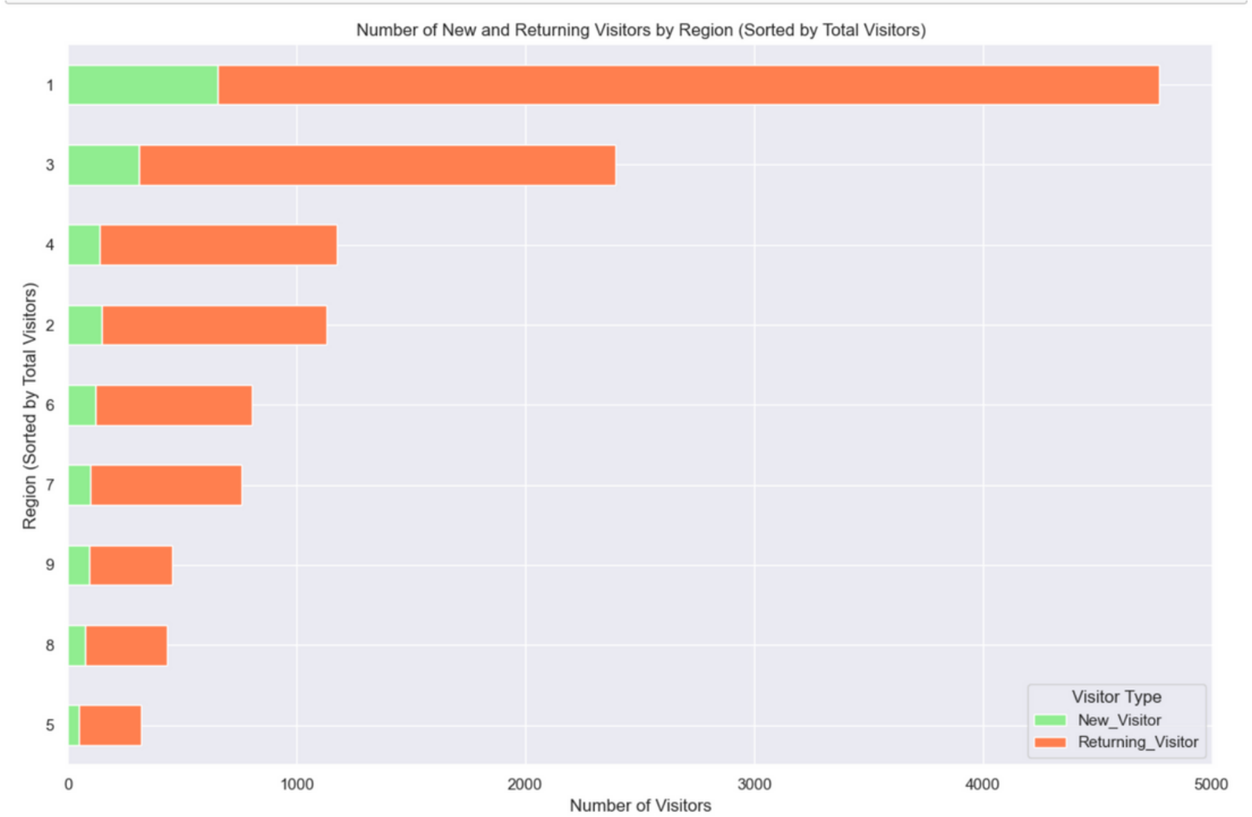
- **Varied User Engagement:** Visitors exhibit varying levels of engagement in administrative, informational, and product-related activities, indicating diverse visiting intents and depth.
- **Room for User Experience Improvement:** The range of bounce and exit rates suggests that while some pages are highly engaging, others may lead to quick user exits, presenting potential areas for user experience enhancement.

Data Exploratory Analysis

Region: The dataset does not contain definition information for 'region'; it simply uses the numbers 1-9 to represent them. In our data exploration, we attempted to identify the characteristics of each region.

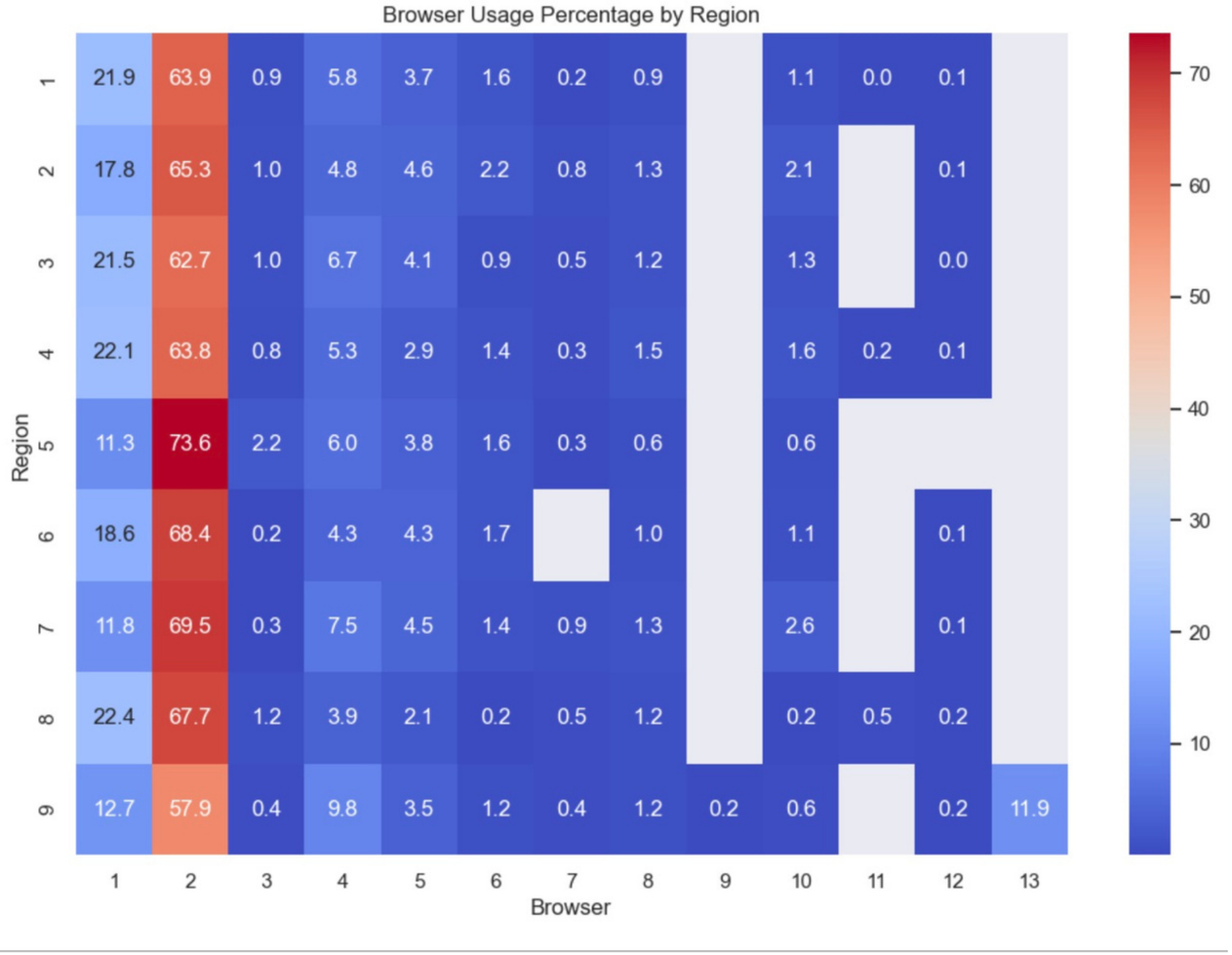
Data Exploratory Analysis

Regions: This graph presents a clear visualization of new versus returning visitors by region. We observe significant disparities in the number of shoppers across different regions. For example, Region 1 has a much higher number of shoppers compared to others



Data Exploratory Analysis

Region vs Browser :
According to this heat, as for all regions, Browser 2 has the highest number of users, followed by Browser 1. We speculate that Browser 2 represents mobile access while Browser 1 represents computer access.



Data Preprocessing

In order to prepare the data for analysis, we have formulated a comprehensive pre-processing plan. Here are the main steps we had completed:

- **Data Type Conversion:** The data types of certain columns (such as “OperatingSystems”, “Browser”, “Region”, “TrafficType”) converted from numerical to categorical (object type).

- **Duplicate Removal:** Duplicate rows in the dataset identified and removed.

- **Missing Value Handling:** Checked for missing values using isnull().sum().

Data Preprocessing - Feature Selection

Total number of features: 62

LASSO

Only 2 features that have coefficients greater than 0.

We first tried out LASSO feature selection but it turns out that only 2 features were being selected, so considering losing less information about the dataset, we are not going with this feature selection method.

Z-Score

Set alpha = 0.05, only 20 features that have a z-score smaller than alpha.

The second feature selection method we tried is Z-score (or linear regression). We've also tried some smaller Z-scores, but meanwhile, the number of the selected features also decreases.

Mutual Information Score

Order the MI score of the features from largest to smallest, and select the top 30 features that have relatively higher MI score.

The last method we tried was the mutual information score. Compared with Z-score, this method is better for exploring the importance of the mix of categorical variables and continous variables. And so, we would like to use MI score for feature selection.

Data Preprocessing - Feature Selection

By the implications of our feature selection method on the previous slide, we would like to see the effect of feature selection by fitting **two logistic regression models**, one with feature selection and one without feature selection. We chose logistic regression to do this comparison because logistic regression takes relatively less time to train the model.

By comparing the **performance of two logistic regression models**, one with feature selection and one without feature selection, the **area under the precision-recall curve** of the model with feature selection turns out to be very low (around 0.22), we thus decided **not to perform any feature selection and dimension deduction**.

Model Selection

Evaluation Metrics

Area under Precision-Recall curve (PR AUC)
Importance due to class imbalance.

Hyperparameter Tuning

Utilization of Grid Search or Random Search
Aim for optimization of model performance.

Model Evaluation & Selection

In-depth Evaluation: Area under Precision-Recall curve (PR AUC).
Select model(s) based on test set efficiency

Feature Importance & Insights

Analyze drivers of shopping behavior.
Gain insights for strategic decisions.

Model Selection

	KNN	Logistic Regression	Gradient Boosting	Random Forest	MLP (0/1/2/3 Hidden Layer)
Area Under Precision-Recall Curve	0.58	0.61	0.73	0.73	0.63
Hyper - parameters of the best model for each algorithm	n_neighbors = 200 p = manhatten	max_iter = 200 tol = 1e-06	learning_rate= 0.02 max_depth = 7 max_features =30 min_samples_leaf = 10 n_estimators = 180	n_estimators = 250 max_features = 13 min_sample_leaf = 5 criterion = entropy	shuffle = True max_iter = 600 batch_size = 256 activation = ReLU alpha = 0.01 hidden_layer_sizes = (512,) learning_rate_init = 0.01 momentum = 0.1

Strategy for Hyperparameter Tuning

The strategy of tuning KNN, Logistic regression, Gradient Boosting, and Random Forest shares the same logic. We first did a random trial, testing through a random set of hyperparameters. Then depending on the results, we narrowed the range for each hyperparameter. We stopped tuning the model when there was no significant change in the PR AUC or the model started to overfit.

- **KNN**
  - Tuned hyperparameters: n_neighbors, p (distance function)
- **Logistic regression**
  - Tuned hyperparameters: tolerance, max_iter
- **Gradient Boosting**
  - Tuned hyperparameters: n_estimators, max_features, min_samples_leaf, learning_rate, max_depth
  - The Gradient Boosting model is one of the models that return the highest PR AUC, but considering its long training time, it might be time-consuming to replicate when we want to improve the model and refit it, so we did not select this model as our best model.

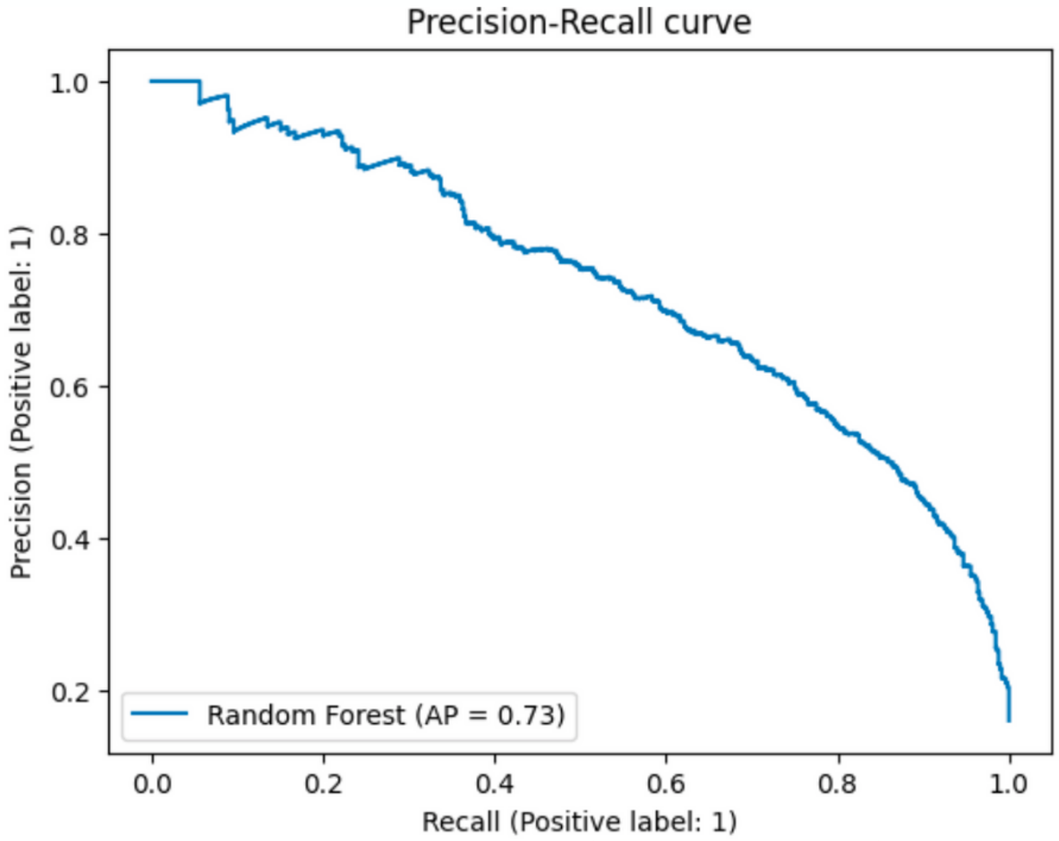
Strategy for Hyperparameter Tuning

- **Random Forest**
  - Tuned hyperparameters: n_estimators, min_sample_leaf, max_depth, criterion (loss function)
  - The Random Forest model is one of the models that return the highest PR AUC. Since it has a shorter training time, it might be convenient to replicate when we want to improve the model and refit it, so we select this model as our best model.
- **MLP (with 0/1/2/3 hidden layers)**
  - Static parameters: This part is where we did some random testing on it. Since fine-tuning an MLP model is time-consuming, we decided to do some random testing on some hyperparameters and make them static during the rest of the hyperparameter tuning process. The strategy of this random testing is that keep all the remaining hyperparameters the same, and tune only one parameter which could help to save the training time at each trial. In other words, the static parameters are all the parameters that reduce the training time.
    - shuffle = True: This is applied to keep the randomness of the minibatch.
    - batch_size = 256: Since our dataset is large, we used Adam optimizer with the minibatch size of 256. Introducing this minibatch could help speed up the training process.
    - max_iter = 200: The default value of this parameter is 200, but this is not enough for the gradient descent algorithm to reach its optimal for our dataset. Thus, we set this value to 600.

Strategy for Hyperparameter Tuning

- **MLP (with 0/1/2/3 hidden layers)**
  - Static parameters: (Continued)
    - activation = ‘relu’: Throughout the process of tuning the activation function, we have tried tanh, sigmoid, ReLU, and identity functions. We found that in most cases, the MLP model with the activation function of either ReLU or tanh outperforms other models. However, ReLU is more efficient for training the model, so we decided to use ReLU as the activation function for the rest of the process of hyperparameter tuning.
  - Tuned hyperparameters:
    - hidden_layer_sizes: We fine-tuned the MLP with 0, 1, 2, and 3 hidden layers respectively. For each hidden layer, we tuned different numbers of hidden units. To decide which number of hidden units will be tuned, we again did some random trials on the model with different hidden units, and so we determined a range from 128 to 512 (all base 2) hidden units that will be tuned for each hidden layer.
    - alpha: We tuned the L2 regularization term between 1e-3 and 1e-2.
    - momentum: We tuned the momentum among 0.1, 0.2, 0.5.
    - learning_rate_init: Again, we did some random trials to determine the possible tuning value for the learning rate. We tuned the learning rate among 1e-2, 2e-2, 5e-3.

Best Model Performance



Accuracy	0.8957
Precision	0.91713222 for Class 0 0.73549884 for Class 1
Recall	0.96285435 for Class 0 0.54280822 for Class 1
F1 score	0.93943729 for Class 0 0.62463054 for Class 1

- Our model achieves an accuracy of 0.8957 which shows that it could classify 0.8957 of the dataset correctly. In addition, from the result of Precision, Recall, and F1 score, we observe that our model classifies Class 0 better than Class 1. This might be due to the imbalance of dataset.

Implications of Model Results for Business Strategy

- **Revenue Forecasting:**

The model's analysis of historical sales data identifies trends and seasonality, aiding in accurate future sales predictions. This insight is crucial for aligning inventory with projected demand, optimizing stock management, and reducing wastage or stockouts.

- **Marketing Optimization:**

The model's predictive capabilities could guide the timing and targeting of marketing campaigns. For instance, it can identify that customers with frequent product browsing behavior are more likely to make a purchase, enabling personalized promotion pushes.

- **Resource Allocation:**

The model offers strategic guidance for resource investment by projecting future sales, highlighting period with high ROI potential, and advising caution in riskier period.

- **Risk Management:**

Provides a data-informed basis for financial decision-making, allowing for proactive budget adjustments to guard against potential downturns and leverage forecasted growth.

Gaining Competitive Edge with Predictive Analytics

- **Market Positioning:**

Predictive analytics help in understanding market dynamics evolution, informing strategic positioning to adapt to market changes faster than competitors.The analysis of 'SpecialDay' and 'Month' columns can determine the effectiveness of promotional campaigns during holiday seasons. For example, if we observe an increase in 'PageValues' during the month of November, it suggests a strong potential for Black Friday sales, and we can position marketing efforts accordingly.

- **Strategy Development:**

Leveraging data-driven insights for crafting strategies that outperform competitors, design effective promotions, and target underserved market segments. The 'ProductRelated' and 'ProductRelated_Duration' data can help in developing strategies for cross-selling and up-selling by identifying products that frequently co-occur in browsing sessions.

- **Customer Insights:**

Enhancing understanding of customer behaviors and preferences to tailor offerings and differentiate the brand in the market.'BounceRates' and 'ExitRates' offer insights into potential pain points on the website. A high bounce rate might indicate that users do not find what they are looking for, prompting a strategy to improve content relevance.

Influencing Market Trends and Future Directions

- **Market Trend Analysis:**

The model's predictive capabilities extend to forecasting market shifts and changing customer needs, enabling businesses to anticipate trends and act proactively.

- **Long-Term Planning:**

Predictive analytics inform product development roadmaps, market expansion, and strategic plans, ensuring alignment with market projections.

- **Continuous Improvement:**

It's crucial to adapt the model to new data and emerging trends to maintain strategic relevance and informed decision-making.

GOT QUESTIONS?

Thank you

Appendix

Data Description

Original Dataset Snippet

online_shoppers_intention

Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0	0	0	0	1	0	0.2	0.2	0	0	Feb	1	1	1	1	Returning_Visitor	FALSE	FALSE
0	0	0	0	2	64	0	0.1	0	0	Feb	2	2	1	2	Returning_Visitor	FALSE	FALSE
0	0	0	0	1	0	0.2	0.2	0	0	Feb	4	1	9	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	2	2.666666667	0.05	0.14	0	0	Feb	3	2	2	4	Returning_Visitor	FALSE	FALSE
0	0	0	0	10	627.5	0.02	0.05	0	0	Feb	3	3	1	4	Returning_Visitor	TRUE	FALSE
0	0	0	0	19	154.2166667	0.015789474	0.024561404	0	0	Feb	2	2	1	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	1	0	0.2	0.2	0	0.4	Feb	2	4	3	3	Returning_Visitor	FALSE	FALSE
1	0	0	0	0	0	0.2	0.2	0	0	Feb	1	2	1	5	Returning_Visitor	TRUE	FALSE
0	0	0	0	2	37	0	0.1	0	0.8	Feb	2	2	2	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	3	738	0	0.022222222	0	0.4	Feb	2	4	1	2	Returning_Visitor	FALSE	FALSE

Granularity: Dataset contains all web-viewing sessions within a one-year period, and each record represents a single session from a unique user.

Categorical Variables: Month, OperatingSystems, Browser, Region, TrafficType, Visitor Type, Weekend, Revenue

Numerical Variables: Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, ProductRelateD_Duration, BounceRates, ExitRates, PageValues, SpecialDay

Data Description

Web Page Information

- Administrative, Informational, ProductRelated: the number of pages within each page category visited by the customer in a particular session
- Administrative_Duration, Informational_Duration, ProductRelated_Duration: the total time spend in each of those webpage categories within a particular session
- Value of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action

Google Analytics Metrics

- BounceRate: Percentage of visitors who enter the site from a specific webpage and then leave(“bounce”) without any other actions
- ExitRate: For page views of a specific webpage, the percentage that this page is the last in a visiting session (i.e., leave the website from this page)
- PageValue: The average value for a web page that a user visited before completing an transaction

Data Description

Individual Customer Information

- OperativeSystem: 1-8
- Browser: 1-8
- TrafficType: 1-20
- VisitorType: New_Visitor, Returning_Visitor, Other
- Except VisitorType, other variables were already transformed from categories in text to numerical representations

Customer Visit Details

- SpecialDay: Ranges from 0 and 1, it indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which a session is more likely to be finalized with transaction. The value of this variable is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. Considering it usually takes a week for items to ship and arrive after placing an order, the variable has maximum value of 1 about 6-7 days before the special day, and 0 on the special day.
- Month: Month of the year a session takes place
- Weekend: Boolean indicating whether a session takes place during the weekend
- Revenue: Dependent variable in boolean

Selected features using different feature selection methods

LASSO:
ExitRates, PageValues

Z-Score (Linear regression):
Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, ProductRelated_Duration, BounceRates, ExitRates, PageValues, SpecialDay, Weekend, Month_May, Month_Nov, OperatingSystems_2, OperatingSystems_3, Browser_2, Browser_3, Region_3, TrafficType_2, VisitorType_Returning_Visitor

Mutual Information Score: (select the top 30 features with higher MI scores and rank by the MI Score, this could give an insight into the significance of the features to the target variable)

Browser_6, VisitorType_Other, Browser_5, Region_7, OperatingSystems_6, OperatingSystems_5, Region_9, TrafficType_7, OperatingSystems_2, Month_Sep, Month_Oct, TrafficType_8, Region_5, Month_June, TrafficType_9, TrafficType_15, TrafficType_20, TrafficType_18, TrafficType_17, Month_Jul, TrafficType_14, TrafficType_13, Region_2, TrafficType_11, SpecialDay, TrafficType_10, Month_Dec, Month_Feb, TrafficType_12, Region_6

Reference

Sakar, C., & Kastro, Y. (2018). Online Shoppers Purchasing Intention Dataset [Dataset]. https://doi.org/10.24432/C5F88Q