Viva Questions

General

1. **What is SmartCart?**

   SmartCart is an e-commerce purchase prediction project designed to forecast whether a customer will make a purchase during a session.

2. **What is the main goal of this project?**

   The goal is to analyze customer behavior data, train ML models, and deploy an application that predicts purchase likelihood, enabling data-driven business decisions.

3. **Who can benefit from this system?**

   E-commerce platforms, digital marketers, product managers, and analysts can use it to optimize conversions and improve customer engagement.

Dataset & Features

1. **What dataset was used?**

   The dataset contains session-level customer interaction data, including demographics, browsing behavior, and purchase outcomes.

2. **What are the key features in the dataset?**

   Features include age, gender, location, membership status, returning customer flag, device type, time spent, pages viewed, cart items, wishlist items, and more.

3. **What is the target variable?**

   The target variable is **purchase**, where `1` indicates a purchase and `0` indicates no purchase.

4. **How was missing data handled?**

   Missing data was addressed during preprocessing by appropriate cleaning techniques to ensure reliable model performance.

Exploratory Data Analysis (EDA)

1. **What were the key insights from EDA?**

Insights included the impact of age, device type, membership, time spent, and cart/wishlist items on purchase probability.

2. **Why is EDA important in this project?**

   EDA helps uncover trends, validate assumptions, and identify influential features before model training.

3. **Which features showed the strongest correlation with purchase behavior?**

   Cart items, session value, and membership status showed high predictive importance.

Preprocessing & Feature Engineering

1. **How were categorical variables encoded?**

   Categorical variables were encoded using LabelEncoder to keep dimensionality low for Random Forest models.

2. **Why was StandardScaler used?**

   StandardScaler was applied to normalize numerical features, stabilizing model performance.

3. **What feature engineering steps were applied?**

   Feature selection, encoding, scaling, and transformation were used to enhance model quality.

Model & Training

1. **Which model was chosen?**

   RandomForestClassifier was selected due to its robustness and ability to handle categorical + numerical data.

2. **Why Random Forest over other models?**

   It handles non-linear relationships, requires minimal preprocessing, and avoids overfitting compared to simpler models.

3. **What evaluation metrics were used?**

   Metrics included precision, recall, F1-score, accuracy, confusion matrix, and ROC curve.

4. **What were the classification results?**

   The model achieved an accuracy of ~73.8%, with good balance between precision and recall.

Results & Interpretation

1. **What does the confusion matrix indicate?**

   It shows how well the model distinguishes between purchase vs. non-purchase sessions, identifying both correct and incorrect predictions.

2. **What does the ROC curve tell us?**

   The ROC curve demonstrates the model's discriminative power, with an AUC indicating strong predictive performance.

3. **Which features were most important in predictions?**

   Cart items, session value, time spent, and membership were top contributors.

Business Impact

1. **How can businesses use this model?**

   They can target high-potential customers with personalized offers, optimize marketing spend, and improve conversion rates.

2. **What business problems does it solve?**

   It reduces wasted ad spend, enhances personalization, and improves customer retention strategies.

3. **How reliable are the predictions in real-world settings?**

   The model performs reliably on test data, but real-world deployment requires continuous monitoring and retraining.

Deployment

1. **How was the model deployed?**

   The model was deployed using **Streamlit**, allowing interactive visualization and predictions via a web interface.

2. **Why Streamlit for deployment?**

It’s lightweight, easy to use, and ideal for rapid prototyping and cloud deployment.

3. **Can this system scale to production?**

   Yes, with enhancements like Dockerization, CI/CD pipelines, and API integration for real-time predictions.

Tools & Technologies

1. **Which technologies were used?**

   Python, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Joblib, and Streamlit.

2. **Why was Joblib used?**

   Joblib was used to save and load trained models and scalers efficiently.

Future Improvements

1. **What are the limitations of this project?**

   The dataset is limited in size; class imbalance and unseen customer behavior may affect generalization.

2. **What future improvements are suggested?**

   Future work includes hyperparameter tuning, gradient boosting models, SHAP explainability, and production-grade deployment with Docker and CI/CD.