Online Shoppers Purchasing Intention

Victory (Victor) Ma

Goal: Predict whether an online customer will generate revenue based on various features related to their time on the website

Objective: Using results from the model, we can then fine-tune the website to maximize profit for the business

(maximizing shareholder profit is definitely the goal in life)

Opening, Challenge, Action, & Resolution

Opening: We are using the *Online Shoppers Purchasing Intention Dataset* from UC Irvine's Machine Learning Repository, which has 18 variables

Challenge: Can we accurately predict if a customer will purchase something during their time on the website?

Action: We will follow the essential workflow outlined by CSCI 200B in order to create an effective model for the company to use

Resolution: Once we have trained and evaluated our models, we will have a greater understanding of what indicators are most important when predicting whether or not a customer purchases the product on the website

Description of the Dataset

This particular dataset from UC Irvine gives us 10 numerical variables, 7 categorical variables, and 1 target variable

- Features such as **“Informational”, “Informational Duration”, “Product Related”** and **“Product Related Duration”** represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories
- The value of **“Bounce Rate”** refers to the percentage of customers who enter the site from that page and then leave without triggering any other requests
- The value of **“Exit Rate”** is calculated as for all pageviews to the page, the percentage that were the last in the session
- The target variable is named **“Revenue”**, indicating whether or not a purchase was made during the session

Exploring The Data

Step 1: Check Data

- Reviewed variable types and checked for missing values
- Noted that “Revenue” is a binary class, 1 indicating a purchase and 0 indicating no purchase
- Initial analysis showcased that...

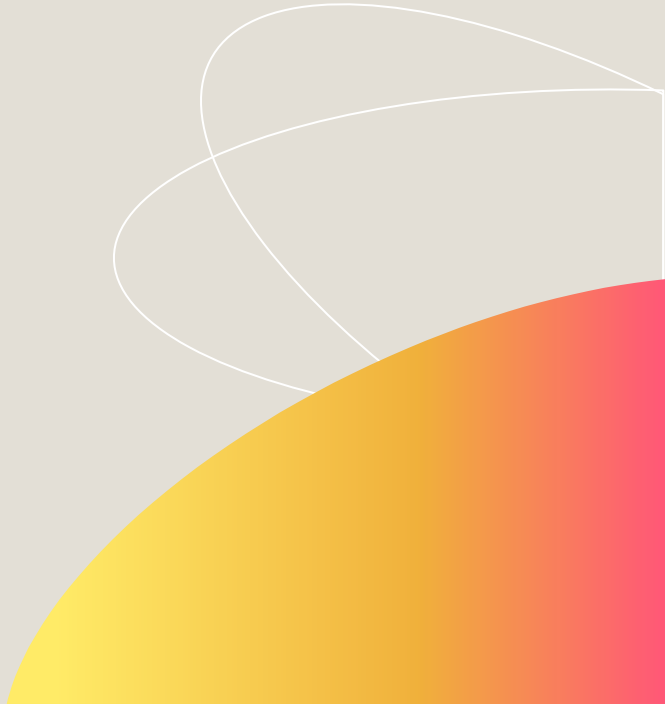
Step 2: Visualizations

- Generated basic bar plots, scatterplots, histograms, etc. to understand the features and relationships much better

Step 3: Statistical Tests/Measures

- Conducted statistical tests (ANOVA, T-Tests, etc.) to identify key relationships between features and the target

Visualizations / Statistical Measures



Data Cleaning & Wrangling

- **Confirmed that there are no missing values in the columns**
- **Checked for duplicates/strange values (i.e. NA, None, etc.)**
- **Checked for outliers or problematic ranges of values**
- **OHE (One-Hot Encoded) categorical features such as "Weekend" and "Revenue"**
- **No other features need transformation, which is great**
- **Based on descriptions and context, we will select 8-10 main features to include**

Selecting Our Models

Our learning approach is Supervised Learning for Classification, because we have a defined target variable—Revenue—indicating a class (purchase or no purchase)

Models That Were Chosen:

Null Model: Assigning all observations to the majority class (in this case, no revenue). This will be our baseline or benchmark for the following models

Logistic Regression: Utilized for its simplicity and interpretability our problem, because its binary classification

Random Forest With Boosting: Utilized for its potential to capture more complex relationships in the data and accuracy for classification tasks

| *Training Our Models*

1

Data Split

PUT METHODS HERE

2

Tuning Hyperparameters

PUT METHODS HERE

3

Fitting The Model

PUT METHODS HERE

4

Measuring Performance

PUT METHODS HERE

Evaluating Our Models

1. Calculate Performance Metrics
2. Compare Models
3. Analyze Confusion Matrices
4. Generate Model Insight

Overall Insights & Concluding Ideas!

...

Any questions?

Thank you for helping me maximize shareholder profit. I am sure that they will be happy.