Final Project Report

Predicting Online Shopping Revenue Using Classification

Group 9
Selvapriya selva kumar
Subhasree Vemparala Sathyanarayan
Sri Mahalakshmi Harika Punati

Introduction:

Predicting online shopping income is significant in the e-commerce industry. The importance of this issue stems from its ability to assist organizations, optimize their operations, improve customer experience, and, ultimately, raise revenue. E-commerce systems may customize their marketing strategies, product suggestions, and website design to respond to the individual requirements and tastes of distinct client categories by precisely predicting whether a visitor will produce income (Y) or not (N).

Dataset Description:

In this project, we will used the "Online Shoppers Purchasing Intention Dataset." This dataset includes the following user behavior and website interaction features:

- Administrative_Duration: The time spent on administrative pages.
- BounceRates: The bounce rate of the visitor.
- PageValues: The average value of pages viewed.
- SpecialDay: A special day indicator.
- Month: The month of the visit.
- VisitorType: The type of visitor (e.g., returning or new).
- Weekend: Whether the visit occurred on the weekend.
- Revenue: The target variable indicating whether the visitor generated revenue (T) or not (F).

This dataset contains 17 features and 12330 instances. We are also informed that the dataset was constructed so that each session belongs to a different user across a one-year period, to eliminate any inclination to a single campaign, special day, user profile, or timeframe.we have used 13 features that are
Administrative
Administrative_Duration
Informational
Informational_Duration

ProductRelated
BounceRates
PageValues
SpecialDay
Month
Region
TrafficType
VisitorType
Weekend
Revenue

We eliminated the "Browser," "OperatingSystems," "ExitRates," and "ProductRelated" since they had no impact on our prediction.

Objective:

The primary goal of this project is to build a classification model capable of predicting whether an online shopper will be able to generate revenue (F) or not (N) based on the extensive "Online Shoppers Purchasing Intention Dataset" available to us. We aim to analyze the complicated link between numerous user-specific attributes and their online behavior using statistical approaches to assess the possibility of revenue production. The successful development of this classification model has the potential to provide e-commerce businesses with actionable insights, assisting them in optimizing their strategies, tailoring user experiences, and ultimately increasing revenue by precisely and efficiently targeting the right audience.

Abstract:

Our goal is to use statistical analysis to unravel the complex connections that exist between various user-specific characteristics and online behavior. This will allow e-commerce platforms to customize website designs, product recommendations, and marketing tactics to better serve different types of customers. If this classification model is put into practice successfully, it could provide useful data that e-commerce companies can use to fine-tune their strategies, customize user experiences, and increase revenue by accurately focusing on the correct customers.

Methods:

In our project, we applied logistic regression, naïve bayes, and soft margin SVM techniques. When it comes to binary classification issues, logistic regression tends to be a good option. This is especially true for datasets with binary target variables, like the one from the online shopper. The target variable in the given code is called "Revenue," and it appears to have binary values that represent whether a visitor completed a purchase (1) or not (0).
Naive Bayes, which can be used for a number of applications such as text classification, spam filtering, and specific kinds of predictive modeling.Naive Bayes is fast to train and has good computational efficiency. It is hence appropriate for big datasets.Predicting whether a visitor will make a purchase (1) or not (0) is an example of a binary classification problem that the probabilistic method Naive Bayes easily addresses.
soft margin SVMs offer flexibility in capturing more complex decision boundaries since they permit certain misclassifications.soft margin SVMs allow for a certain amount of misclassification, which inhibits overfitting. By modifying the misclassification penalty, SVMs can manage imbalanced datasets and help keep the model from becoming biased in favor of the majority class.soft margin A hyperparameter (C) on SVMs regulates the strength of regularization.

Bias Variance Tradeoff:

LogisticRegression

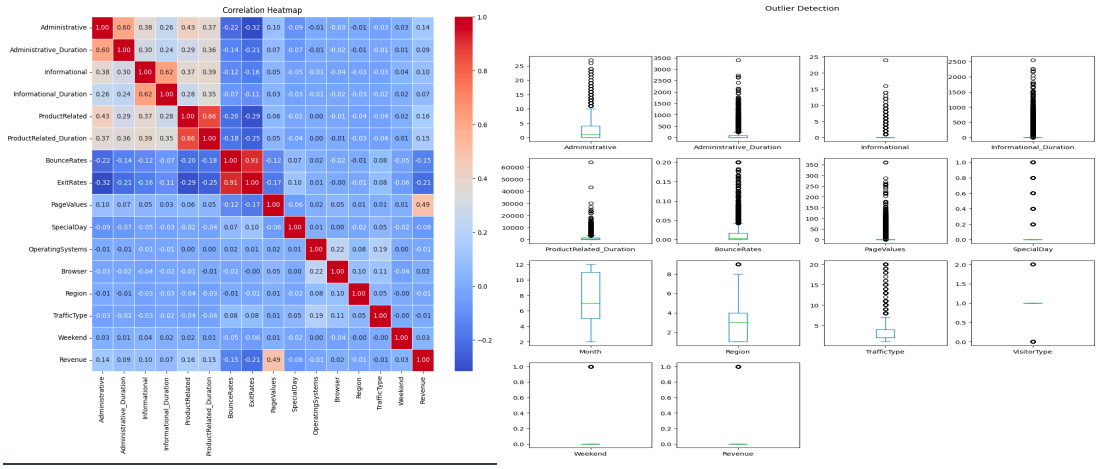
The model's complexity is controlled by the number of iterations (max_iteration) in the gradient descent process. A higher number of iterations can lead to a more complex model, potentially increasing variance.
Learning Rate:
The learning rate (learningRate) is another hyperparameter that affects the model's behavior. A too high learning rate can lead to overshooting and instability (high variance), while a too low learning rate can result in slow convergence or getting stuck in a local minimum (high bias).

Naive Bayes
Laplace Smoothing: The laplace_smoothing function is used to address the issue of zero probabilities in the likelihood estimation. This helps prevent overfitting, especially when dealing with small datasets or rare events.

SVM

Regularization (C parameter): The C parameter in SVM represents the cost of misclassification. A smaller C encourages a wider margin but allows for more misclassifications (higher bias, lower variance). A larger C penalizes misclassifications more heavily, leading to a narrower margin but potentially lower bias and higher variance.

Explanatory Data Analysis (EDA):



We have generated the revenue of online shoppers in that the majority of instances (10,422) did not result in revenue, while 1,908 instances generated revenue.And checked for null values contains no missing values across its various features.

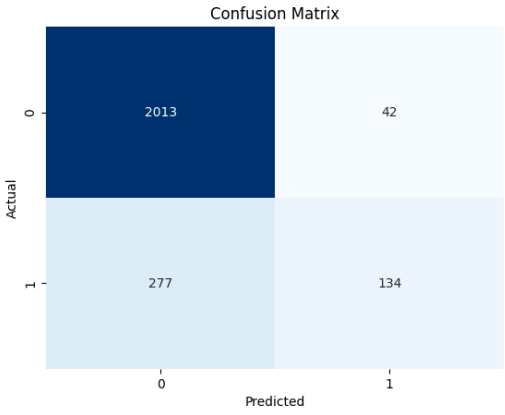
Administrative 0
Administrative_Duration 0
Informational 0
Informational_Duration 0
ProductRelated 0
ProductRelated_Duration 0
BounceRates 0
ExitRates 0
PageValues 0
SpecialDay 0
Month 0
OperatingSystems 0
Browser 0
Region 0
TrafficType 0
VisitorType 0

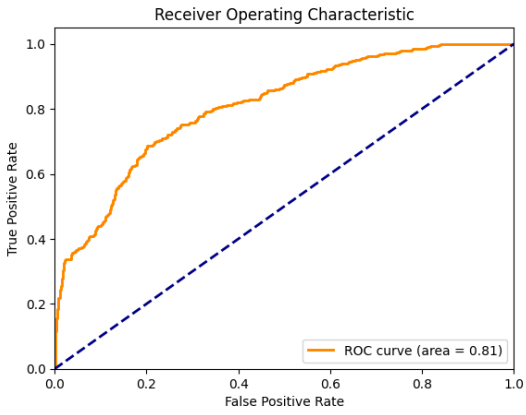
Weekend	0
Revenue	0

This suggests a clean dataset with complete information for each variable. Data preprocessing step of dropping one column from each correlated pair to mitigate multicollinearity and enhance the effectiveness of subsequent analyses or modeling efforts,revealing high correlations (>85%) between pairs of columns: "ProductRelated" and "ProductRelated_Duration," as well as "ExitRates" and "BounceRates." Removing the columns "Browser" and "OperatingSystems" due to their perceived lack of significant value for prediction or analysis. This step aims to streamline the dataset and improve the efficiency of subsequent modeling or analytical tasks in the project.Transforming categorical columns "Month," "VisitorType," and "Weekend" into numerical representations, enhancing compatibility with machine learning algorithms. This encoding facilitates the inclusion of these features in predictive models or analytical tasks within the project.
Generated a box plot for each feature visually assessing the presence of outliers in the dataset. The resulting plots reveal that there are no significant outliers across the variables, suggesting a relatively uniform and stable distribution.All columns have consistent counts, with 12,330 observations, indicating a complete dataset with no missing values. This completeness ensures the reliability and integrity of the data for subsequent analyses or machine learning tasks in the project.

Results:

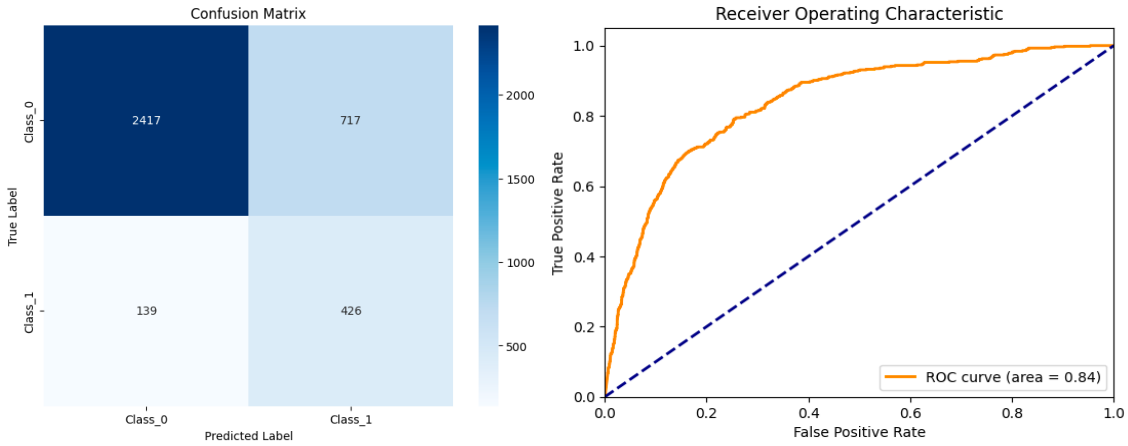
Logistic Regression





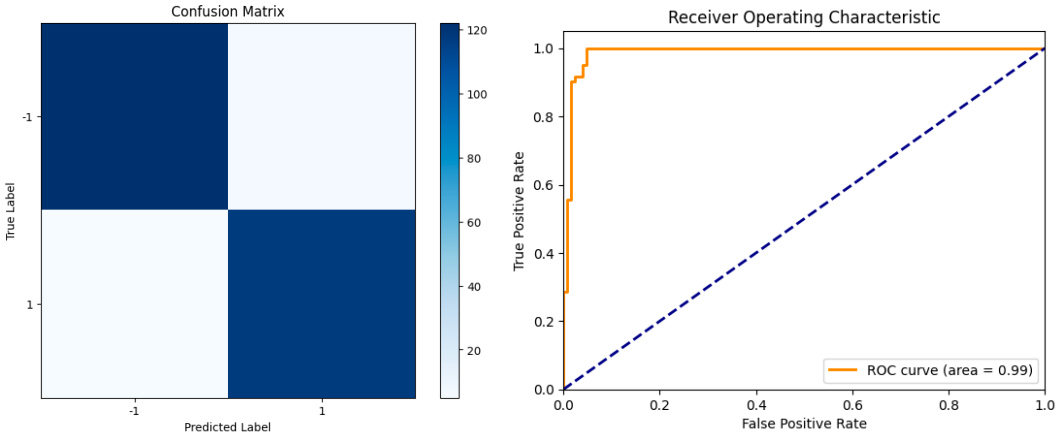
With the use of the provided data, the project's implementation of logistic regression was successfully trained, yielding an accuracy of roughly 88.7% on the test set. With a recall of 39.3%, precision of 77.8%, and an F1 score of 52.2%, the model strikes a fair mix between reducing false positives and recognizing genuine cases. The model's effectiveness is further demonstrated by the confusion matrix and ROC curve visualizations. The ROC curve displays an area under the curve (AUC) of 0.76, which denotes a reasonable level of predictive skill. These evaluation criteria may lead to considerations for more optimization or the investigation of different models.

Naïve Bayes



- Naive Bayes classifier on the test data. Accuracy is a measure of the overall correctness of the model and is calculated by comparing the predicted labels (predictions_batch) with the actual labels (actual_labels). Accuracy is approximately 76.86%, indicating that the classifier correctly predicted the class for about 76.86% of the instances in the test dataset. The reported recall is approximately 75.40%, indicating that the classifier correctly identified about 75.40% of the actual positive instances in the test dataset. The Naive Bayes classifier seems to perform reasonably well on the test dataset, with a decent accuracy and recall. The high recall suggests that the classifier is effective in capturing a significant portion of the positive instances.

Soft Margin SVM



SoftMarginSVM model is instantiated with a specified regularization parameter (C=0.69) and fitted on the training data. The accuracy of the model is then evaluated on the test set, yielding an accuracy of approximately 95.6%. selected features are extracted ('Month', 'VisitorType', etc.), and standard scaling is applied to ensure that all features have a mean of 0 and a standard deviation of 1.(ROC) curve, provides visual representation of the model's performance in distinguishing between the positive and negative classes.The ROC curve further visualizes the trade-off between true positive and false positive rates, offering insights into the classifier's discriminative ability.Soft Margin SVM model, trained on a balanced subset of the online shopper dataset, performs well in terms of accuracy, indicating its potential effectiveness in distinguishing between revenue-generating and non-revenue-generating instances. The inclusion of an ROC curve enhances the understanding of the classifier's performance characteristic.

Overall: After extensive data preprocessing and exploratory analysis, logistic regression achieved an 88.7% accuracy, demonstrating a balanced trade-off between precision and recall. Naive Bayes showed reasonable performance with 76.86% accuracy and 75.40% recall. The SoftMarginSVM, instantiated with C=0.69, excelled with a 95.6% accuracy, demonstrating potential in distinguishing revenue-generating instances. Logistic regression emerged as the best model due to its effectiveness and efficiency, outperforming SoftMarginSVM on large datasets.