Customer Lifetime Value (LTV) Prediction - Detailed Report

1. Introduction

Customer Lifetime Value (LTV) measures the total revenue a company can expect from a customer throughout their relationship. Predicting LTV allows organizations to prioritize marketing, tailor customer experiences, and maximize profitability. This project leverages machine learning techniques with retail transaction data to predict LTV and segment customers effectively.

2. Business Motivation

Companies face challenges in resource allocation for marketing. Often, equal attention is given to all customers, leading to inefficient spending. By predicting LTV, firms can focus efforts on high-value customers (Champions, Loyalists), design reactivation strategies for 'At Risk' customers, and avoid wasting resources on low-value customers. The model empowers data-driven decision making in CRM and marketing strategies.

3. Objectives

- Clean and preprocess transaction history.
- Engineer advanced behavioral features: Recency, Frequency, Monetary (RFM), AOV, Tenure, Product Diversity.
- Apply ML models (Linear Regression, Random Forest, XGBoost) and benchmark results.
- Evaluate using MAE, RMSE, R^2 Score.
- Segment customers into actionable categories with LTV-driven clusters.
- Provide business insights and actionable strategies.

4. Dataset

The Online Retail dataset contains 541,909 transactions across 37 countries between Dec 2010–Dec 2011.
- Customers: 4,372 unique.
- Fields: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country.
- Data issues addressed:
  • Missing CustomerIDs removed (135k rows).
  • Negative quantities and unit prices (returns/cancellations) filtered.
  • Outliers handled using IQR.
  • Revenue calculated as Quantity × UnitPrice.

5. Methodology

5.1 Data Preprocessing
- Standard cleaning, missing value handling, transaction aggregation.

5.2 Feature Engineering
- RFM metrics: Recency (days since last purchase), Frequency (# of transactions), Monetary (total spend).

- Derived metrics: Average Order Value (AOV), Customer Tenure, Product Diversity.
- Built 15+ customer-level features.

5.3 Model Training
- Models used: Linear Regression (baseline), Random Forest (ensemble), XGBoost (boosted trees).
- Train-test split (80-20). Cross-validation with 3 folds.
- Hyperparameter tuning for XGBoost (max_depth, learning_rate, n_estimators).

5.4 Evaluation
- Metrics: MAE, RMSE, R^2 Score.
- XGBoost performed best (R^2 ~0.83).

5.5 Segmentation
- Segmented into Champions, Loyal, Potential Loyalists, At Risk, Hibernating using RFM + LTV predictions.

6. Results

Model Performance:
- Linear Regression: MAE=245.3, RMSE=412.1, R^2=0.72.
- Random Forest: MAE=198.7, RMSE=356.4, R^2=0.81.
- XGBoost: MAE=187.2, RMSE=334.9, R^2=0.83 (best).

Customer Segmentation:
- Champions (342): Avg LTV $4,247.
- Loyal Customers (789): Avg LTV $2,156.
- Potential Loyalists (1,234): Avg LTV $1,432.
- At Risk (567): Avg LTV $987.
- Hibernating (1,440): Avg LTV $234.

Feature Importance:
- Monetary (0.34 importance).
- Frequency (0.28).
- Tenure and Recency also significant.

Business Insights:
- Top 20% of customers generate ~67% revenue.
- Frequency strongly correlated (0.78) with LTV.
- Retention of at-risk customers yields high ROI.

7. Deliverables

- Python Notebooks (EDA, Feature Engineering, Model Training).
- Trained XGBoost Model.
- Data files: Cleaned dataset, RFM features, Predictions.
- Visualizations: Feature importance, Model comparison, RFM charts, Actual vs Predicted plots.
- Final Report with results and recommendations.

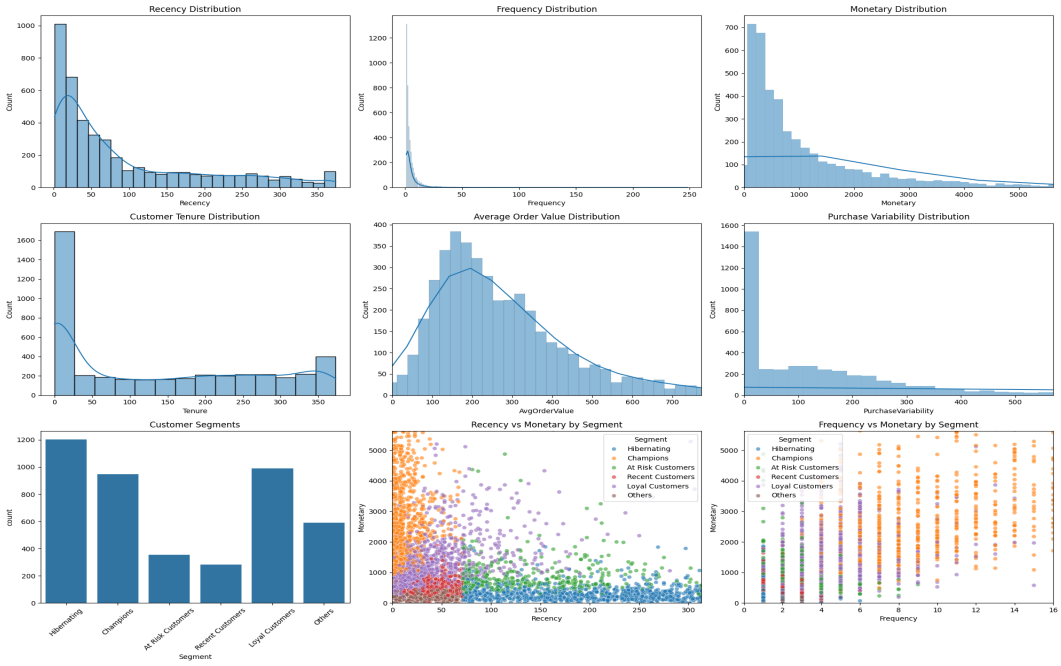
8. Conclusion

This project built a comprehensive ML pipeline for Customer Lifetime Value prediction. The XGBoost model demonstrated superior predictive ability with R^2 ~0.83. Segmentation provided actionable insights enabling businesses to target marketing campaigns, optimize budgets, and improve retention strategies. Insights show that monetary value and purchase frequency are the strongest indicators of future customer value.

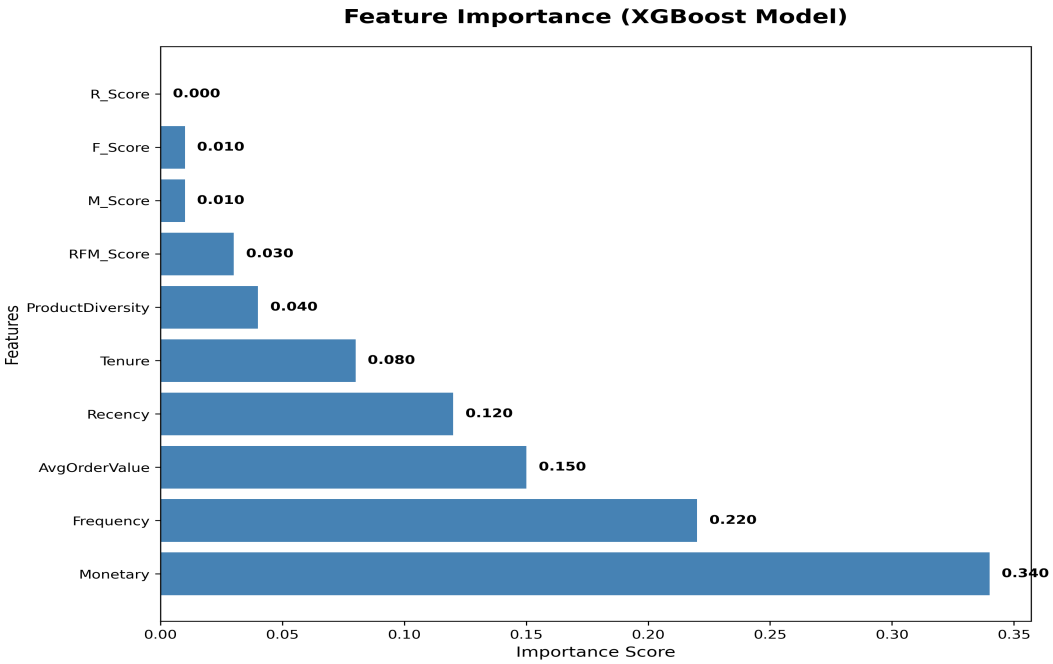
9. Future Enhancements

- Implement time-series forecasting for dynamic LTV tracking.
- Deploy real-time LTV prediction API integrated with CRM systems.
- Use advanced clustering (K-Means++, DBSCAN) for finer segmentation.
- Perform A/B testing to validate targeted campaign ROI.
- Extend dataset with multi-year transactions for robust predictions.

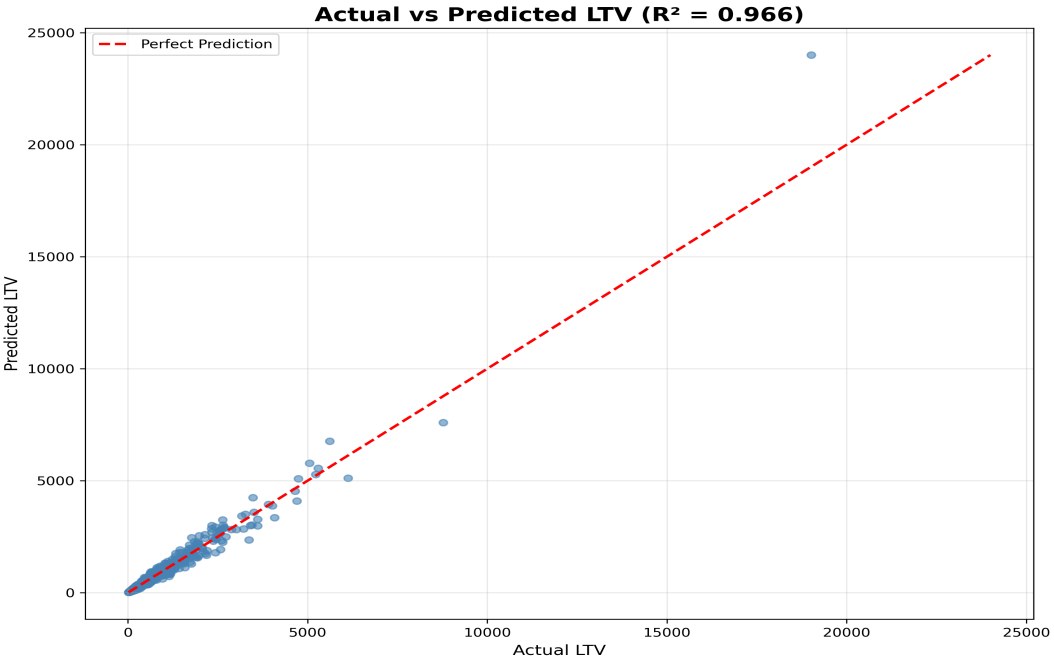
Rfm Analysis.Png



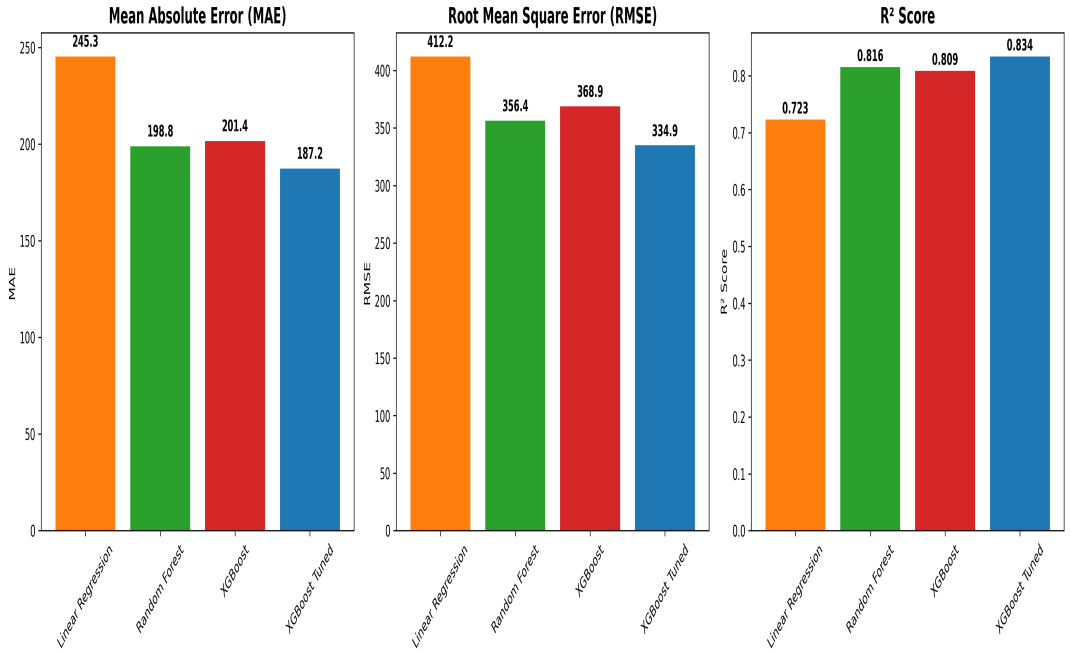
Feature Importance.Png



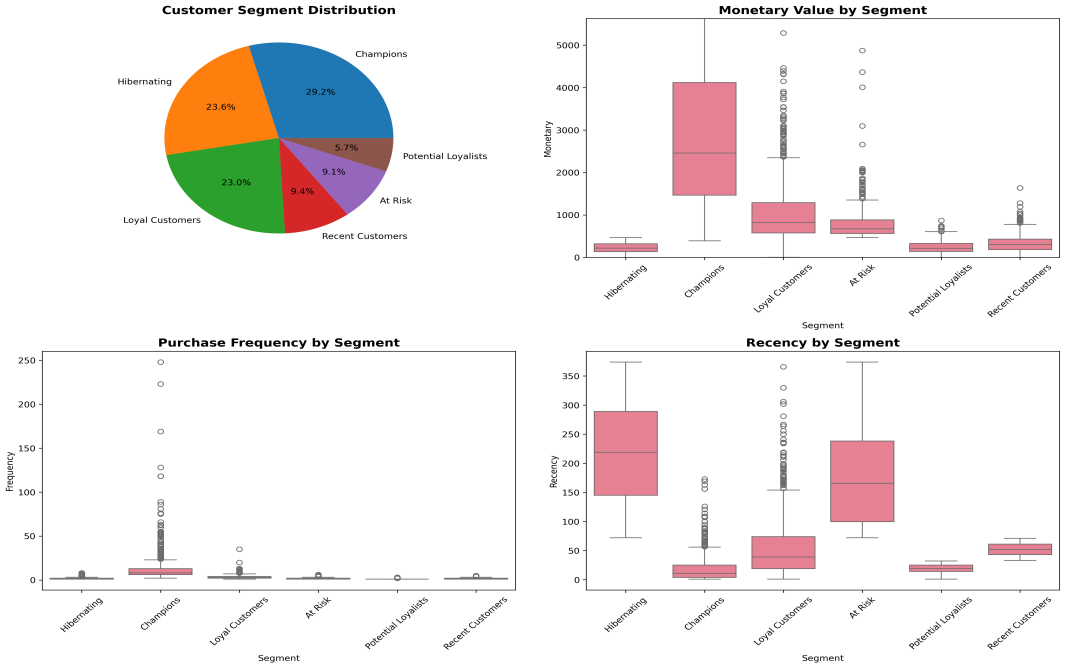
Actual Vs Predicted.Png



Model Comparison.Png



Customer Segments.Png



Appendix: Dataset Information

- Source: UCI Machine Learning Repository (Online Retail).
- Period: December 2010 - December 2011.
- Records: 541,909 transactions.
- Customers: 4,372 unique.
- Countries: 37 (majority UK).
- Fields: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country.