BUAN 6341.003 - APPLIED MACHINE LEARNING

PROJECT REPORT

ONLINE SHOPPERS PURCHASE INTENTION

Group 4:
Akshara Athirala
Pravallika Chekka
Sai Pavan Egiteela
Subhash Chandra Gannamraju
Ruthvik Mekala
Yamini Avula

ABSTRACT

Many people who visit e-commerce websites might not plan to buy anything and this could be the result of several factors. However, we can determine if a person is likely to make a purchase or not based on their actions on the e-commerce website. In this project, we will investigate if user behavior on an e-commerce website can be used to anticipate a customer's desire to purchase using Google Analytics data by applying machine learning techniques to create exact prediction models. The ability to predict customer purchase intention can be invaluably helpful to e-commerce businesses as it helps in the better understanding of the digital retail space.

MOTIVATION

Global buying patterns are changing due to the transition from retail to online shopping. E-commerce has already become a major form of retail market. Online customers often browse pages of e-commerce sites before placing orders or abandon browsing without purchase. By offering items that are tailored to each individual customer, this information can assist businesses in better meeting the preferences of their clients and benefiting both parties. As a result, businesses' sales can increase. However, most of the time customers visiting these online websites may not make any purchase at all. This could be due to various factors i.e., the price of the product or window shopping. It is important to predict customers' purchasing intentions so that retention measures like e.g., recommending suitable products can be taken to convert potential customers into purchasers. Currently, the closest existing solution to this problem has been the recommendation system where previous purchase information from a customer is processed to predict the types of products a customer would be interested in. There is evidence to suggest that retention measures such as an apt recommendation system plays an extremely important part in converting sales. Here the prediction of customer purchase intention can help strategize different marketing strategies and could be added to the mechanism of recommendation system of an e-commerce retailer. An example could be that if the ML solution predicts a strong customer purchase intention, then may be the recommend system could recommend a higher quality or a more expensive product as it can be inferred that the user would be willing to consider a better or more expensive product if the intention to purchase a type of product is very strong. If the solution predicts a low purchase intention, then recommendation system could recommend products that are on discount or products with special offers i.e., “Buy one get one free”. Later this historical data of how customer intention changes with such recommendation can also be studied and be applied to improve the recommendation system itself further. However, this report focuses only on the extent of predicting customer purchase intentions.

OBJECTIVE:

The purpose of this project is to make use of user data or browsing history traces that users may leave behind when they visit an online retailer. The research uses session information data to forecast online buyers' purchase intentions with the use of this data. The project's goal is to use this data to build a machine learning model that will forecast customers' intention to make purchases.

SCOPE OF THE PROJECT:

The scope of this project is only limited to predicting customer purchase intentions and evaluating and measuring the accuracy of these predictions.

BACKGROUND KNOWLEDGE:

As mentioned in the introduction, the problem is classified as a Machine Learning problem. Machine learning is a type of computer algorithm that improves automatically through interaction and collecting new data. It is a subset of artificial intelligence. This machine learning model comes under supervised learning, providing us with the essential information required to make data-driven decisions, boost sales, and thrive in the ever-changing e-commerce industry. In the provided dataset, the target variable (or dependent variable) is the "Revenue" attribute. This attribute indicates whether a session resulted in an e-commerce transaction. It serves as the class label for classification tasks. The possible values for the target variable are typically binary, representing whether a transaction was made (e.g., "Yes" or "No"). We would like to build different classification models such as Logistic Regression, Decision trees, K-Nearest Neighbors (KNN) etc. to predict customer purchase intention when he visits the website using metrics such as accuracy, sensitivity, f1 score, etc.

DATA

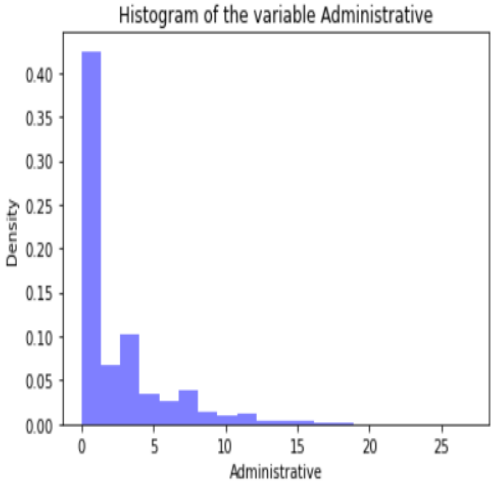
The dataset titled “Online Shoppers Purchasing Intention” from UCI Machine Learning Repository was used for our project. The dataset consists of 18 variables and 12,330 records. Of the 18 variables, 10 are numerical variables, 7 are categorical variables and 1 is Boolean variable.

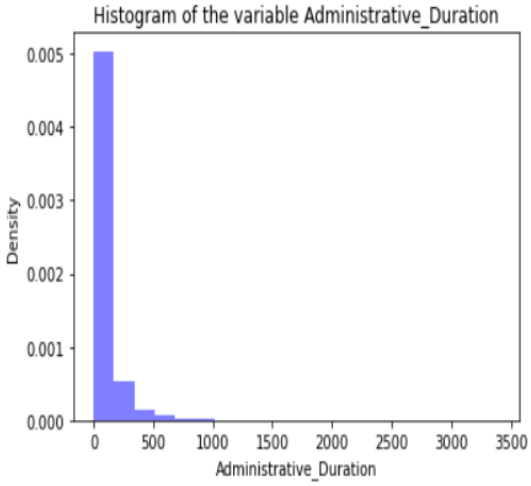
Attribute Name	Data Type	Description
Administrative	Numerical	Number of different types of administrative pages visited by the visitor in the session.
Administrative Duration	Numerical	Total time spent by the visitor on administrative pages in the session.
Informational	Numerical	Number of different types of informational pages visited by the visitor in the session.
Informational Duration	Numerical	Total time spent by the visitor on informational pages in the session.
Product Related	Numerical	Number of different types of product-related pages visited by the visitor in the session.
Product Related Duration	Numerical	Total time spent by the visitor on product-related pages in the session.
Bounce Rate	Numerical	The percentage of visitors who enter the site from a specific page and then leave without further interaction.
Exit Rate	Numerical	The percentage of pageviews that were the last in a session for a specific page.
Page Value	Numerical	The average value of a web page that a user visited before completing an e-commerce transaction.
Special Day	Numerical	Indicates the proximity of the site visiting time to a specific special day likely to result in transactions.
Operating System	Categorical	The operating system used by the visitor.
Browser	Categorical	The web browser used by the visitor.
Region	Categorical	The geographical region or location of the visitor.
Traffic Type	Categorical	The type of traffic source that brought the visitor to the site.
Visitor Type	Categorical	Indicates whether the visitor is a returning or new visitor.
Weekend	Boolean	Indicates whether the date of the visit falls on a weekend.
Month	Categorical	The month in which the visit occurred.
Revenue	Categorical	Indicates whether the session resulted in an e-commerce transaction.

EXPLORATORY DATA ANALYSIS

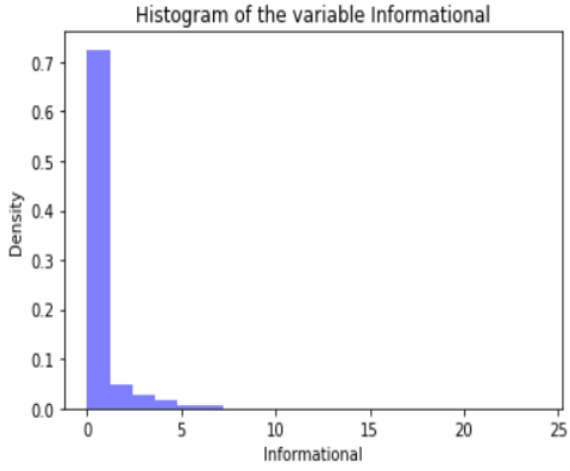
This section delves into the comprehensive exploration of our dataset, offering a detailed examination of the variables, relationships, and potential patterns that may influence online shoppers' purchase decisions. Through visualizations, summary statistics, and correlation analyses, our exploratory data analysis aims to unveil the hidden patterns that underlie the observed data. By identifying outliers, understanding the distribution of key features, and discerning any notable trends, we lay the groundwork for better modeling and decision-making in subsequent stages of the project.

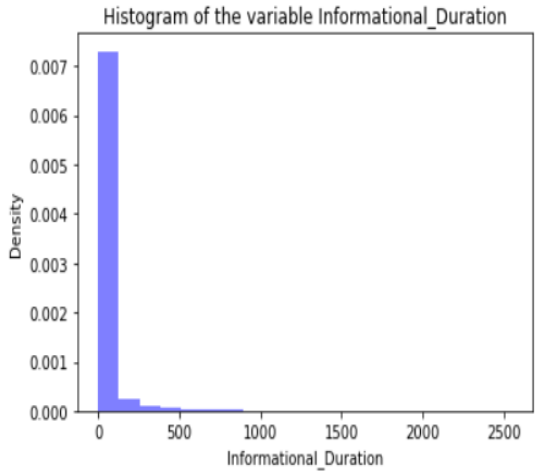
The following graphs shows the distribution of number of pages visited by visitors about account management and Website, communication and address information of shopping site and the time spent on them.





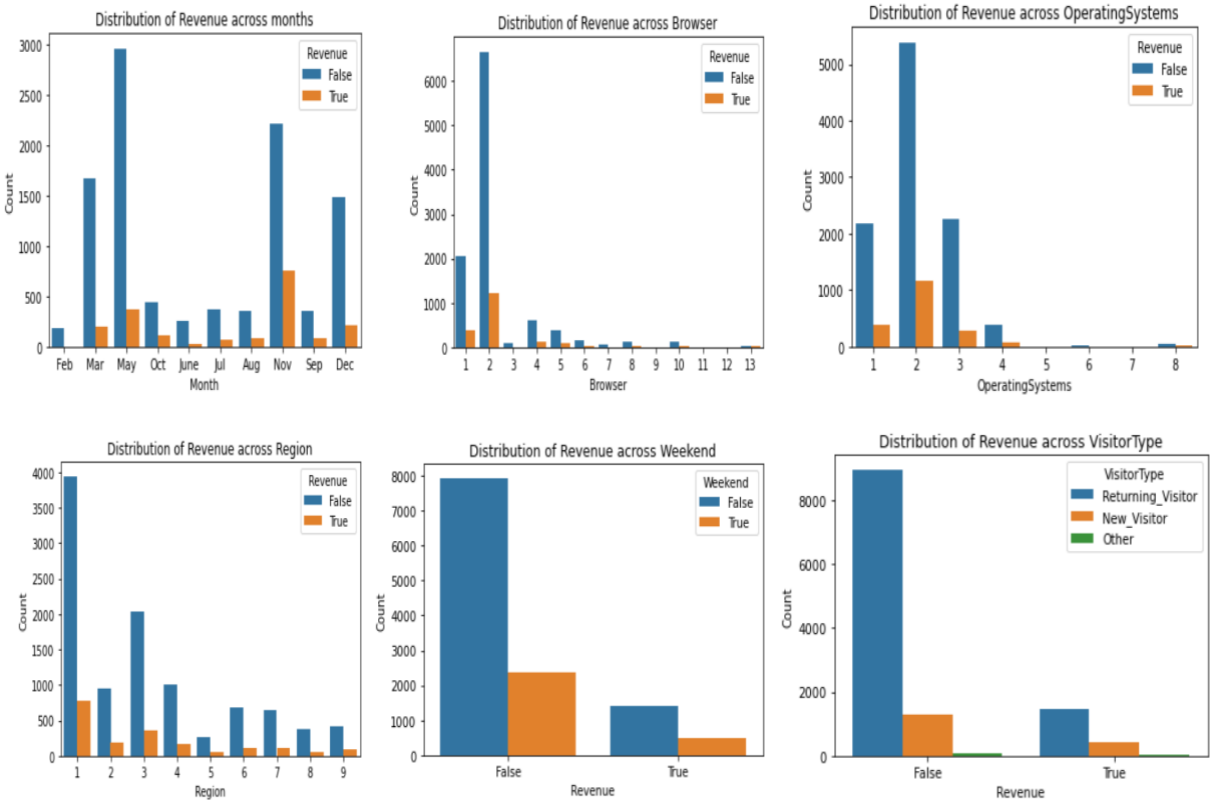
According to the histogram, 45% of visitors only view one or two account management-related pages. Interestingly, most of these visitors stay on each page for more than 0 to 120 seconds (about 2 minutes), which suggests a quick but significant interaction with the account management information.





Regarding pages that provide communication, address, and website information, the histogram indicates that 70 percent of users only look through one or two pages. The predominant tendency is that visitors spend more than 0 to 120 seconds (about 2 minutes) on each of these pages, much like they do with the account administration pages.

Revenue Analysis



Monthly Transaction Trends: The analysis of transaction data indicates that visitors made the highest number of transactions in the month of November. This suggests a noteworthy surge in user activity during that specific month.

Browser Contribution to Revenue: Browser analysis reveals that Browser Number 2 has significantly contributed to the highest revenue. This could imply that visitors using Browser 2 engage in transactions that contribute more to the overall revenue.

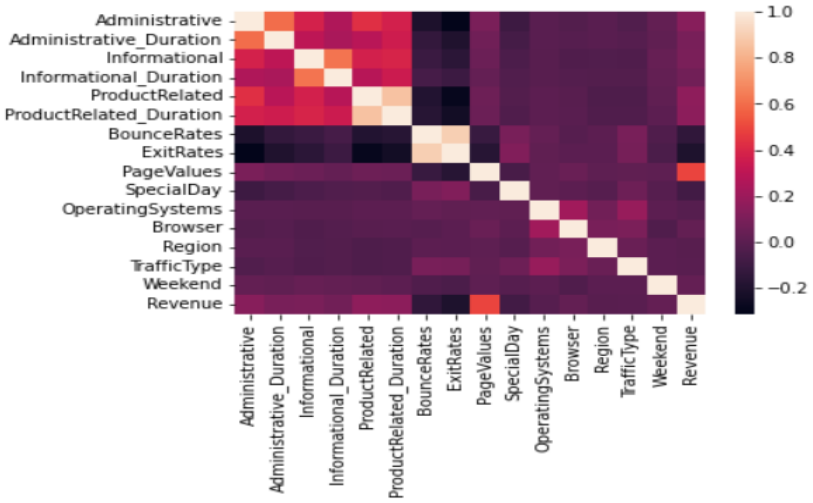
Prevalent Operating System for Transactions: Operating System 2 emerged as the predominant choice among visitors, with this OS being utilized for the majority of transactions. This highlights the importance of optimizing the platform and user experience for Operating System 2.

Regional Revenue Distribution: Geographically, Region 1 stands out as the top contributor to revenue, indicating that this specific region plays a pivotal role in generating transactions and income for the business.

Weekday vs. Weekend Revenue Analysis: The analysis of revenue patterns across the days of the week suggests that weekdays have a substantial impact on revenue generation. This underscores the importance of understanding and catering to user behavior variations during weekdays compared to weekends.

Visitor Type and Transaction Volume: Among different visitor types, returning visitors exhibited the highest transaction volume. This insight is crucial for tailoring marketing strategies or user experiences to encourage repeat visits and transactions.

CORRELATION MATRIX



The correlation matrix shows that the following variables are highly correlated with each other:

Administrative and Administrative_Duration

Informational and Informational_Duration

ProductRelated and ProductRelated_Duration

This suggests that websites with higher levels of administrative traffic, informational traffic, or product-related traffic are also more likely to have longer average session durations.

PREPROCESSING

We did the following preprocessing steps to our data set to improve our dataset for better modelling:

- **Null values check**: We checked for null values in the data and did not find any null values in the data.
- **Removal of Duplicate Entries**: We found duplicate entries in the dataset and eliminated them methodically. By ensuring that every row is distinct, this phase guards against biases or errors that can arise in subsequent studies.
- **Data Type Modification**: Certain variables data types were carefully examined and changed to more suitable ones, like float or integers.

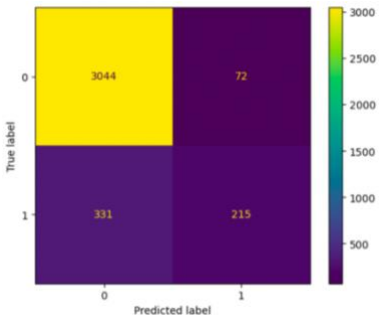
- **Scaling for Standardization of Magnitude**: Scaling was applied to variables with varying magnitudes, such as administrative and product-related durations. Standardizing these variables to a single scale is important in order to avoid modelling problems, guarantee fair comparisons, and preserve model correctness.
- **Establishing Dummy Variables for Classification Features:** Dummy variables were created for certain variables (Month, Visitor Type, Region). This makes it easier to incorporate categorical data.

MODELS & PERFORMANCE EVALUATION:

Logistic Regression

Logistic regression is a statistical modeling technique used for binary classification. It predicts the probability of an event occurring given a set of independent variables. It outputs a probability between 0 and 1. Since we are trying to predict if a purchase occurs or not, logistic regression is an appropriate choice.

The model score is 87.91% and the accuracy score is 88.99%, indicating it can accurately classify a significant portion of the data.



The confusion matrix shows the number of correctly and incorrectly predicted instances for a classification model.

- **Accuracy**: Proportion of correctly classified instances. Overall accuracy is 0.89, which suggests the model performs well on the dataset.
- **Precision**: Measures the fraction of instances predicted as class X that are truly class X. High precision for class 0 (0.90) indicates the model rarely misclassifies other instances as class 0. Lower precision for class 1 (0.75) suggests some instances from other classes are mistakenly predicted as class 1
- **F1-score**: Harmonic mean of precision and recall, providing a balanced assessment. High F1-score for class 0 (0.94) indicates the model performs well overall for this class. Lower

F1-score for class 1 (0.52) suggests the model's performance on this class requires improvement.

- **Recall**: Measures the fraction of actual class X instances that are correctly predicted as class X. High recall for class 0 (0.98) shows the model effectively identifies most instances of class 0. Low recall for class 1 (0.39) indicates the model misses a significant portion of actual instances belonging to class 1.

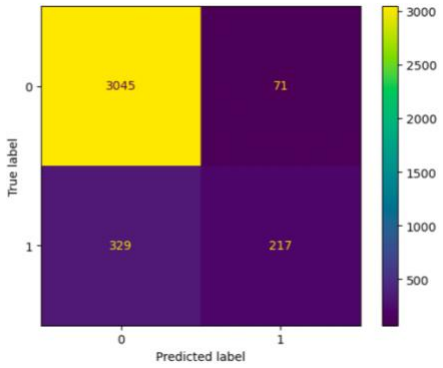
Overall, the logistic regression model performed well on the test set. However, there is room for improvement in the recall of the model for the negative class. Future work could focus on improving the recall of the model for the negative class without sacrificing the precision of the model for the positive class.

Lasso Classifier

The Lasso classifier, regarded as Logistic Regression with L1 regularization, functions by establishing a linear model and introducing a penalty mechanism that effectively shrinks the coefficients. It shrinks coefficients towards zero, effectively removing features with weak contributions. This leads to a more interpretable model by highlighting the most important features influencing the target variable. By penalizing non-zero coefficients, Lasso reduces model complexity and prevents overfitting, especially on datasets with high dimensionality or collinearity issues.

Lasso acts as a regularizer, adding stability to the model and preventing excessive sensitivity to noise or outliers in the data.

The model score is 89.08%, indicating it can accurately classify a significant portion of the data.



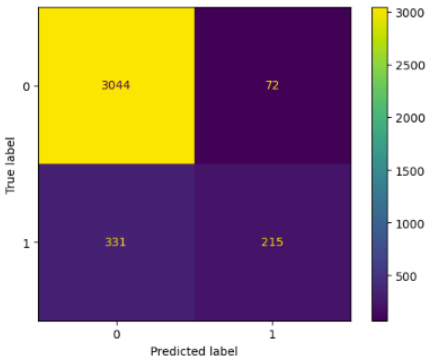
- The AUC score of 0.6873 suggests a moderate ability to distinguish between the positive and negative classes.
- The confusion matrix shows that the model correctly identifies most instances of class 0 (majority class) with high precision (0.90) and recall (0.98).
- However, the model struggles with class 1 (minority class), exhibiting lower precision (0.75) and recall (0.40).
- This implies the model might be biased towards the majority class due to the significant class imbalance.

The Lasso Regression model demonstrates good overall accuracy in classification. However, further analysis and tuning are necessary to improve its performance on the minority class and address potential bias introduced by class imbalance. By investigating feature importance and optimizing the regularization parameter, the model's effectiveness and interpretability can be enhanced.

Ridge classifier

The Ridge classifier, also known as Logistic Regression with L2 regularization, fits a linear model to the data, and introduces a regularization term based on the squared magnitudes of the coefficients, mitigating the risk of overfitting by penalizing larger coefficients. This regularization not only contributes to the model's generalization performance but also provides stability and resilience in the face of multicollinearity among the input features.

The model score is 89% and the AUC score is 0.6853.



- Precision: 0.75 - Moderate precision suggests some misclassification of other instances as class 1.

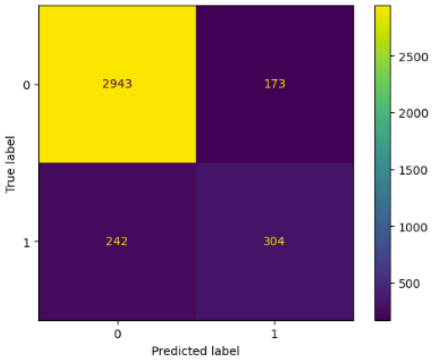
- Recall: 0.39 - Low recall reveals the model misses a significant portion of actual instances belonging to class 1.
- F1-score: 0.52 - Low score confirms the model's performance is weaker for this class.

The Ridge classifier exhibits a more balanced performance, maintaining high precision and recall for non-purchase instances, while delivering a relatively balanced trade-off for purchase instances. On the other hand, the Lasso classifier excels in capturing non-purchase instances with a remarkable recall but struggles with precision and recall balance for purchase instances. If achieving a more balanced predictive performance across both classes is crucial, the Ridge classifier may be preferred. However, if accurately identifying instances of non-purchase is a top priority, the Lasso classifier may be favored, albeit at the expense of a less balanced performance for purchase instances.

KNN Classifier

The k-Nearest Neighbors (KNN) classifier is versatile and intuitive for classification tasks. KNN classifies data points by considering the majority class of their k-nearest neighbors in the feature space. It adapts dynamically to the underlying patterns in the data, making it suitable for scenarios with complex decision boundaries.

The model score is 88.67% and AUC score is 0.7506.



- Precision: Class 0: 0.92 - The model correctly identifies 92% of class 0 instances as class 0. Class 1: 0.64 - The model correctly identifies 64% of class 1 instances as class 1.
- Recall: Class 0: 0.94 - The model identifies 94% of all actual class 0 instances. Class 1: 0.56 - The model identifies 56% of all actual class 1 instances.

- F1-score: Class 0: 0.93 - This balanced measure confirms the model performs well for class 0. Class 1: 0.59 - This lower score indicates the model's performance needs improvement for class 1.

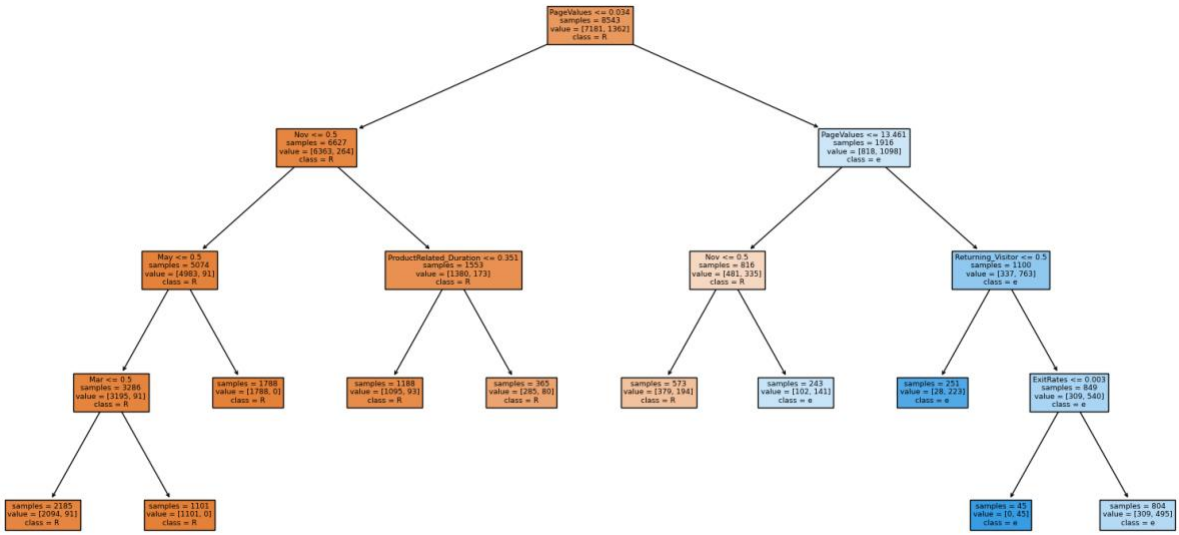
The model performs well on the majority class (class 0) with high precision and recall. However, the performance of the minority class (class 1) is weaker, with lower precision and recall. This suggests the model might be biased towards the majority class due to the significant class imbalance.

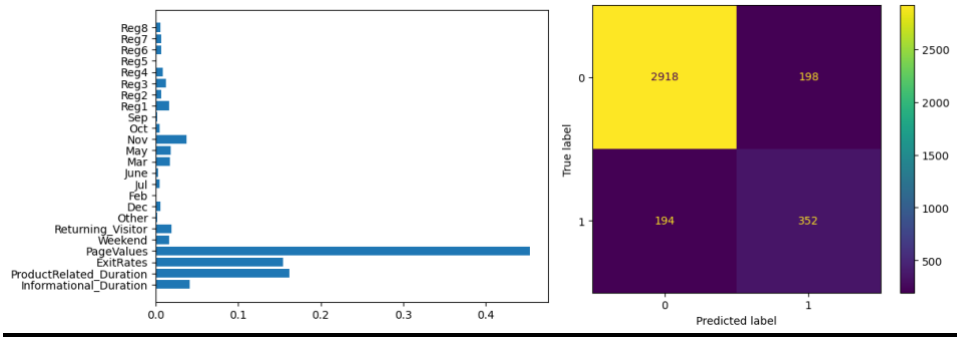
Decision Tree

The Decision Tree algorithm is employed for both classification and regression tasks. It breaks down complex decision scenarios into a series of binary choices based on input features. Each decision node represents a feature, and the branches emanating from it signify the possible outcomes. It is valuable for understanding the underlying patterns in data.

We used both entropy method and gini method to calculate each decision. The score for full model without tuning is 86.61% for entropy method and 85.77% for gini method.

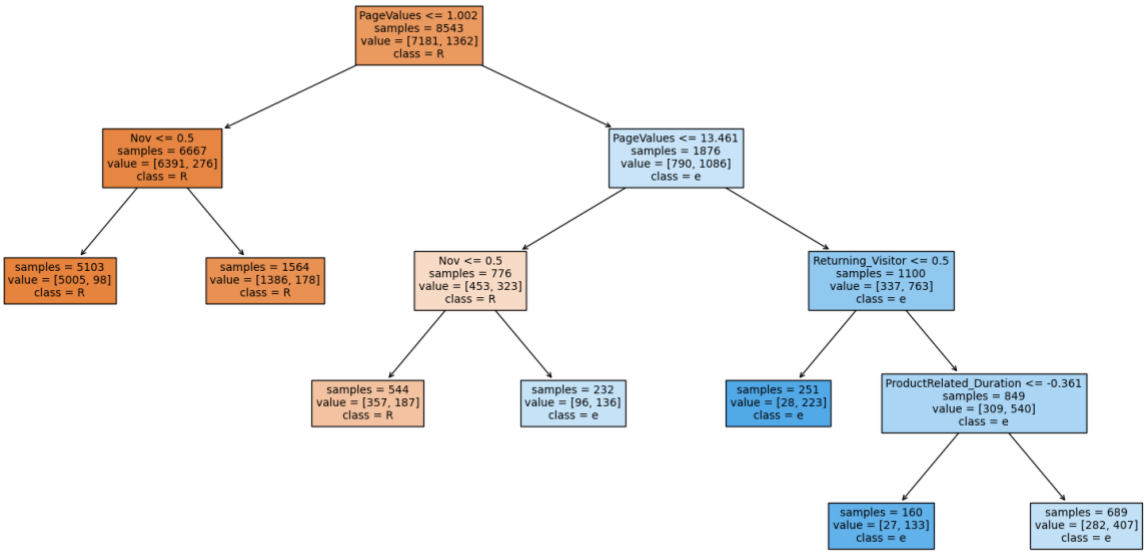
We tuned the model to find out the best parameters.
The best parameters for entropy method is {'max_depth': 4, 'max_leaf_nodes': 10, 'min_samples_split': 2}.
The best parameters for gini method is {'max_depth': 4, 'max_leaf_nodes': 7, 'min_samples_split': 2}.
(Note – the criterion specified while running the model was incorrect, the minimum sample split should have been higher.)
The score for entropy method is 89.29% and for gini method is 89.37%.
DT based on entropy:

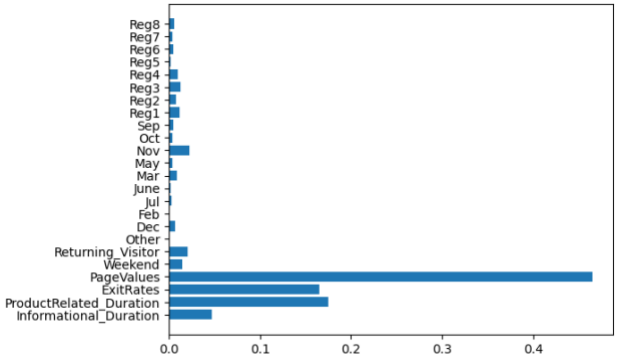


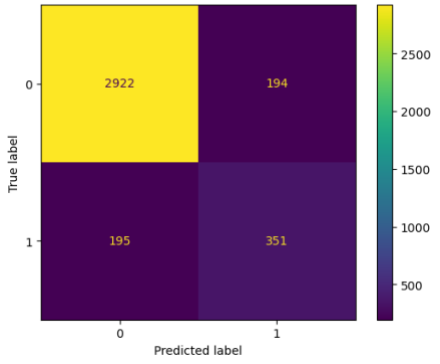


- Precision (for class 0): 94% — This indicates that 94% of instances predicted as belonging to class 0 (e.g., non-purchase) were indeed correct, reflecting a high level of accuracy in classifying non-purchase instances.
- Precision (for class 1): 64% — This means that 64% of instances predicted as belonging to class 1 were correct, reflecting a moderate level of accuracy in classifying purchase instances.
- Recall (for class 0): 94% — This suggests that the model correctly identified 94% of all actual instances belonging to class 0, indicating a strong ability to capture non-purchase instances.
- Recall (for class 1): 64% — The model correctly identified 64% of all actual instances belonging to class 1, indicating a moderate ability to capture purchase instances.
- F1-score (for class 0): 94% — A score of 94% suggests a robust overall performance for class 0, considering both precision and recall.
- F1-score (for class 1): 64% — A score of 64% suggests a moderate overall performance, considering both precision and recall.

DT based on gini:







- Precision (for class 0): 94% — This indicates that 94% of instances predicted as belonging to class 0 (e.g., non-purchase) were indeed correct, reflecting a high level of accuracy in classifying non-purchase instances.
- Precision (for class 1): 64% — This means that 64% of instances predicted as belonging to class 1 were correct, reflecting a moderate level of accuracy in classifying purchase instances.
- Recall (for class 0): 94% — This suggests that the model correctly identified 94% of all actual instances belonging to class 0, indicating a strong ability to capture non-purchase instances.
- Recall (for class 1): 64% — The model correctly identified 64% of all actual instances belonging to class 1, indicating a moderate ability to capture purchase instances.
- F1-score (for class 0): 94% — A score of 94% suggests a robust overall performance for class 0, considering both precision and recall.

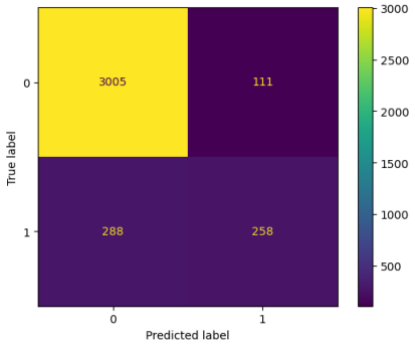
- F1-score (for class 1): 64% — A score of 64% suggests a moderate overall performance, considering both precision and recall.

Page Values is the most importance feature for decision making in both the methods.

Since score of gini method is slightly better and all other scores are same as entropy method, we can go ahead with DT using gini method for our project.

Linear SVC

Linear SVC is a versatile and efficient classifier for binary classification tasks. By strategically positioning a hyperplane to separate different classes, the algorithm maximizes the margin between instances, enhancing its generalization ability. Its simplicity, interpretability, and effectiveness in dealing with high-dimensional data make it a valuable tool for various machine learning applications.



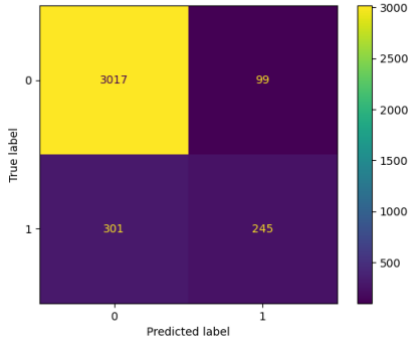
- Precision: Class 0: 0.91 - The model correctly identifies 91% of class 0 instances as class 0. Class 1: 0.70 - The model correctly identifies 70% of class 1 instances as class 1.
- Recall: Class 0: 0.96 - The model identifies 96% of all actual class 0 instances. Class 1: 0.47 - The model identifies 47% of all actual class 1 instances.
- F1-score: Class 0: 0.94 - This balanced measure confirms the model performs well for class 0. Class 1: 0.56 - This lower score indicates the model's performance needs improvement for class 1.

Linear SVC demonstrates good overall accuracy and a well-balanced performance across classes. However, further investigation and optimization can be beneficial to improve performance on the minority class and address any potential bias introduced by class imbalance.

Linear SVC with GridSearchCV

GridSearchCV is a hyperparameter tuning technique that systematically searches for the best combination of parameters for a machine learning model. In this case, GridSearchCV was used to find the optimal values for the C parameter of the Linear SVC model.

The model score is 89.07% and C = 1 is the best parameter for the given dataset.



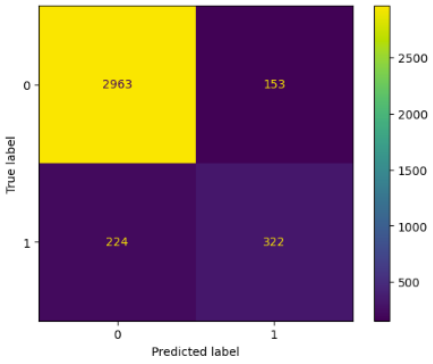
- Precision: Class 0: 0.91 - The model correctly identifies 91% of class 0 instances as class 0. Class 1: 0.71 - The model correctly identifies 71% of class 1 instances as class 1.
- Recall: Class 0: 0.97 - The model identifies 97% of all actual class 0 instances. Class 1: 0.45 - The model identifies 45% of all actual class 1 instances.
- F1-score: Class 0: 0.94 - This balanced measure confirms the model performs well for class 0. Class 1: 0.55 - This lower score indicates the model's performance needs improvement for class 1.

GridSearchCV effectively optimized the Linear SVC model, leading to a slight improvement in overall performance. While further adjustments might be necessary to address class imbalance and improve performance on the minority class, the optimized model demonstrates good accuracy and generalizability for this binary classification task.

SVM Kernel Functions

Kernel functions play a crucial role in SVM models by transforming the input data into a higher-dimensional space where the data is linearly separable. This allows the SVM to find an optimal hyperplane for classification in the original space even if the data is non-linearly separable.

The model score is 89.70% and the best parameters are {'C': 10000, 'gamma': 0.001}.



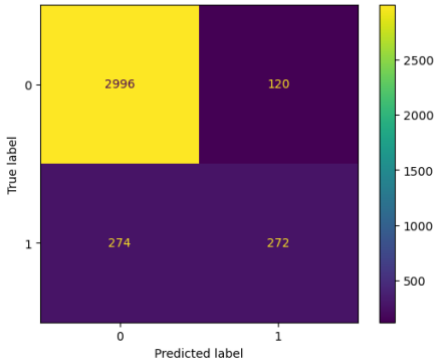
- Precision:Class 0: 0.93 - The model correctly identifies 93% of class 0 instances as class 0. Class 1: 0.68 - The model correctly identifies 68% of class 1 instances as class 1.
- Recall: Class 0: 0.95 - The model identifies 95% of all actual class 0 instances. Class 1: 0.59 - The model identifies 59% of all actual class 1 instances.
- F1-score: Class 0: 0.94 - This balanced measure confirms the model performs well for class 0. Class 1: 0.63 - This lower score indicates the model's performance needs some improvement for class 1.

The SVM with kernel functions demonstrates good overall accuracy and a well-balanced performance across classes. The use of a kernel function allows the model to handle the non-linearity in the data and achieve better performance compared to the Linear SVC model. However, it's important to consider the potential trade-offs in terms of computational cost and tuning complexity.

Voting Classifier

The Voting Classifier, a powerful ensemble learning technique, combines the predictions of multiple individual classifiers to produce a more robust and accurate final prediction. This collaborative approach leverages on the strengths of each to collectively enhance overall performance. The Voting Classifier operates through a majority voting mechanism, where each classifier "votes" for a particular class, and the class with the most votes is selected as the final prediction.

The model score is 89.24%



- Precision for Class 0 (non-purchase): The precision for class 0 is 92% indicates that 92% of instances predicted as non-purchase were accurate.
- Precision for Class 1 (purchase): The precision for class 1 is 69%. This implies that 69% of instances predicted as purchase were accurate.
- Recall for Class 0 (non-purchase): The recall for class 0 is 96%. This suggests that the model successfully captured 96% of all actual non-purchase instances.
- Recall for Class 1 (purchase): The recall for class 1 is 50%. This indicates that the model successfully identified 50% of all actual purchase instances.
- F1-Score for Class 0 (non-purchase): The F1-score for class 0 is 94%.
- F1-Score for Class 1 (purchase): The F1-score for class 1 is 58%.

In summary, the model demonstrates strong performance in correctly classifying instances of non-purchase (Class 0), with high precision and recall. However, its performance on predicting instances of purchase (Class 1) is characterized by a moderate level of accuracy, as reflected in the lower precision and recall values.

Bagging Classifier

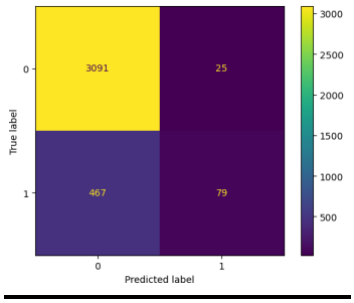
The model score is 88.83% for bagging with Linear SVC.

The model score is 89.2% for bagging with Logistic regression.

The model score is 88.6% for bagging with Decision Tree.

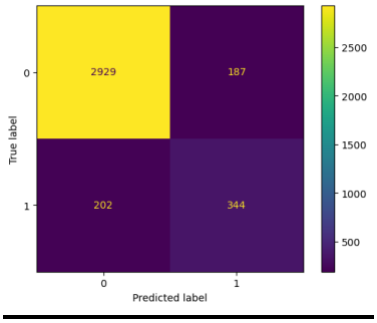
Random Forest Classifier

The model score is 86.56%



Adaboost Classifier

The model score is 89.37%



OVERALL SUMMARY

Performance ranking of each Algorithm on Prediction Customer Purchase Intention

Highest Accuracy achieved: 89.7% using **SVM with Kernel functions**

Machine Learning Algorithm	Accuracy
Logistic Regression	0.889
Lasso	0.890
Ridge	0.890
KNN Classifier	0.885
Decision Tree - Entropy	0.892
Decision Tree - Gini Index	0.893
Linear SVC	0.892
Linear SVC with gridsearch	0.890

SVM with Kernel functions	0.897
Voting Classifier	0.892
Bagging classifier (Decision tree)	0.892
Bagging classifier (Logistic Regression)	0.886
Bagging classifier (Linear SVC)	0.888
Random Forest classifier	0.865
Adaboost	0.893

Classification model (SVM with kernel functions model, Decision Tree, Adaboost) give better results for our data set. From the classification models used, SVM with kernel functions model gives result with higher accuracy.

CONCLUSION

The best machine learning algorithm for a particular task will depend on the specific data set and the desired outcome. The project aim was to build a solution that can predict customer purchase intention with as high an accuracy as possible. The highest accuracy it has managed to achieve was 89.7% accuracy. It was recorded as the highest accuracy by comparing and ranking the various model's performances with each other. It can be seen that SVM with Kernel functions performed the best among the various algorithms implemented.

INSIGHTS

- Predicting customer purchase intention: This project has the potential to provide valuable insights into user behavior and purchase patterns on e-commerce websites. By analyzing session data, it can reveal trends and indicators that differentiate between browsing and purchasing customers. This information can be leveraged to:
- Improve marketing campaigns: Tailor marketing messages and offers to individual customers based on their predicted purchase intention.

- Personalize product recommendations: Recommend products that are most likely to appeal to individual customers based on their browsing history and predicted purchase intention.
- Optimize website design and user experience: Identify website features and functionalities that influence purchase intention and optimize the user experience accordingly.
- Reduce cart abandonment: Identify potential roadblocks in the purchase journey and implement measures to reduce cart abandonment.
- Increase sales and conversion rates: By understanding and influencing customer behavior, businesses can increase sales and conversion rates.

LIMITATIONS

- Data quality and availability: The accuracy and effectiveness of the model will depend heavily on the quality and availability of data. Session data might be incomplete or inaccurate, impacting the model's performance.
- Model bias: Machine learning models can be susceptible to bias if the training data is biased. This can lead to inaccurate predictions for certain demographics or customer segments.
- Limited scope: This project only focuses on predicting purchase intention based on session data. Other factors influencing purchase decisions, such as external economic conditions or individual preferences, are not considered.
- Model interpretability: Some machine learning models, especially complex ones, can be difficult to interpret. This can make it challenging to understand why the model makes certain predictions and limit its practical application.
- Ethical considerations: Collecting and analyzing user data raises ethical concerns regarding privacy and data security. Businesses need to ensure they comply with data privacy regulations and use user data responsibly.