Final Project Report: [Predicting Online Shopping Purchasing Intention Using Random Forest and AdaBoost Classifier]

Course: CS383 - Introduction to Machine Learning
Semester: Fall 2024
Team Members: Amy Li, Cecilia Chen
Instructor: Adam Poliak

Table of Contents

Abstract

This project aimed to predict online shopping purchasing intentions by analyzing customer browsing patterns. Utilizing a dataset from the UCI Machine Learning Repository, we employed ensemble learning methods, specifically Random Forest and AdaBoost, to identify key factors influencing purchase decisions and to develop accurate predictive models. Our methodology involved data preprocessing, hyperparameter tuning using techniques like GridSearchCV, and model evaluation through train-test splits and cross-validation. We used Python packages such as pandas, scikit-learn, and matplotlib for data manipulation, model training, and visualization.

Key Findings:

1. Model Performance: Both Random Forest and AdaBoost demonstrated high accuracy, with Random Forest showing a slight edge in precision and AdaBoost in recall and F1 Score.
2. Hyperparameter Tuning: For AdaBoost, an optimal range for n_estimators was found to be between 200 and 400, and a learning rate of around 0.02. Random Forest benefits from around 45 n_estimators, a min_samples_split of 12 to 17, and a max_depth of 6 to 16. 3.Feature Importance: PageValues, ExitRates, and ProductRelated_Duration were identified as the most influential features in predicting purchasing intentions.

Introduction

Online shopping has become a dominant mode of consumer behavior, making it critical for online shopping platforms to understand the factors influencing purchasing decisions. Predicting whether a shopper will complete a purchase based on their browsing behavior can enable businesses to optimize their platforms designs and marketing strategies, leading to increased revenue and customer satisfaction.

This project explores the prediction of online shopping purchasing intentions using machine learning. By analyzing a dataset that captures customer browsing patterns, the project aims to identify key factors influencing purchase decisions and develop models that accurately predict purchase outcomes. Applying ensemble learning methods, Random Forest and AdaBoost, we aim to discover actionable insights while improving prediction accuracy.

Problem Statement

The central problem addressed in this project is:

How can machine learning models predict online shopping purchasing intentions, and what are the key factors driving purchase decisions?

The hypotheses for this study are as follows:

1. Features like time spent on product-related pages and the value of pages viewed are likely to have a strong influence on purchasing behavior.
2. Ensemble learning algorithms, such as Random Forest and AdaBoost, are expected to outperform simpler models because they can effectively capture complex interactions within the dataset.

This project focuses on the following tasks:

- Preparing and analyzing a dataset of online shopping sessions for better finding meaningful patterns.
- Conducting hyperparameter tuning to observe the influence of different hyperparameters at varying levels and optimize model performance.
- Developing and evaluating machine learning models (Random Forest and AdaBoost) with the best hyper parameter combination to predict purchasing intentions with high accuracy.
- Exploring the top features that drive purchasing decisions by analyzing feature importance scores generated by the models.

Related Work

- In "Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers", Wyner et al. offer a theoretical explanation for the success of AdaBoost and Random Forests. It suggests that both algorithms work well due to their ability to interpolate the training data completely without error, combined with a self-averaging property that leads to low generalization error. The paper challenges conventional wisdom about the need for regularization or early stopping in boosting algorithms and proposes using AdaBoost like Random Forests, with large decision trees and without regularization or early stopping. This inspired us to use both Random Forests and AdaBoost to evaluate the Online Shopping Purrchasing Intention dataset.

- The paper "Hyperparameters and Tuning Strategies for Random Forest" written by Probst et al. introduces model-based optimization (MBO) as an effective tuning strategy and presents the tuneRanger R package to automate this process. Benchmark studies demonstrate the improved prediction performance and efficiency of tuneRanger compared to default settings and other tuning methods. The paper emphasizes the importance of tuning hyperparameters like n_estimators (the number of trees), max_features (variables drawn for each split), min_samples_split (minimum samples for a node to split), and min_samples_leaf (minimum samples in a leaf node) for optimizing Random Forest performance. It confirms that while Random Forests perform well with default settings, hyperparameter tuning can further enhance model accuracy. Therefore, we decided to include n_estimators, max_features, and min_samples_split tuning for hyper parameter tuning.

Data Description

Describe the dataset(s) you used, including:

- **Source(s)**: The dataset in question is sourced from the UCI Machine Learning Library , as mentioned by its creators, C. Okan Sakar from the Department of Computer Engineering, Bahcesehir University, and Yomi Kastro from Inveon Information Technologies Consultancy and Trade . It is specifically designed for the subject area of business and is associated with tasks such as classification and clustering.

- **Size and Format**: The dataset is available in a CSV format, which is a common structured data format that can be easily used for machine learning tasks. It consists of 12,330 instances, each representing a user session over a one-year period. There are 17 features in total, with 10 being numerical and 7 being categorical. The numerical features include metrics like 'Administrative Duration', 'Informational Duration', and 'ProductRelated Duration', which are derived from the URL information of the pages visited by the user. The categorical features include attributes such as 'Month', 'OperatingSystems', 'Browser', 'Region', 'TrafficType', 'VisitorType', and 'Weekend' .

- **Preprocessing**: The dataset is clean and doesn't have any missing data.

Methodology

1. In our approach, we utilized two prominent ensemble learning algorithms, AdaBoost and Random Forest, to tackle classification tasks. We structured our training process with a meticulous train-test split to assess the model's performance on unseen data, and we also implemented cross-validation to bolster the reliability of our models. For hyperparameter tuning, we meticulously adjusted key parameters: for AdaBoost, we focused on n_estimators and learning_rate, while for Random Forest, we fine-tuned n_estimators, max_features, min_samples_split, and min_samples_leaf. We employed a variety of Python packages including ucimlrepo for dataset retrieval, pandas for data manipulation, scikit-learn for model training and evaluation, and matplotlib.pyplot for data visualization. Our analysis focused on the AdaBoostClassifier and RandomForestClassifier from scikit-learn to perform classification tasks, with numpy and random supporting numerical operations and randomization needs.
2. We performed Model Comparision Based on Best Parameter Prediction, which applied Random Forest and AdaBoost algorithms to predict online shopping purchasing intentions, using a 90-10 train-test split and 5-fold cross-validation to ensure generalization. Hyperparameter tuning was performed with GridSearchCV to optimize model parameters, leveraging tools like scikit-learn, pandas, and Matplotlib for modeling and analysis.

Results

Hyperparameter tuning

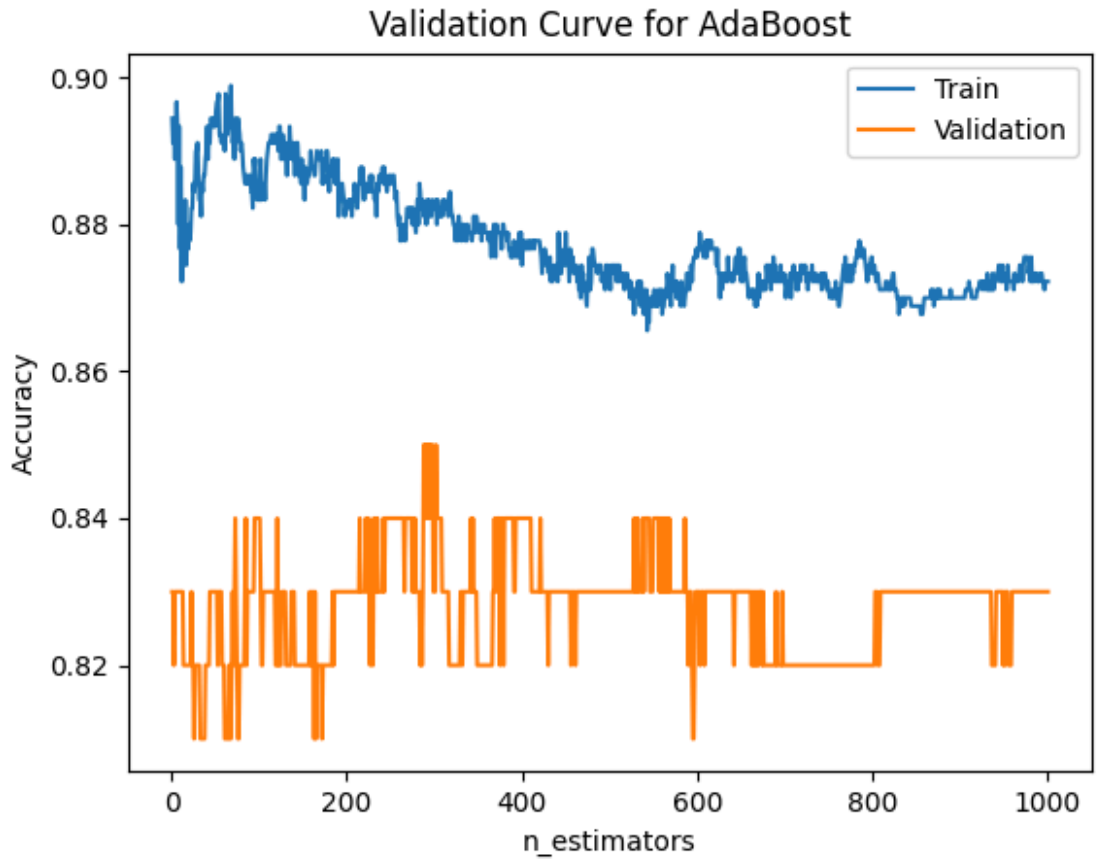
Here we will include the results for hyperparameter tuning for both AdaBoost and Random Forest. We will use validation curve with both training accuracy and validation accuracy to show the sensitivity between changes in AdaBoost and Random Forest's accuracy with changes in hyperparameters of the model.

AdaBoost

n_estimators

The AdaBoost validation curve demonstrates that as the number of estimators increases, the training accuracy initially remains high but then shows a slight downward trend, which is a common sign of the model becoming more complex and possibly overfitting. The validation accuracy, which is crucial for assessing the model's performance on new data, stays relatively consistent and does not significantly improve with additional estimators, indicating that the model is already generalizing well.

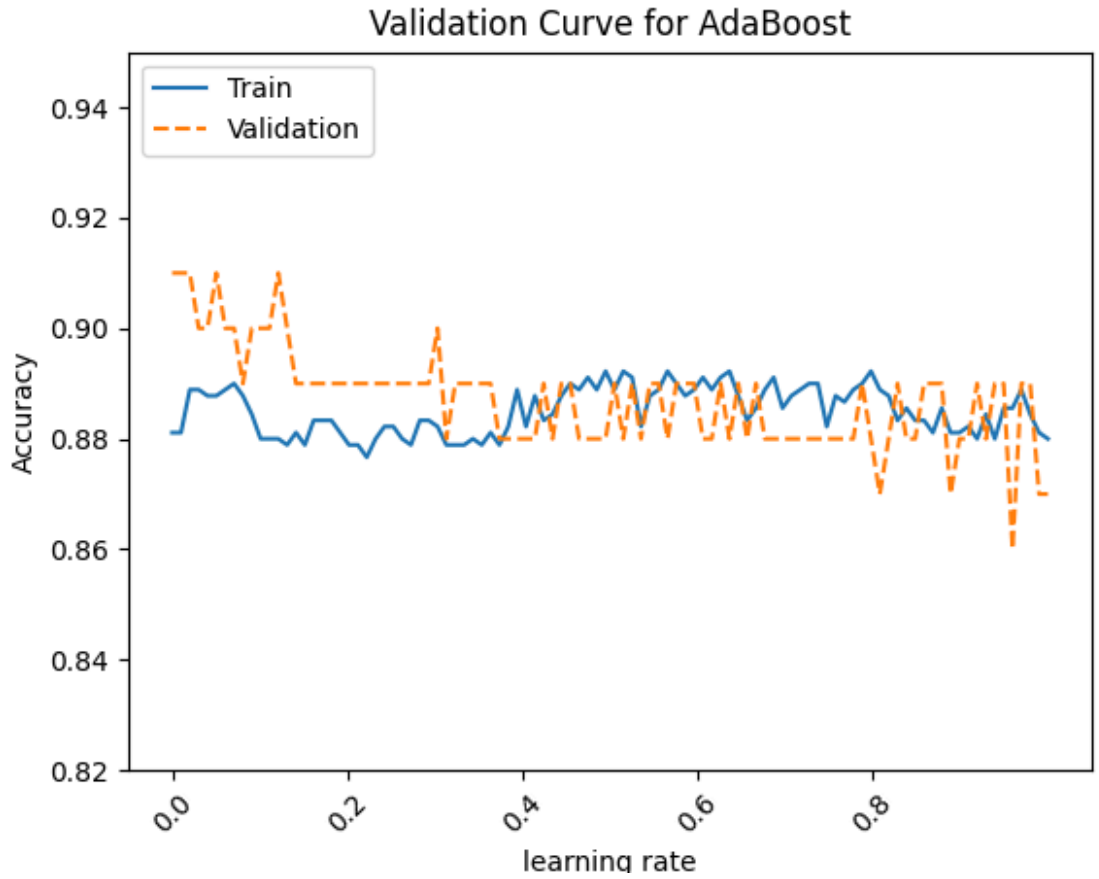
However, there is a noticeable dip in validation accuracy at around 600 estimators, which could be an indication of overfitting or a point of instability. To avoid this and to maintain a good balance between bias and variance, a recommended number of estimators might be between late 200 and 400. This range seems to offer a stable and high validation accuracy, suggesting that the model is performing well without being overly complex.



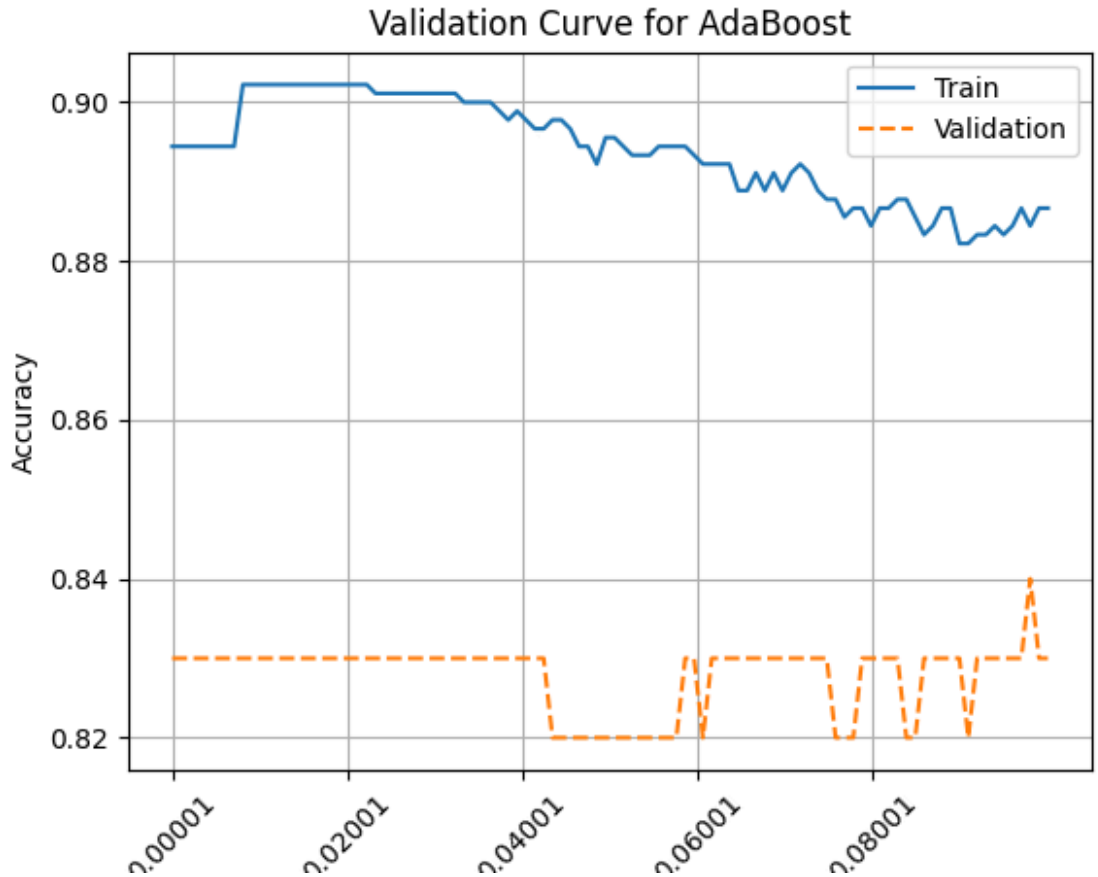
learning rate

The plot shows the validation curve for an AdaBoost model as the learning rate increase from 0.0001 to 1.

The blue line represents the training accuracy, while the dashed orange line represents the validation accuracy. As the learning rate increases, the training accuracy generally improves, but the validation accuracy initially increases, reaches a peak around a learning rate of 0.53, and then starts to decline. This pattern suggests this may not be the right range for tuning best learning rate.



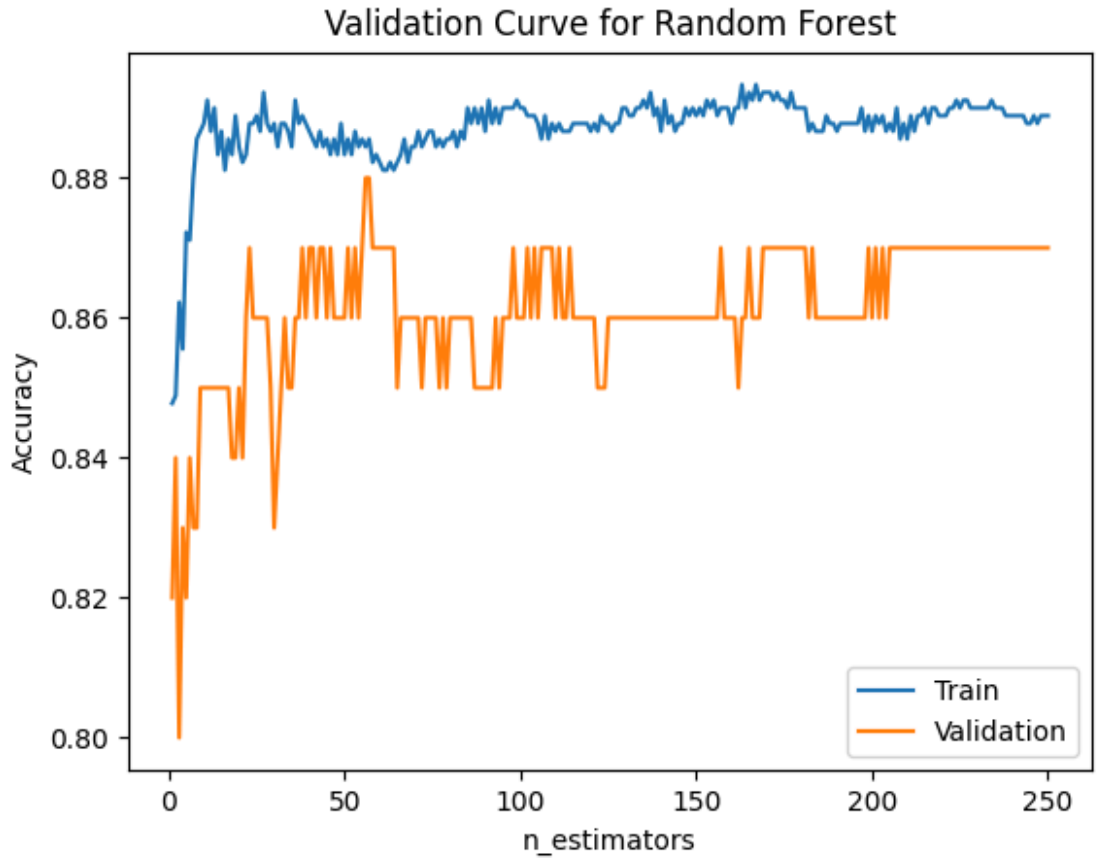
Therefore, we have the other experiment done, which focused on learning rate from 0.00001 to 0.1 with plot shown below. We can see that the training accuracy increases first, then remains a little bit above 0.90 for a while, and then drop. This suggests the best learning rate should be around 2e-2.



Random Forest

n_estimators

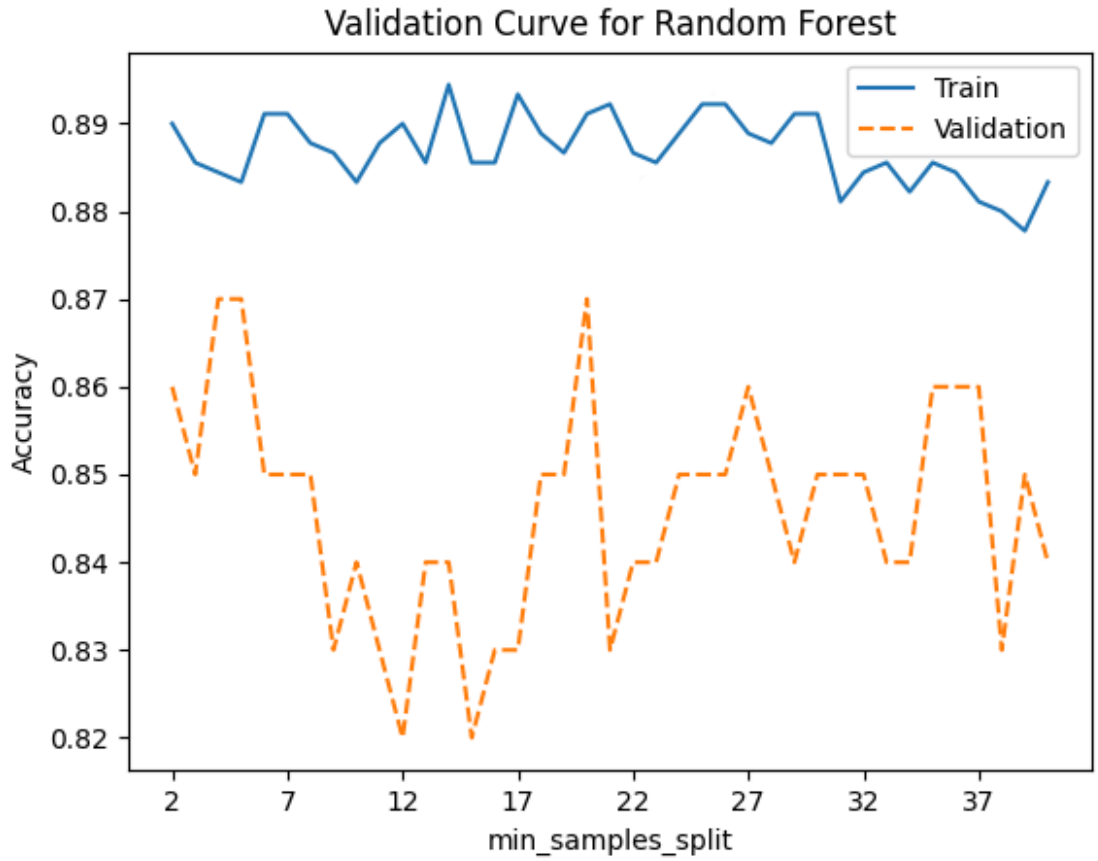
The plot illustrates the validation curve for a Random Forest model, showcasing the impact of the number of n_estimators on both training and validation accuracy. As the number of n_estimators increases, the training accuracy steadily improves, while the validation accuracy initially increases and then stabilizes around a value a little bit higher than 0.88. This suggests that increasing the number of estimators beyond a certain point (around 50) provides diminishing returns in terms of validation accuracy. Therefore, a model with approximately 45 estimators might be a good balance between model complexity and generalization performance.



min_samples_split

The plot displays the validation curve for a Random Forest model, focusing on the min_samples_split hyperparameter. The training accuracy remains relatively stable and high, around 0.89, indicating the model is learning well on the training data. In contrast, the validation accuracy fluctuates more and is generally lower, around 0.85 to 0.87, suggesting some overfitting.

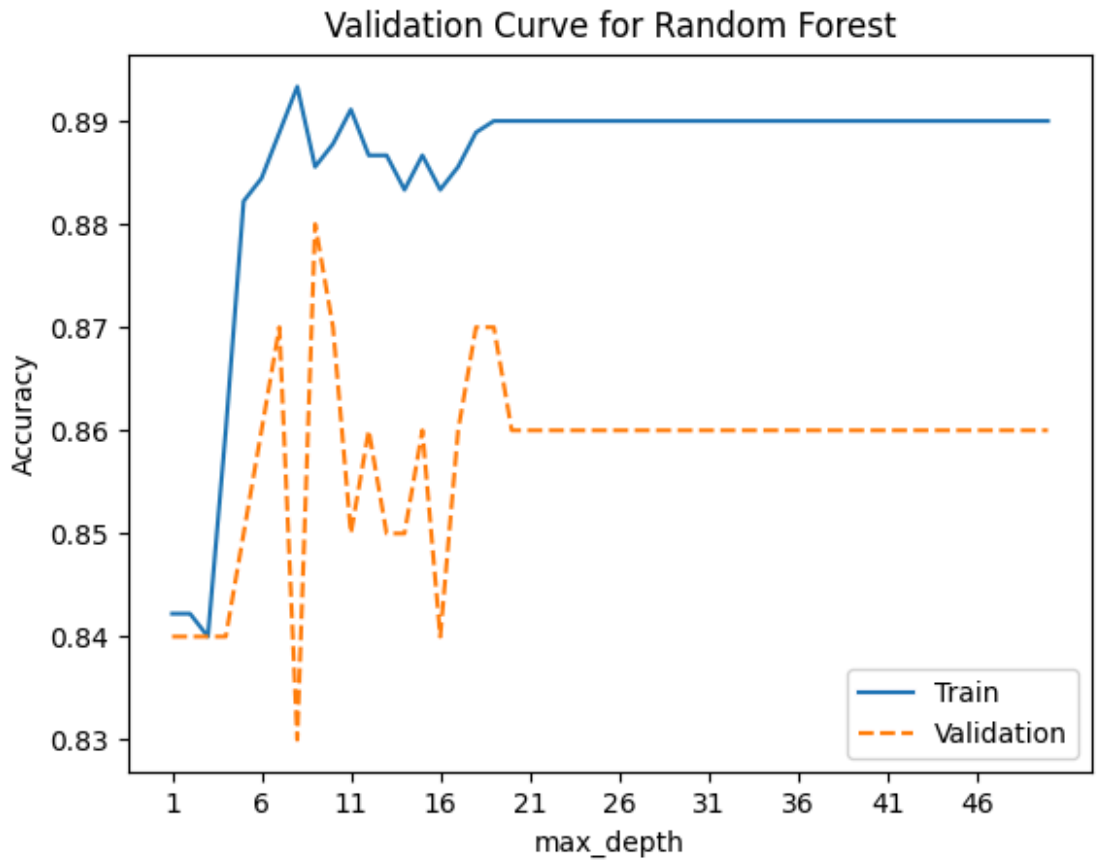
The best balance between bias and variance seems to occur around min_samples_split values of 12 to 17, where the validation accuracy peaks. A suggested value for min_samples_split could be in this range to achieve a good generalization performance on unseen data.



max_depth

The Random Forest validation curve indicates that training accuracy rises sharply and then plateaus as max_depth increases, showing the model's robust learning on training data. In contrast, validation accuracy peaks at a max_depth of around 6, after which it declines, hinting at overfitting with deeper trees.

Post the initial peak, validation accuracy levels off, maintaining a consistent, albeit lower, performance from max_depth of 16 onwards. This flattening suggests diminishing returns on increasing depth. A max_depth of 6 to 16 is thus optimal, balancing model complexity and generalization for the best validation accuracy.



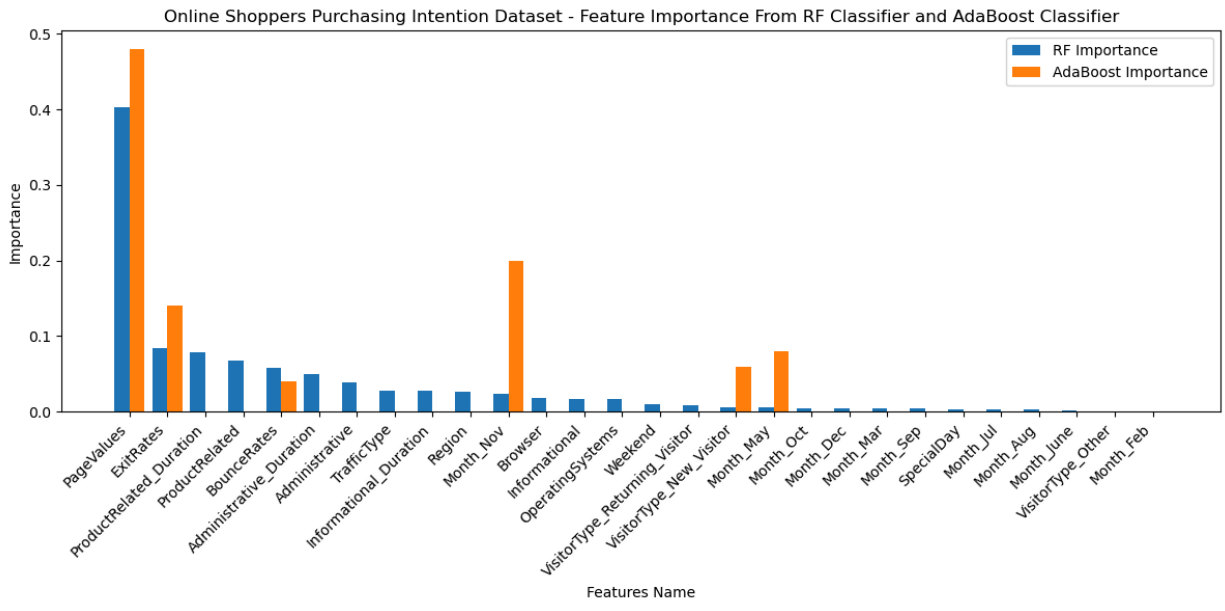
Model Comparision Based on Best Parameter Prediction

Feature Importance

Feature	RF Importance	AdaBoost Importance
PageValues	0.402614	0.48
ExitRates	0.0846152	0.14
ProductRelated_Duration	0.0788203	0
ProductRelated	0.0672446	0
BounceRates	0.0577837	0.04
Administrative_Duration	0.0503969	0
Administrative	0.0391559	0
TrafficType	0.0283468	0
Informational_Duration	0.0280264	0
Region	0.0270764	0
Month_Nov	0.0235772	0.2

Feature	RF Importance	AdaBoost Importance
Browser	0.0175859	0
Informational	0.0170328	0
OperatingSystems	0.0166456	0
Weekend	0.00945192	0
VisitorType_Returning_Visitor	0.00852322	0
VisitorType_New_Visitor	0.00635391	0.06
Month_May	0.0063041	0.08
Month_Oct	0.00500593	0
Month_Dec	0.00498314	0
Month_Mar	0.00471067	0
Month_Sep	0.00387399	0
SpecialDay	0.00348955	0
Month_Jul	0.00327302	0
Month_Aug	0.00261917	0
Month_June	0.00183814	0
VisitorType_Other	0.000426073	0
Month_Feb	0.000225018	0

The feature importance comparison for Random Forest and AdaBoost is shown in a side-by-side bar chart:



The feature importance scores derived from Random Forest and AdaBoost are shown in the chart below. PageValues is the most influential feature for both models, with Random Forest assigning an importance of 40.26% and AdaBoost assigning 48%. Other key features include ExitRates and ProductRelated_Duration, which play significant roles in Random Forest. For AdaBoost, Month_Nov and ExitRates are more emphasized. To make a hyposthesis, Black-Friday could be related with this high importance of November.

Model Performance

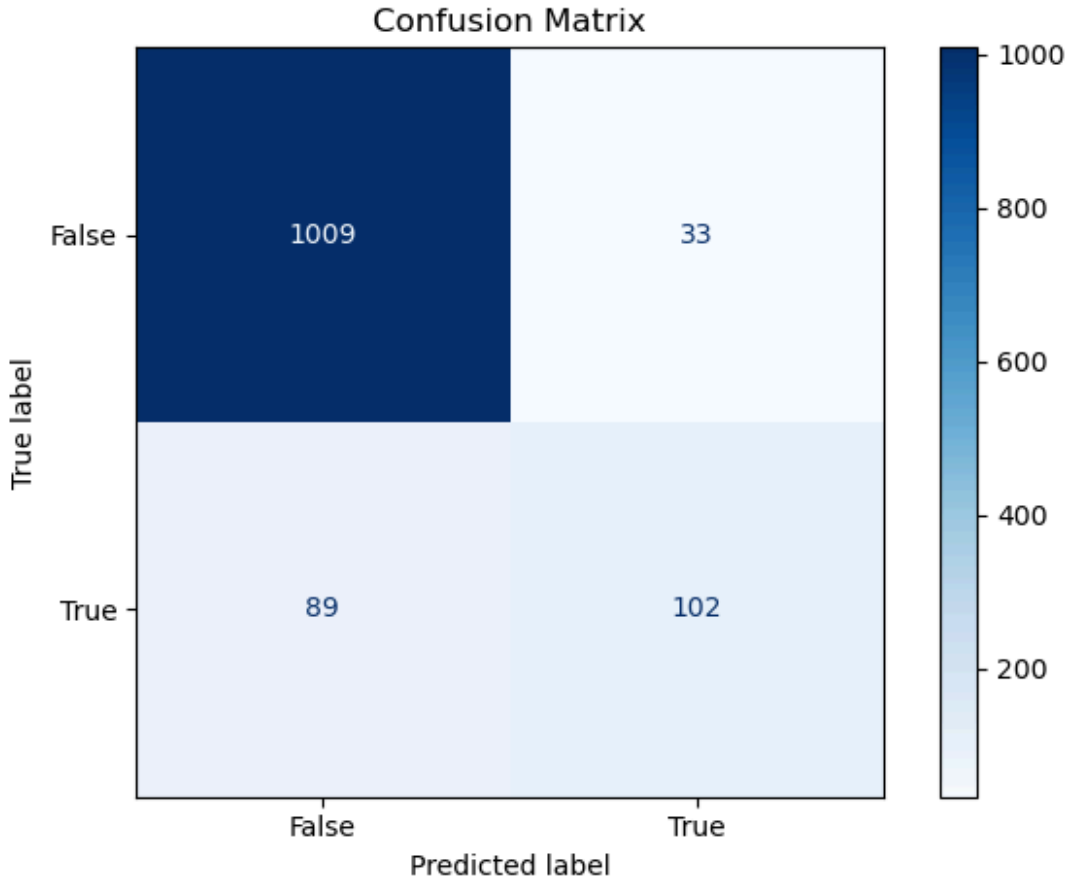
A comparison of the performance metrics for Random Forest and AdaBoost is shown below:

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.901054	0.755556	0.534031	0.625767
AdaBoost	0.898621	0.726027	0.554974	0.62908

Both models have similar accuracy, with Random Forest slightly outperforming AdaBoost in precision, while AdaBoost performed better in recall and F1 Score. This indicates that AdaBoost may be better at identifying actual purchasers due to its emphasis on learning from misclassified instances during training, whereas Random Forest focuses on overall accuracy and precision. Despite these differences, both models demonstrate comparable performance overall.

Confusion Matrix

The performance of the best-performing model, Random Forest, is summarized in a confusion matrix:



The performance of the two models is summarized in their confusion matrices. For Random Forest, the model correctly identified 1009 instances where a purchase did not occur (True Negatives) and 102 instances where a purchase occurred (True Positives). However, there were 89 False Negatives, where actual purchases were missed, and 33 False Positives, where purchases were incorrectly predicted. These results indicate that while the model performs well in identifying non-purchases, it struggles with predicting purchase outcomes.

Discussion

What Worked Well

- The ensemble learning models, Random Forest and AdaBoost, proved effective in predicting online shopping purchasing intentions. Both models demonstrated strong performance, with accuracies of approximately 90%. Random Forest excelled in precision, indicating its ability to minimize false positive predictions, while AdaBoost performed slightly better in recall and F1 Score, suggesting its strength in identifying actual purchasers.

- The feature importance analysis provided clear insights, highlighting PageValues, ExitRates, and ProductRelated_Duration as the most significant predictors of purchasing behavior.

Challenges and Limitations

Despite the overall success of the models, several challenges were encountered:

1. **Class Imbalance**: The dataset contained more non-purchase cases than purchase cases, which affected the models' ability to predict true positives effectively.
2. **Hyperparameter Tuning Complexity**: Observing the effects of different hyperparameter levels required extensive experimentation, which was computationally intensive and time-consuming.
3. **Model Bias**: Random Forest tended to focus more on precision at the cost of recall, while AdaBoost's boosting mechanism sometimes led to overfitting to misclassified instances.

Addressing the Problem Statement

The results successfully address the problem statement by:

- Identifying the top three features influencing purchasing behavior: PageValues, ExitRates, and ProductRelated_Duration. These findings support the idea that certain browsing behaviors, like spending more time on specific pages or viewing high-value pages, are closely linked to whether a shopper makes a purchase.

- Demonstrating that ensemble models like Random Forest and AdaBoost are well-suited for classification problems, as they effectively handle complex interactions within the dataset.

Conclusion and Future Work

This project successfully applied Random Forest and AdaBoost algorithms to predict online shopping purchasing intentions, achieving high accuracy and providing actionable insights into consumer behavior. Key features such as PageValues, ExitRates, and ProductRelated_Duration were identified as significant predictors of purchase outcomes. The findings suggest that ensemble learning methods are effective for e-commerce analytics, offering a balance between precision and recall that is crucial for business decision-making.

Future Work

To further enhance the findings:

- Addressing class imbalance through techniques such as oversampling, undersampling, or synthetic data generation could improve recall.
- Exploring additional ensemble methods or deep learning models might uncover new patterns and improve predictive accuracy.
- Incorporating temporal or behavioral data could provide deeper insights into shopper behavior, especially for time-sensitive or recurring purchasing patterns.

References

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The Annals of Statistics, 28(2), 337-407. Institute of Mathematical Statistics.

Wyner, A., Olson, M., Bleich, J., & Mease D. (2017). Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers. DOI: https://doi.org/10.48550/arXiv.1504.07676

Probst, P., Wright, M., & Boulesteix, A. (2019). Hyperparameters and Tuning Strategies for Random Forest. DOI: https://doi.org/10.48550/arXiv.1804.03515

Sakar, C., & Kastro, Y. (2018). Online Shoppers Purchasing Intention Dataset. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5F88