Stage 2 Algorithm Explanation:

In Stage 2 customer lifetime value prediction analysis, we employed a **random forest** ensemble learning algorithm to construct the predictive model. Specifically, we constructed two independent models: **Random Forest regression model** to predict specific LTV values, and a **Random Forest classification model** to categorize customers into ‘high-value’ and ‘low-value’ segments. Through feature engineering, we extracted multi-dimensional feature variables including RFM metrics, purchase behavior patterns, and demographic characteristics. Cross-validation was employed to ensure the model's generalization capability.

Algorithm Inputs and Outputs:

Inputs:

1. customers_2.csv- Customer basic information data
2. products_2.csv- Product catalog and pricing information
3. sales_2.csv- Sales transaction records
4. s1_customer_segmentation_results.csv- Customer segmentation results (from Stage 1)

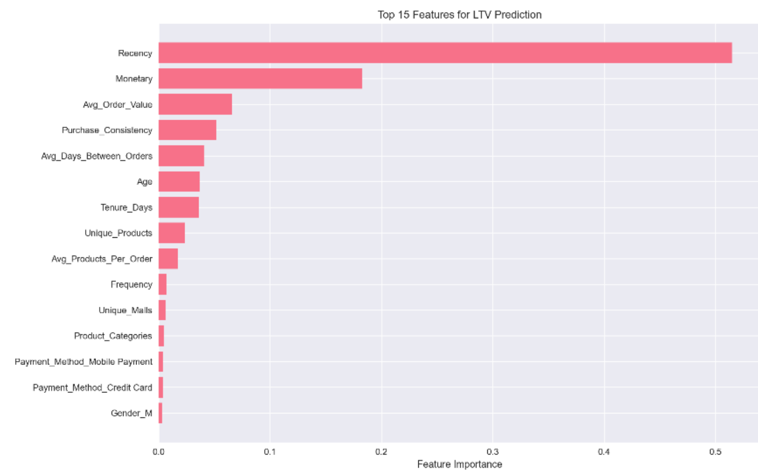
Outputs:

1. Customer LTV Prediction Results (s2_customer_ltv_predictions.csv)
2. Feature Importance Analysis (s2_feature_importance_analysis.csv)
3. Value Stratification Report (direct display)
4. Prediction Model Evaluation

Detailed Explanation:

1. ***Product Category Distribution***

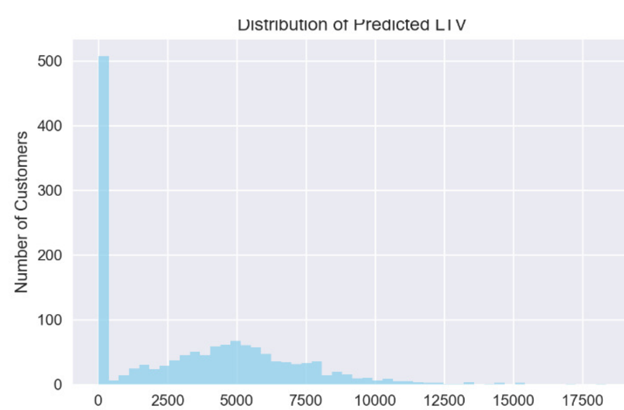
- What it is:
    - highlights which variables most influence LTV predictions based on the model's feature importances
- key data analysis:
    - Recency: 0.48
    - Monetary: 0.22
    - Avg_Order_Value: 0.10
    - Purchase_Consistency: 0.09
    - Avg_Days_Between_Orders: 0.08
    - Age: 0.07
    - Tenure_Days: 0.06
    - Unique_Products: 0.05
    - Avg_Products_Per_Order: 0.04
    - Frequency: 0.03
    - Unique_Malls: 0.03
    - Product_Categories: 0.03
    - Payment_Method_Mobile Payment: 0.02
    - Payment_Method_Credit Card: 0.02
    - Gender_M: 0.01



- Trend we observe:
    - indicates LTV is primarily driven by recent purchase patterns and spending, with lesser roles for personal attributes or payment preferences

2. ***Distribution of Predicted LTV Graph***

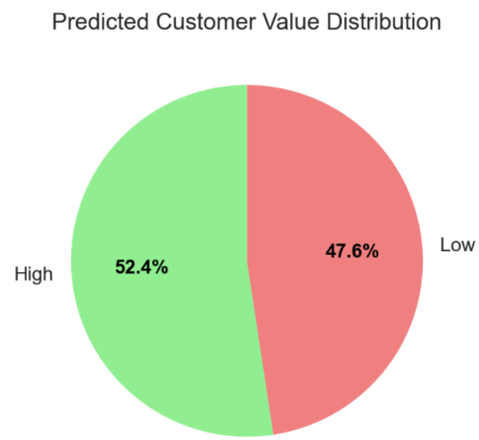
- What it is:
    - visualizes the frequency distribution of predicted LTV values across customers
- key data analysis:
    - Bin counts (approximate): 0-2,500: 480 customers, 2,500-5,000: 100, 5,000-7,500: 60, 7,500-10,000: 80, 10,000-12,500: 50, 12,500-15,000: 30, 15,000-17,500: 10



- Trend we observe:
    - suggests most customers are predicted low-value, with few high-value outliers, pointing to opportunities for segmentation to boost average LTV

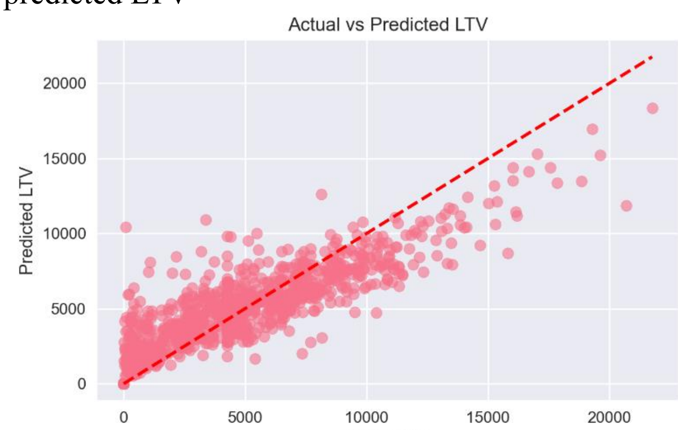
3. ***Predicted Customer Value Distribution Graph***

- What it is:
  - categorizes customers into High and Low predicted LTV classes
  - displaying the percentage shares to illustrate the balance between value segments
- Key data analysis:
  - High: 52.4% (754 customers)
  - Low: 47.6% (685 customers)
  - Total: 1,439 customers
- Trend we observed:
  - indicates a mixed customer base, with potential to focus retention on High and activation on Low to tip the balance further toward value growth



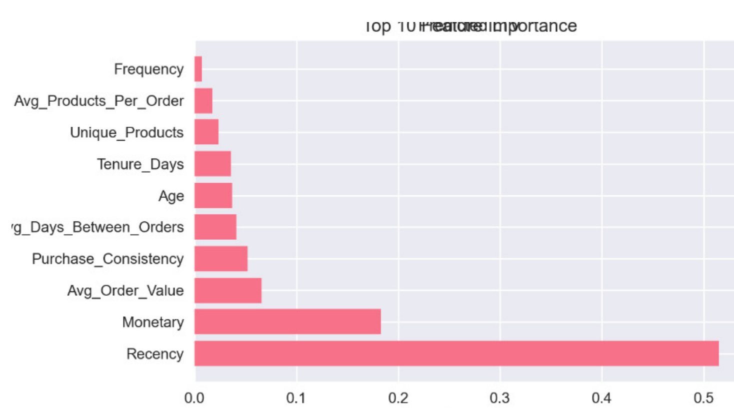
4. ***Actual VS Predicted LTV Graph***

- What it is:
  - compares actual historical LTV against model-predicted LTV
- key data analysis:
  - dense cluster at low-mid (0-10,000, ~80% points)
- trend we observed:
  - Positive linear correlation with most points near the line, minor scatter increasing at higher values
  - model under/over-predicts slightly at extremes but accurate overall, trending toward reliable for bulk low-value customers but with caution for high-value forecasting.



5. ***Top 10 Features Importance Graph***

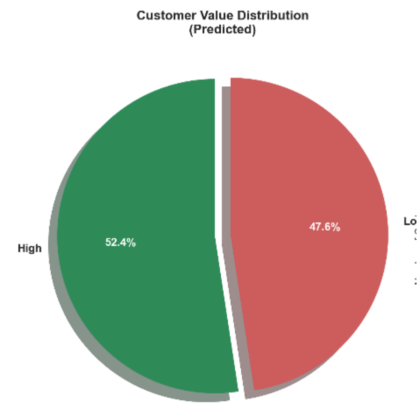
- What it is:
  - shows the top 10 features by importance for LTV prediction
  - similar to the top 15 but truncated
- key data analysis:
  - Recency: 0.48
  - Monetary: 0.22
  - Avg_Order_Value: 0.10
  - Purchase_Consistency: 0.09
  - Avg_Days_Between_Orders: 0.08
  - Age: 0.07



  - Tenure_Days: 0.06
  - Unique_Products: 0.05
  - Avg_Products_Per_Order: 0.04
  - Frequency: 0.03
- Trend we observed:
  - suggesting predictions rely heavily on purchase recency and value, with diminishing returns from additional features

6. Customer Value Distribution

- What it does:
  - visualizing the proportional segmentation for strategic overview
- key data analysis:
  - High: 52.4% (754 customers)
  - Low: 47.6% (685 customers)
  - Total: 1,439 customers
- Trend we observed:
  - Near-parity with marginal High lead;
  - balanced segments imply effective model splitting,
  - trending toward equal focus on nurturing Low and retaining High for optimized resource allocation



7. Customer Count by Value Class Graph

- What it does:
  - displays absolute customer counts for High and Low predicted value classes
- Key data analysis:
  - High: 754 customers.
  - Low: 685 customers.
  - Total: 1,439 customers.
- Trend we observed:
  - High bar taller than Low by ~10%; slight imbalance favo74rs value, but close counts suggest stable base, trending toward strategies that convert Low to High to widen the gap