PROJECT REPORT
DATE : 09-11-2023
TEAM ID : Team-592521

ML PROJECT

PROJECT : ONLINE SHOPPERS INTENTION PREDICTION

TEAM MEMBERS:
21BRS1682 - N.ANIRUDDHAN
21BCE1124 - V.SHASHAANK
21BAI1535 - S ADITHYA VARDHAN

MENTOR : Saumya Mohandas

CONTENTS

INTRODUCTION

Project Overview:

The aim of this project is to predict customer behavior in online shopping, specifically whether a customer will make a purchase or engage in window shopping. By leveraging classification algorithms, including Logistic Regression, Random Forest, and K-Means clustering, the project seeks to develop a model that can accurately predict customer behavior.

Key Steps:

Data Collection: Gather relevant data on customer behavior in online shopping. This data may include customer demographics, browsing history, past purchase records, and other relevant features.

Data Preprocessing: Clean and preprocess the collected data to ensure its quality and suitability for analysis. This step may involve handling missing values, encoding categorical variables, and normalizing numerical data.

Feature Selection: Identify the most relevant features that contribute to predicting customer behavior. This process may involve techniques such as statistical analysis, correlation analysis, or feature importance ranking.

Model Training: Utilize the selected classification algorithms (Logistic Regression, Random Forest, and K-Means clustering) to train the prediction model. This involves splitting the data into training and testing sets and fitting the algorithms to the training data.

Model Evaluation: Evaluate the performance of each algorithm using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. Compare the results to determine the best-performing algorithm.

Model Selection and Saving: Select the best-performing model based on evaluation results and save it in pickle (.pkl) format for future use.

Model Deployment: Deploy the selected model in a suitable environment or integrate it into an existing system for real-time prediction of customer behavior in online shopping.

By following these steps, the project aims to provide businesses with valuable insights into customer behavior, enabling them to make informed decisions regarding marketing strategies, personalized recommendations, and improving overall customer satisfaction in the online shopping domain.

Online Shoppers Dataset

Purpose:

Online shopping is the activity or action of buying products or services over the Internet. It means going online, landing on a seller's website, selecting something, and arranging for its delivery. The buyer either pays for the good or service online with a credit or debit card or upon delivery. The term does not only include buying things online but also searching for them online.

LITERATURE SURVEY

Existing problem:

In the field of online shoppers prediction using machine learning (ML) techniques, there are several existing challenges and problems that researchers and practitioners encounter. Some of the key problems include:

1. Imbalanced Data: Imbalance in the dataset, where the number of positive (purchasing) instances is significantly lower than negative (non-purchasing) instances, can lead to biased predictions. This can result in models favoring the majority class and performing poorly in accurately predicting the minority class.

2. Feature Engineering: Selecting and engineering relevant features from the available data is crucial for accurate prediction. However, identifying the most informative features and transforming them appropriately can be challenging, especially when dealing with high-dimensional and heterogeneous data.

3. Model Overfitting: Overfitting occurs when a model learns the training data too well, resulting in poor generalization to new, unseen data. This can happen when models are too complex or when there is insufficient data to train the model effectively.

4. Model Interpretability: Many ML algorithms, such as Random Forest or deep learning models, are considered black-box models, making it difficult to interpret their decision-making process. This lack of interpretability can hinder understanding of the factors influencing predictions and limit trust in the model's results.

5. Handling Dynamic Behavior: Online shopping behavior is dynamic and can change over time due to various factors like seasonality, trends, or external events. Developing models that can adapt to these changes and provide accurate predictions in real-time is a significant challenge.

6. Data Privacy and Security: Online shopping involves sensitive customer information, such as personal details and payment data. Ensuring data privacy and security while collecting, storing, and processing this information is critical but can be complex due to regulatory requirements and potential security breaches.

Customer behavior in online shopping: Investigating the factors that influence customer decision-making in online shopping, such as product reviews, website design, pricing strategies, and personalization techniques.

Classification algorithms for customer prediction: Exploring the application of classification algorithms like Logistic Regression, Random Forest, and K-Means clustering in predicting customer behavior in online shopping. Understanding their strengths, limitations, and performance in different scenarios.

Feature selection and model evaluation: Examining different approaches to selecting relevant features for prediction models and evaluating their effectiveness in accurately predicting customer behavior. This may involve techniques like feature engineering, dimensionality reduction, and performance metrics.

Model deployment and practical implications: Exploring how the selected prediction model can be implemented in real-world scenarios and the potential benefits for businesses in terms of targeted marketing strategies, personalized recommendations, and improved customer satisfaction.

References:

Kaggle
Flask
Supervised Learning

UnsupervIsed LearnIng

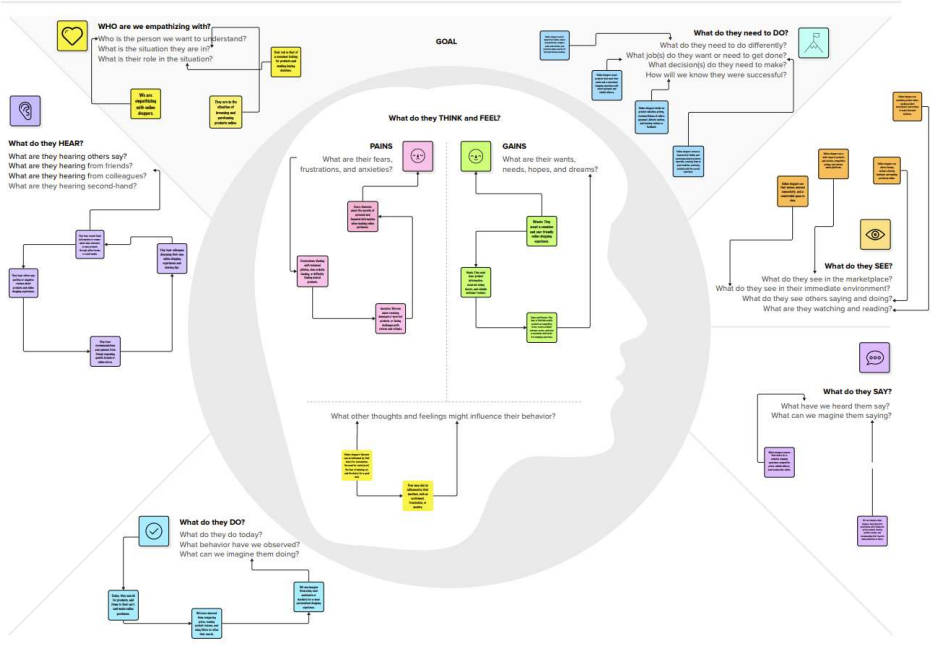
Problem Statement Definition :

Online shopping is the activity or action of buying products or services over the Internet. It means going online, landing on a seller's website, selecting something, and arranging for its delivery. The buyer either pays for the good or service online with a credit or debit card or upon delivery. The term does not only include buying things online but also searching for them online. In other words, I may have been engaged in online shopping but did not buy anything.

We are going to predict whether the customer will buy the product or just go window shopping . Here, We will be using classification algorithms such as Logistic Regression , Random forest, & Clustering algorithm K-Means.We will train and test the data with these algorithms. From this the best model is selected and saved in pkl format.

IDEATION & PROPOSED SOLUTION

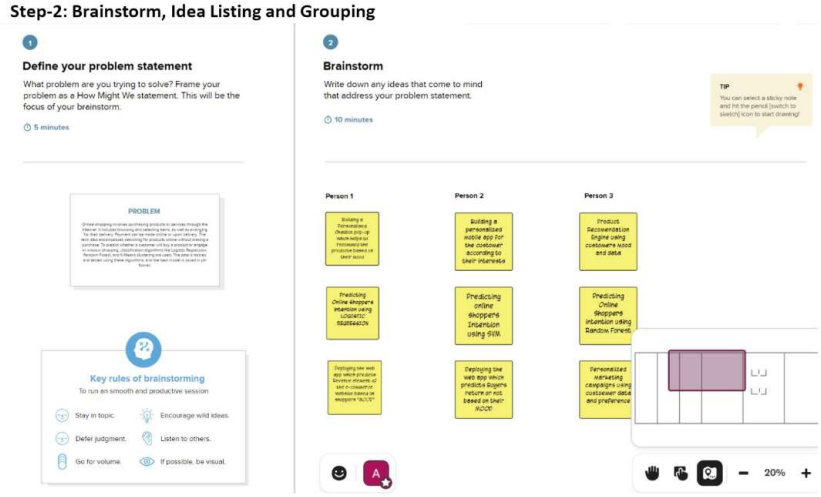
Empathy Map Canvas:



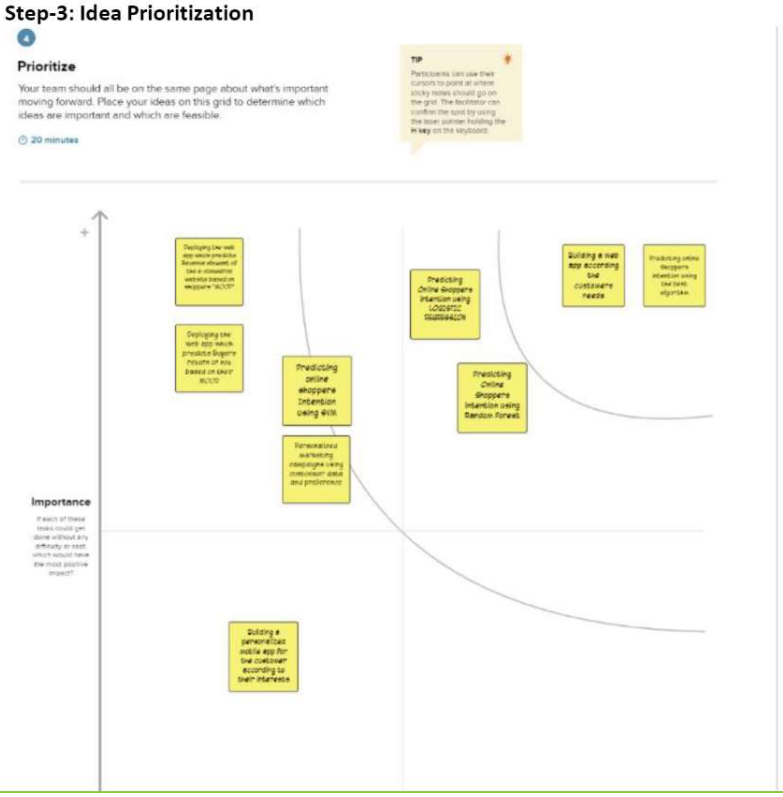
Empathy Canvas Map

Ideation & Brain storming:

Step-2: Brainstorm, Idea Listing and Grouping



Step-3: Idea Prioritization



Ideation and Brainstorming

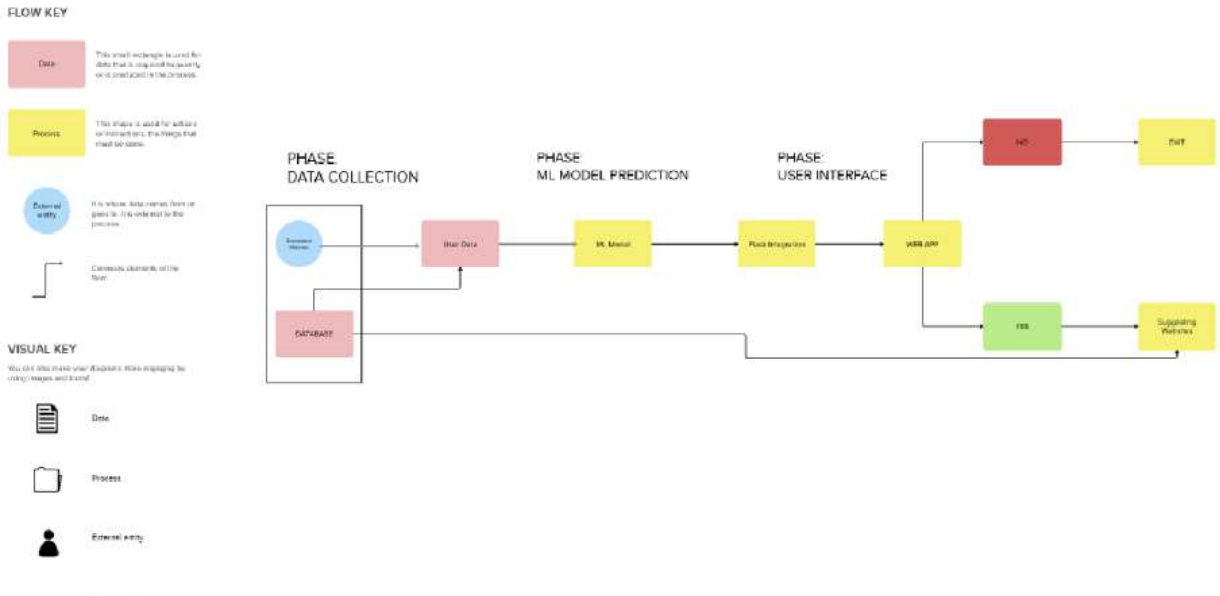
REQUIREMENT ANALYSIS

Functional requirement:

Functional Requirement (Epic)
Registration
Login
Dashboard

PROJECT DESIGN

Data Flow Diagrams & User Stories:



User Stories

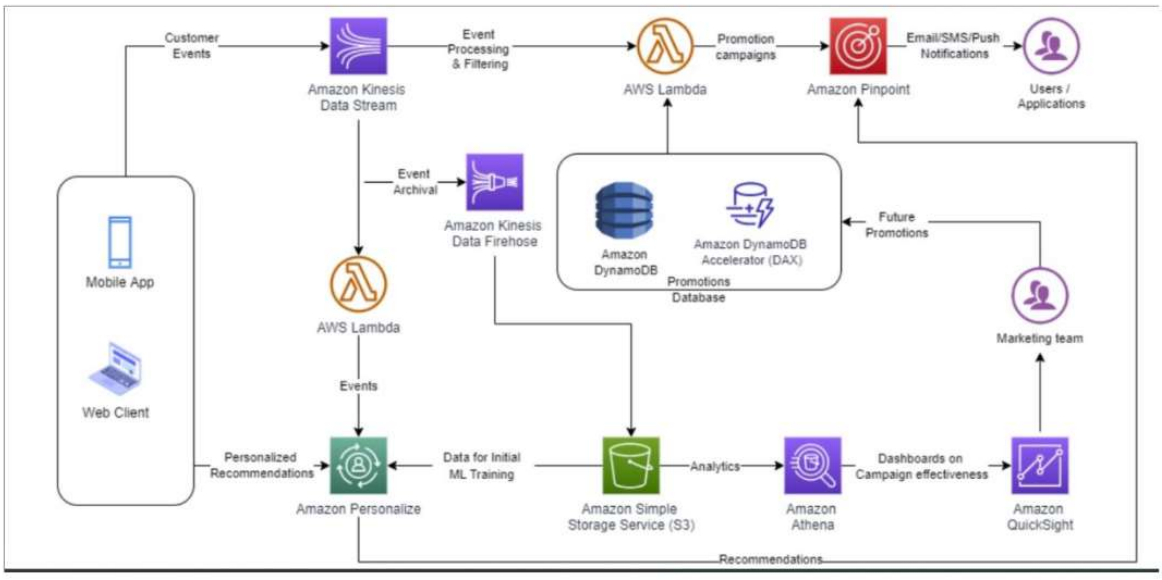
Use the below template to list all the user stories for the product.

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook	I can register & access the dashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail		Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password		High	Sprint-1
	Dashboard					
Customer (Web user)						
Customer Care Executive						
Administrator						

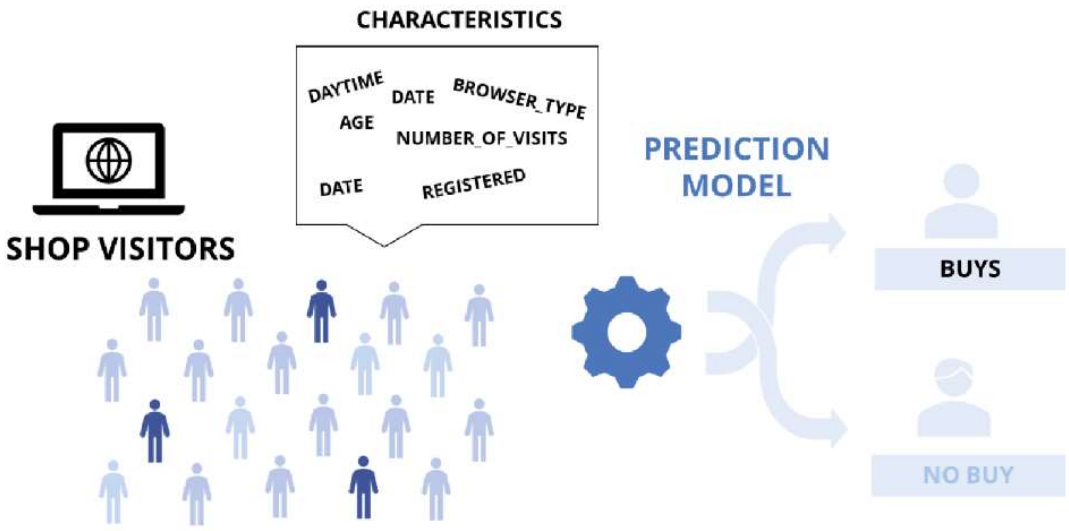
Data flow diagrams and User stories

Solution Architecture:

Example - Solution Architecture Diagram:



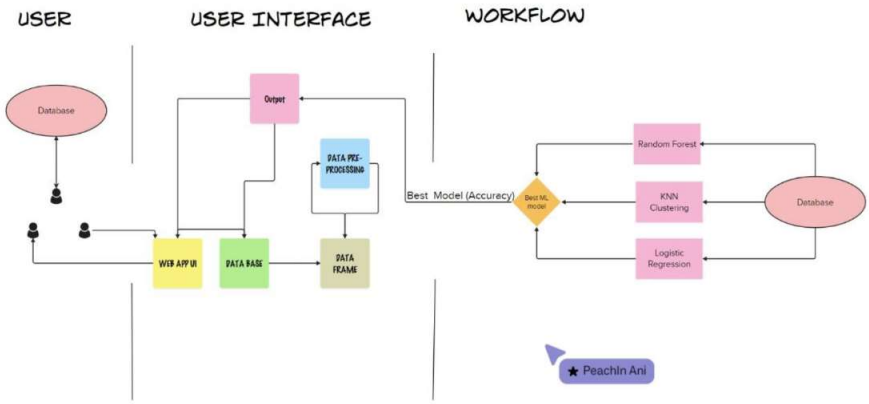
SIMPLIFIED PROTOTYPE OF PREDICTION:



Solution Architecture

PROJECT PLANNING & SCHEDULING

Technical Architecture:



Technical Architecture

Sprint Planning & Estimation, Sprint Delivery Schedule:

Sprint Planning

CODING & SOLUTIONING

Feature 1: Logistic Regression

```
def logreg(X_train,X_test,Y_train,Y_test):
  lr=LogisticRegression()
  lr.fit(X_train,Y_train)
  ypred=lr.predict(X_test)
  print('*******Logistic Regression*******')
  print(' Confusion Matrix ')
  print(confusion_matrix(Y_test,ypred))
  print(' Classification Report ')
  print(classification_report(Y_test,ypred))
```

Feature 2: Random Forest

```
def randfor(X_train,X_test,Y_train,Y_test):
  rf=RandomForestClassifier()
  rf.fit(X_train,Y_train)
  ypred=rf.predict(X_test)
  print('*******Random Forest Classifier*******')
  print(' Confusion Matrix ')
  print(confusion_matrix(Y_test,ypred))
  print(' Classification Report ')
  print(classification_report(Y_test,ypred))
```

PERFORMANCE TESTING

Performace Metrics:

```
def compare_model(x_train,x_test,y_train,y_test):
  logreg(x_train,x_test,y_train,y_test)
  print('-'*100)
  randfor(x_train,x_test,y_train,y_test)
```

```
compare_model(X_train,X_test,Y_train,Y_test)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs fa
iled to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
   https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
   https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
******Logistic Regression*******
Confusion Matrix
[[3065   81]
 [ 355  198]]
Classification Report
              precision    recall  f1-score   support

       False       0.90      0.97      0.93      3146
        True       0.71      0.36      0.48       553

    accuracy                           0.88      3699
   macro avg       0.80      0.67      0.70      3699
weighted avg       0.87      0.88      0.87      3699

----------------------------------------------------------------------------------------------------

*******Random Forest Classifier*******
 Confusion Matrix
[[3042  104]
 [ 240  313]]
 Classification Report
              precision    recall  f1-score   support

       False       0.93      0.97      0.95      3146
        True       0.75      0.57      0.65       553

    accuracy                           0.91      3699
   macro avg       0.84      0.77      0.80      3699
weighted avg       0.90      0.91      0.90      3699
```

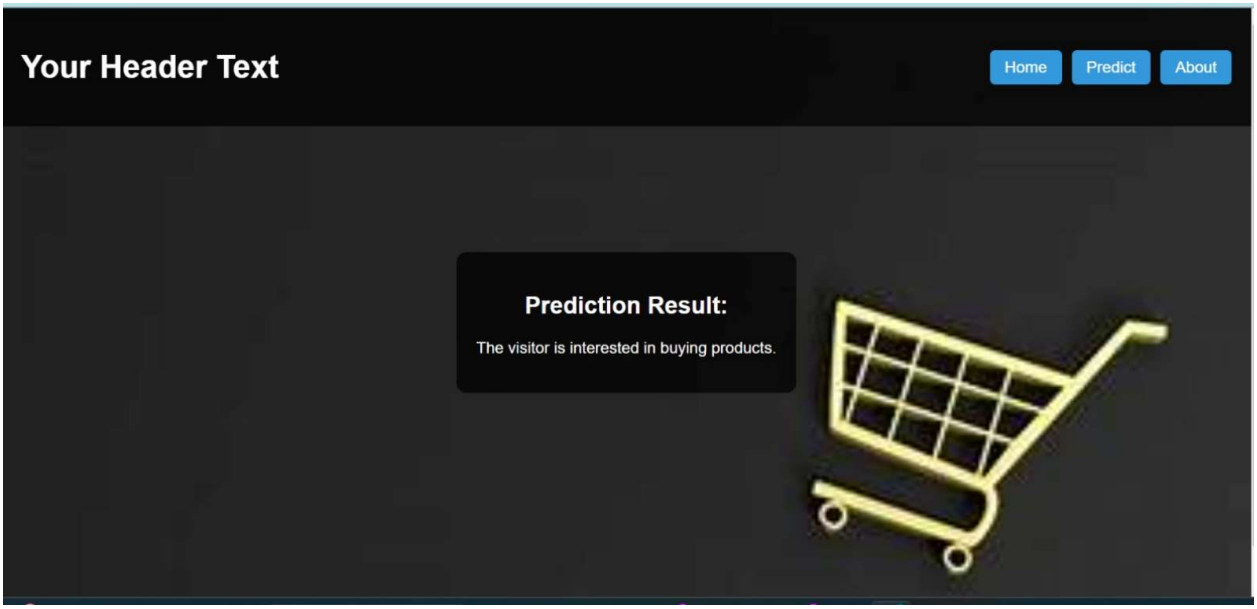
[ ]:
```
rf = RandomForestClassifier()
rf.fit(X_train,Y_train)
ypred = rf.predict(X_test)

cv = cross_val_score(rf,X,Y,cv = 5)
np.mean(cv)
```

t[ ]: 0.8966747769667478

RESULTS

Output Screenshots:




```
rf = RandomForestClassifier()
rf.fit(X_train,Y_train)
ypred = rf.predict(X_test)

cv = cross_val_score(rf,X,Y,cv = 5)
np.mean(cv)
```

```
0.8966747769667478
```

ADVANTAGES AND DISADVANTAGES

Advantages of using machine learning for online shoppers intention prediction include:

- Personalized shopping experiences: ML algorithms can analyze past shopping behavior and preferences to recommend products that a shopper is more likely to buy.

- Increased sales: By predicting shoppers' intentions, online retailers can offer targeted promotions and discounts, which can lead to more sales.

- Improved customer satisfaction: ML algorithms can help retailers optimize their website and user experience, leading to higher customer satisfaction and loyalty.

Disadvantages may include:

- Privacy concerns: Collecting and analyzing personal data to predict shopping behavior may raise privacy concerns among some shoppers.

- Bias: ML algorithms may be biased towards certain demographics or products, leading to unfair recommendations or promotions.

- Limited accuracy: While ML algorithms can make predictions with a high degree of accuracy, they are not perfect and may make mistakes. This can lead to frustrated shoppers and lost

larger datasets and is often more accurate.

Both algorithms have their strengths and weaknesses, and the choice between them depends on the specific problem being addressed. Logistic regression may be a better choice when the goal is to understand the relationship between variables, while random forest may be better when accuracy is the primary concern and there are many variables to consider.

CONCLUSION

Best model having higher accuracy after paramtere tuning and cross validation is **RANDOM FOREST CLASSIFICATION MODEL.**

FUTURE SCOPE

Online shopper intention prediction using machine learning has a lot of potential in the future. As more and more people turn to online shopping, businesses are looking for ways to better understand their customers and provide personalized experiences. By analyzing data on customer behavior, ML algorithms can learn to predict what shoppers are likely to do next, such as which products they are likely to buy or which pages they are likely to visit. This can help businesses optimize their online stores and marketing strategies to better meet the needs of their customers.

APPENDIX

Source Code:

Source Code for WEB APP

Project Demo link :

ML PART: ML DEMONSTRATION

WEB APP PART : WEB APP DEMONSTRATION