Customer Lifetime Value Prediction
Presentation Notes

The Starks
Othmane Zizi (261255341)
Fares Joni (261254593)
Tanmay Giri (261272443)

February 26, 2026

Repository

`https://github.com/othmane-zizi-pro/nwa`

Slide 1 — Title

- Introduce the team: The Starks — Othmane Zizi, Fares Joni, Tanmay Giri.
- State the project: Predict Customer Lifetime Value using RFM features, regularized regression, and causal inference.
- Point to the GitHub repository for full reproducibility.

Slide 2 — Business Context & Our Approach

- **The Problem**: Not all customers are equal; e-commerce businesses invest equally across their customer base, but a small segment drives the majority of revenue. Without predictive CLV, marketing spend is misallocated.
- **Our Approach**: Predict → Explain → Act. Build a regression model on RFM + behavioral features to predict 6-month forward CLV. Use SHAP for explainability, causal inference for treatment effects, and monitoring for production readiness.
- Highlight the four key numbers: 4,266 customers, 1.07M transactions, 8 features, 6-month prediction window.

Slide 3 — Research Hypotheses

- **H_1 — Monetary Features Dominate**: Past spending behavior (Monetary, AvgOrderValue) will be the strongest predictor of future CLV, outweighing frequency and recency signals. *Verdict: Confirmed* — Monetary importance = 2,136.

- **H_2 — Regularization Improves Generalization**: Regularized models (Lasso, Ridge) will generalize better than tree-based methods due to moderate dimensionality and potential collinearity. *Verdict: Confirmed* — Lasso $R^2 = 0.81$ vs Random Forest 0.50.
- **H_3 — High-Frequency $\neq$ High-CLV**: Causal analysis will show that purchase frequency alone does not causally increase CLV; it may reflect selection bias from already-loyal customers. *Verdict: Supported* — ATE ≈ -0.17 (log scale).

Slide 4 — Data Pipeline

- **Source**: UCI Online Retail dataset — UK e-commerce, 2010–2011; 1,067,371 transaction rows.
- Fields: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country.
- **Cleaning steps**: Remove cancelled orders (Invoice starts with 'C'), drop null CustomerID rows ($\sim$25%), filter Quantity > 0 and UnitPrice > 0, remove outliers beyond 99th percentile.
- **Result**: 805,549 clean rows $\rightarrow$ 4,266 unique customers. Observation period ending Dec 2010; 6-month prediction window to Jun 2011.

Slide 5 — Feature Engineering

- **RFM Core**: Recency (days since last purchase), Frequency (total number of orders), Monetary (total revenue generated).
- **Behavioral Extensions**: AvgOrderValue (mean spend per order), AvgBasketSize (items per transaction), NumUniqueProducts (product diversity), AvgTimeBetweenPurchases (purchase cadence), Tenure (days as customer).
- Walk through the feature importance bars: Monetary dominates at 2,136, followed by Frequency (681), AvgOrderValue (534), AvgBasketSize (407), NumUniqueProducts (276), Recency (55).
- **Target variable**: FutureCLV — sum of customer revenue in the 6-month prediction window (Dec 2010 – Jun 2011).

Slide 6 — Modeling Strategy

- Five candidate models: Linear Regression (baseline), Ridge Regression (L2), Lasso Regression (L1, **best model**), Random Forest (ensemble), Gradient Boosting (ensemble).
- **Train/Test split**: 80/20 stratified split preserving CLV distribution; all features standardized via StandardScaler.
- **Cross-validation**: 5-fold CV on training set. Best CV $R^2 = 0.53 \pm 0.18$ (Lasso).
- **Hyperparameter tuning**: Randomized search over alpha values. Best Lasso $\alpha = 4.90$.

Slide 7 — Model Performance

- **Winner**: Tuned Lasso Regression — $R^2 = 0.810$, RMSE = £1,756, MAE = £545, CV $R^2 = 0.53 \pm 0.18$.
- Linear models outperform tree ensembles because the target is dominated by Monetary (a near-linear relationship).
- Walk through the model comparison chart (RMSE/MAE/R^2 across all 5 models) and the best model analysis (predicted-vs-actual scatter + residual distribution).

Slide 8 — Feature Importance & SHAP Analysis

- **Feature importance**: Shown across all models. Monetary dominates with 3× the importance of Frequency.
- **SHAP summary**: Confirms a positive, near-linear relationship — higher past spending directly predicts higher future CLV.
- **Key insight**: This explainability is critical for stakeholder trust. Monetary is the clear dominant driver, validating H_1.

Slide 9 — Lift Chart & Business Impact

- Lift chart shows model-guided targeting dramatically outperforms random selection.
- 5.55× lift at top 10%: top decile captures 5.5× more value than random targeting.
- 3.62× at top 20%: top quintile still highly efficient for campaign targeting.
- 2.76× at top 30%: nearly 3× improvement over untargeted outreach.
- 1.83× at top 50%: model adds value even at broader targeting.
- Business implication: allocating campaign budget to the top two deciles is highly efficient.

Slide 10 — Section Divider: Causal Inference

- Transition from prediction to causation.
- Motivate: "We can predict CLV, but does purchase frequency *cause* higher CLV, or is it just correlated?"
- Introduce the three key elements: Treatment (High Frequency), Outcome (Future CLV), Method (Meta-Learners).

Slide 11 — Causal Inference Setup

- **Treatment**: Binary indicator — customer is above the median purchase frequency. Simulates a "loyalty program" or "re-engagement" intervention.
- **Outcome**: `log(1 + FutureCLV)` — log-transformed to handle right-skewed distribution and make treatment effects interpretable on a relative scale.
- **Controls**: All pre-treatment covariates — Recency, Monetary, Tenure, AvgOrderValue, AvgBasketSize, NumUniqueProducts, AvgTimeBetweenPurchases.
- **S-Learner** (LRS Regressor): Single model with treatment as feature. Estimates ATE via counterfactual prediction.
- **T-Learner** (XGB Regressor): Separate models for treated/control groups. More flexible, captures heterogeneous effects.

Slide 12 — Causal Inference Results

- **S-Learner (LRS)**: ATE $= -0.179$, 95% CI $[-0.378, +0.021]$ — not statistically significant (CI includes zero).
- **T-Learner (XGB)**: ATE $= -0.167$, 95% CI $[-0.258, -0.077]$ — statistically significant negative effect.
- **Interpretation**: High purchase frequency does *not* causally increase CLV. The negative ATE (~ -0.17 on log scale) suggests that after controlling for confounders, high-frequency buyers may be **deal-seekers** with lower per-transaction value. Correlation $\neq$ causation.
- CATE distribution shows heterogeneity: different customers respond differently.
- Causal feature importance: Frequency is the top driver of treatment heterogeneity.

Slide 13 — Threats to Validity

- **Selection Bias in Treatment Assignment**: Frequency is observed, not randomized. The median split may conflate inherently loyal customers with those influenced by external factors. Propensity score matching could help, but unobserved confounders remain.
- **CV vs Test Gap (0.53 → 0.81)**: The gap between cross-validation R^2 (0.53) and test R^2 (0.81) may indicate a favorable test split or temporal patterns. Additional temporal CV would strengthen confidence.
- **Single-Geography, Single-Period**: Data comes from one UK retailer in 2010–2011. Generalization to other markets, product categories, or time periods is not guaranteed.
- **Limited Feature Set**: 8 RFM-based features capture purchase behavior but miss demographics, marketing touchpoints, seasonality, and product affinity — factors that likely influence CLV.

Slide 14 — Monitoring & Deployment Plan

- **Production checks**: Input schema validation, feature range checks, prediction distribution monitoring, latency & throughput SLAs.
- **Alert thresholds**: PSI > 0.1 (moderate drift, investigate); PSI > 0.25 (severe drift, trigger retrain); R^2 drop $> 15\%$ (performance degradation, retrain).
- **Maintenance**: Quarterly scheduled retraining, automated drift-triggered retrain, champion/challenger testing, model versioning & rollback.
- **Current status**: All features stable — max PSI $= 0.035$. No feature exceeds the 0.1 warning threshold. Model is production-ready.

Slide 15 — Key Takeaways & Thank You

- **Takeaway 1**: Lasso achieves $R^2 = 0.81$ — regularized regression outperforms tree-based models with strong predictive power and interpretability.
- **Takeaway 2**: 5.55× lift in top decile — model-driven targeting captures 5× more value than random selection, direct ROI for marketing spend.
- **Takeaway 3**: Frequency $\neq$ causal driver — causal analysis reveals high frequency does not increase CLV, challenging the "more visits = more value" assumption.
- **Takeaway 4**: Production-ready pipeline — full monitoring with PSI drift detection (all stable at < 0.035), automated alerts, and quarterly retraining schedule.
- **Next steps**: Add temporal cross-validation, integrate product-level features, A/B test targeting strategies, expand to multi-market.
- Close with "Thank You" and open for questions.