INSY 674 FINAL PROJECT

Customer Lifetime Value Prediction

Predictive modeling, explainability, causal inference, and monitoring for an end-to-end enterprise data science workflow.

Othmane Zizi (261255341) | Fares Joni (261254593) | Tanmay Giri (261272443) | https://github.com/othmane-zizi-pro/nwa

CONTEXT

Business Context and Objective (5.9.1)

Problem

- Marketing budgets are often distributed uniformly across unequal customers.
- High-value customers can churn without proactive intervention.
- Revenue concentration means ranking quality drives business impact.

Objective

- Predict 6-month CLV per customer from historical transaction behavior.
- Use model output to prioritize retention and campaign spend.
- Track value capture with lift and production monitoring.

HYPOTHESIS

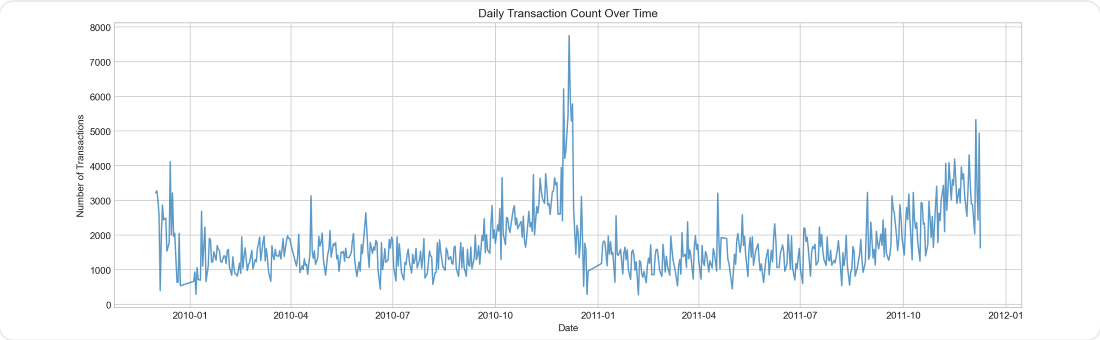
Hypotheses and Testing Framing (5.9.2)

- H1: RFM + behavioral features can predict future CLV.
- H2: Regularized linear models generalize better than tree models on this feature set.
- H3: Causal uplift analysis identifies who should receive treatment.
- Null framing: treatment has no measurable effect on log1p(CLV).

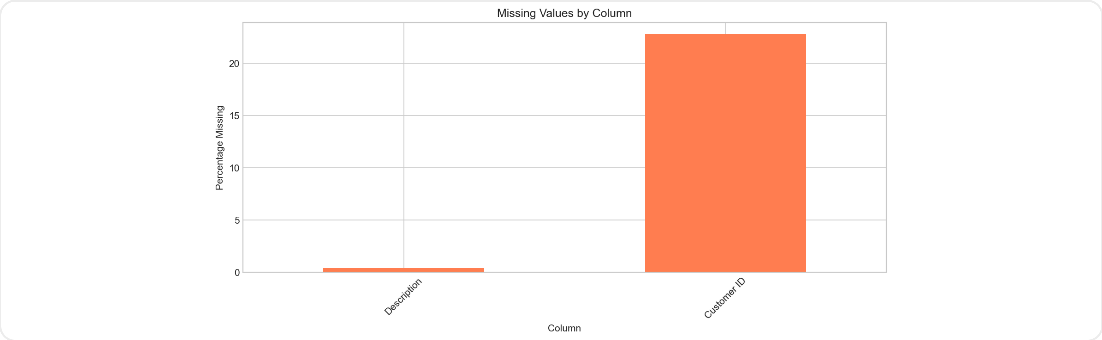
EDA

Exploration Visuals: Coverage and Quality (5.3)

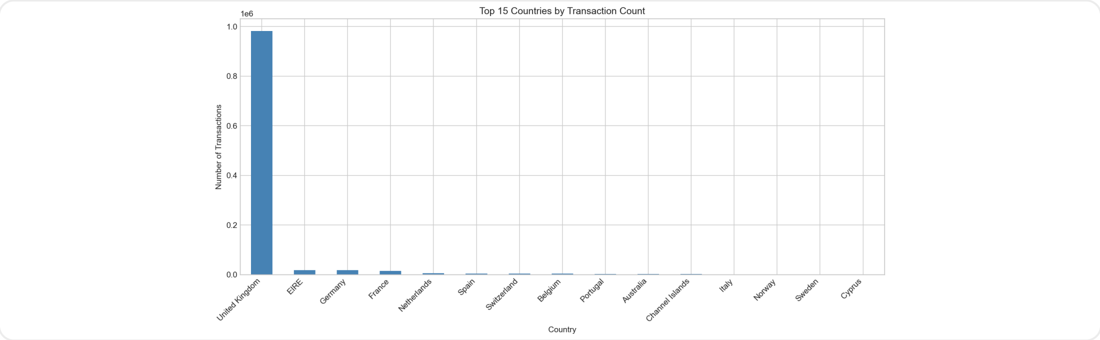
- Initial EDA covered time trends, missingness, geography, and transactional behavior.
- These checks informed cleaning decisions before feature engineering.



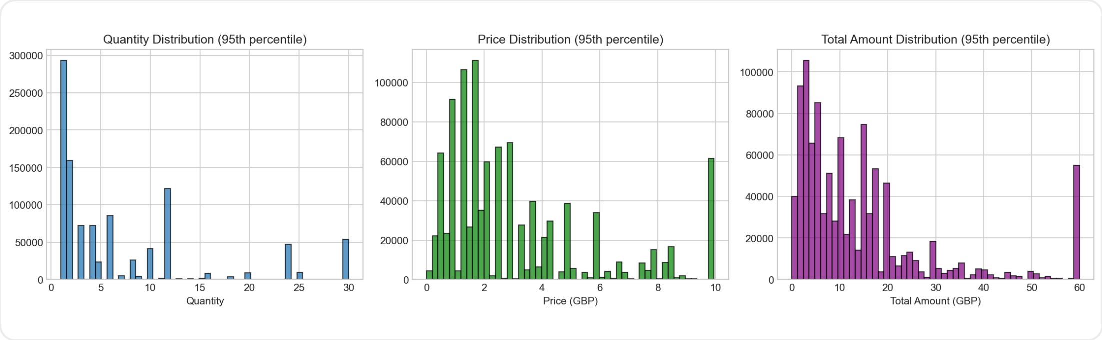
Daily transaction volume



Missing values profile



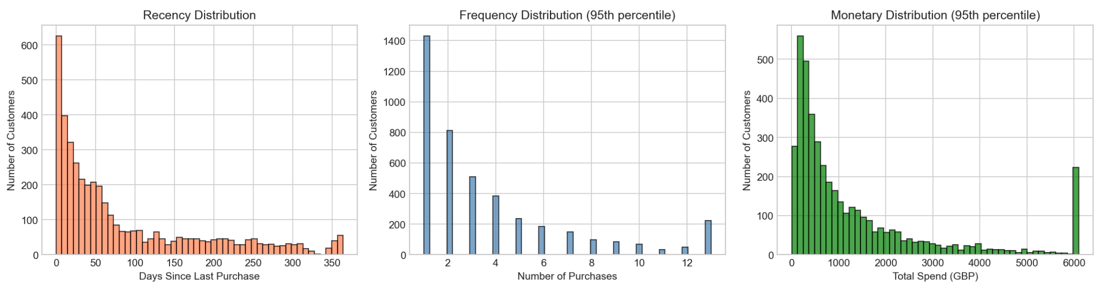
Top purchasing countries



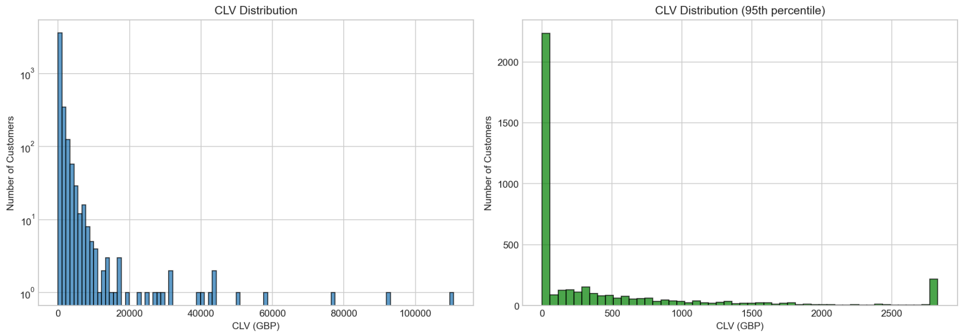
Price and quantity distribution

Feature and Target Distributions (5.4)

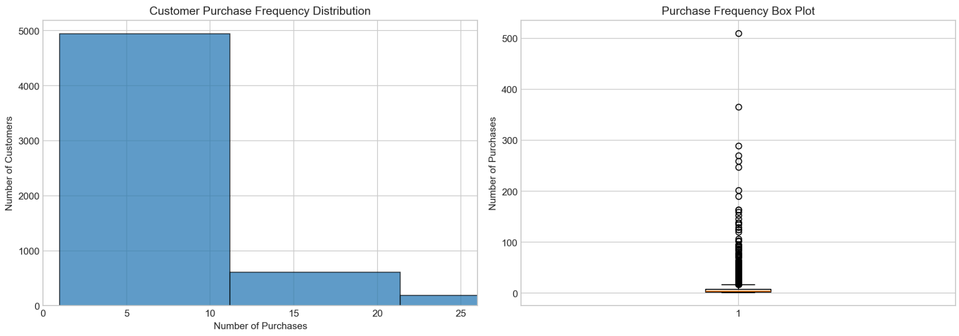
- RFM and behavioral features show heavy skew and strong concentration effects.
- Correlation and distribution checks shaped model family and scaling decisions.



RFM feature distributions



Future CLV distribution



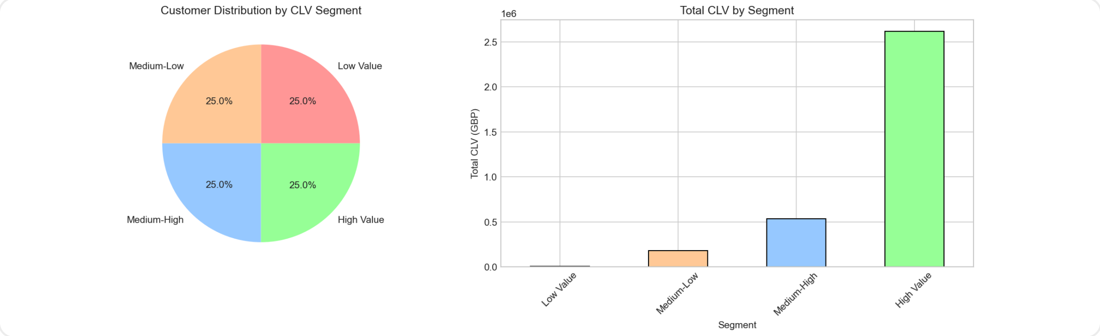
Customer purchase frequency

Feature correlation matrix

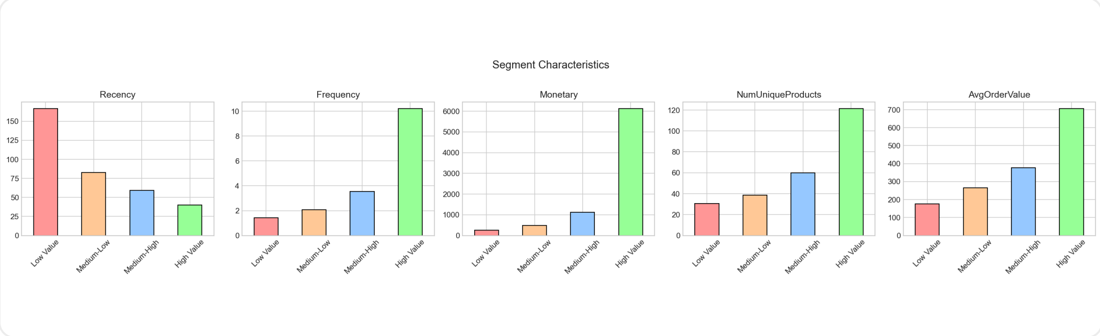
SEGMENTATION

Segmentation and Behavioral Clusters

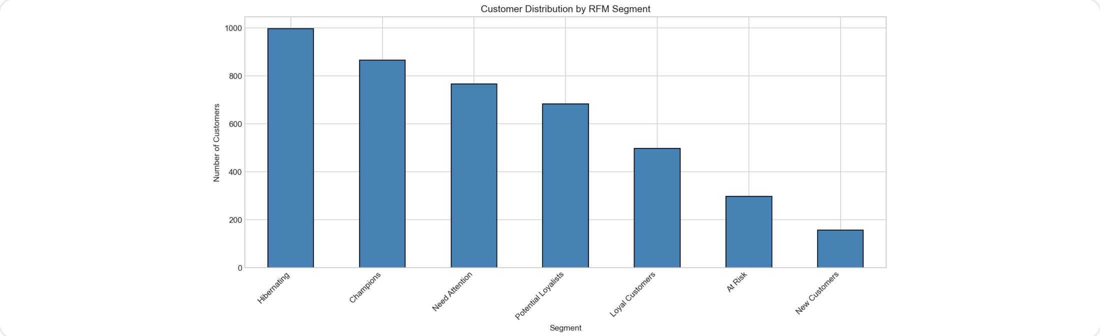
- Segmentation complements prediction by converting scores into action buckets.
- Profiles and cluster diagnostics guide campaign strategy design.



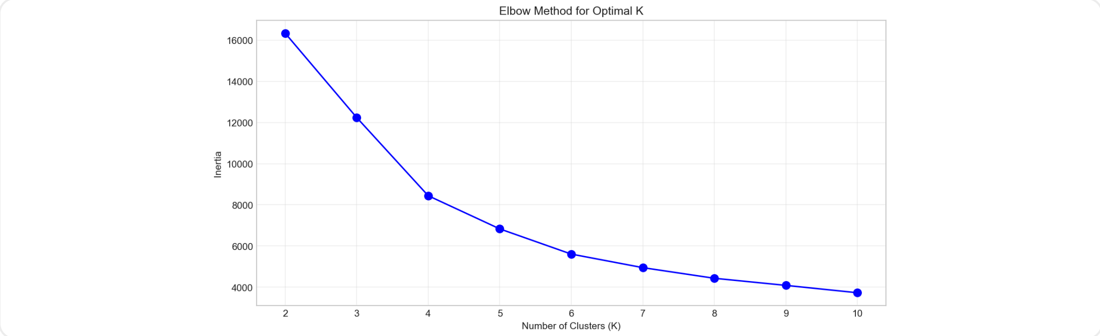
Segment size distribution



Segment profile comparison



RFM segment map



K-means elbow method

MODELING

Modeling Approach and Evaluation (5.5, 5.6)

- Train/val/test split: 80/10/10 at customer level; validation for tuning, test as final holdout.
- Compared Linear, Ridge, Lasso, Random Forest, Gradient Boosting, XGBoost.
- Metrics: RMSE, MAE, R2, MAPE and lift for business prioritization.
- Randomized search used for fine-tuning top candidates.

RESULTS

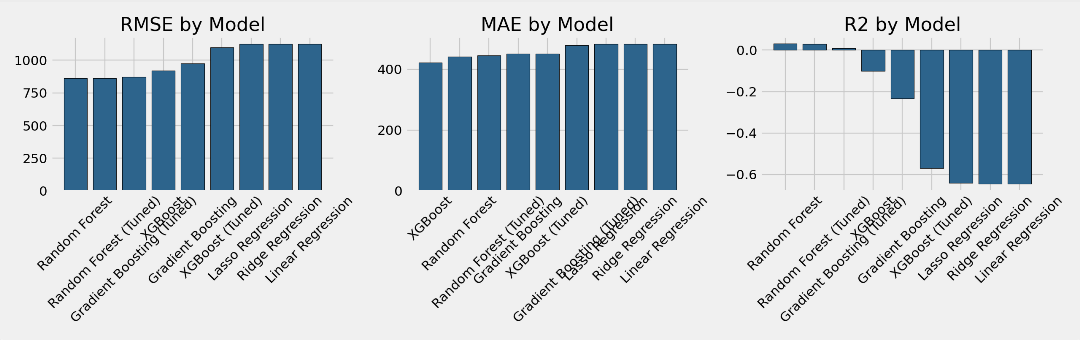
Predictive Performance (5.9.5)

MODEL	RMSE	MAE	R2
Random Forest	862.46	441.21	0.032
Lasso Regression	1761.83	547.71	0.808
Linear Regression	1762.46	547.99	0.808
Ridge Regression	1762.94	547.98	0.808
XGBoost	2387.82	534.66	0.648

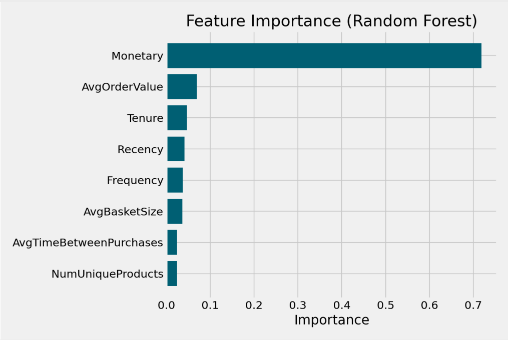
MODEL DIAGNOSTICS

Diagnostics and Business Utility (5.7, 5.8, 5.9.6)

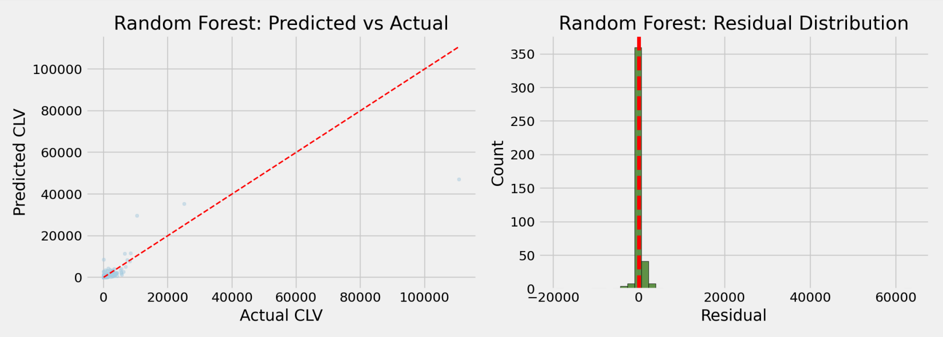
- Model ranking, residual diagnostics, and lift all support selection of a tuned regularized model.
- The top-decile lift demonstrates strong practical targeting value.



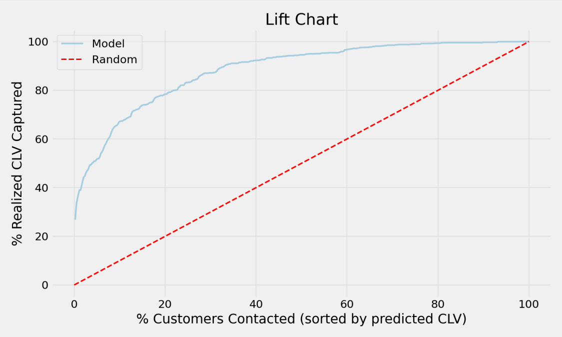
Model metric comparison



Global feature importance



Residual and fit diagnostics

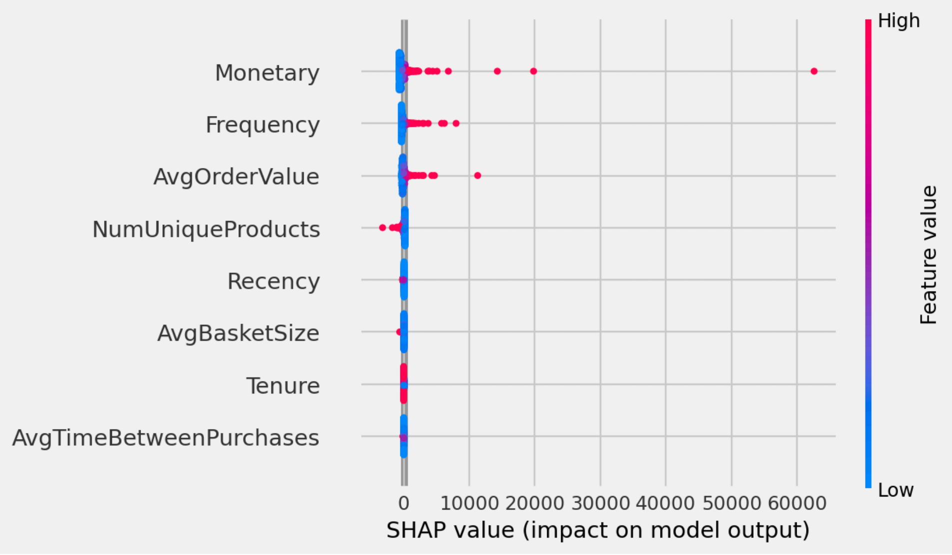


Lift chart (business impact)

EXPLAINABILITY

SHAP Explainability (5.9.6)

- SHAP confirms direction and magnitude of global drivers.
- Monetary and frequency behavior dominate prediction contributions.
- Interpretability supports stakeholder trust and campaign governance.
- Lift@10% = 6.69x over random targeting.



PHASE 2

CausalML Phase

From prediction to intervention impact estimation

CAUSAL SETUP

Target, Treatment, Controls

Definitions

- Outcome: log1p(CLV).
- Treatment: campaign vs control assignment.
- Controls: RFM + behavioral features.

Required Methods

- LRSRegressor (S-learner style).
- XGBTRegressor (T-learner style).
- Feature importance and CATE distribution analyzed in notebook 06.

CAUSAL RESULTS

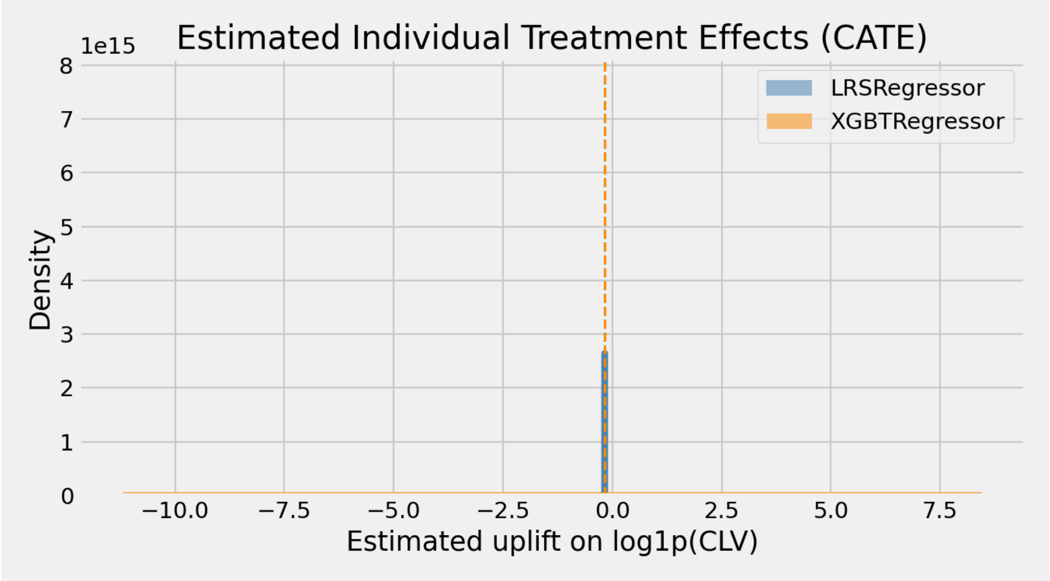
Average Treatment Effects

MODEL	ATE (LOG1P CLV)	CI LOWER	CI UPPER
LRSRegressor	-0.1785	-0.3783	0.0213
XGBTRegressor	-0.1673	-0.2575	-0.0770

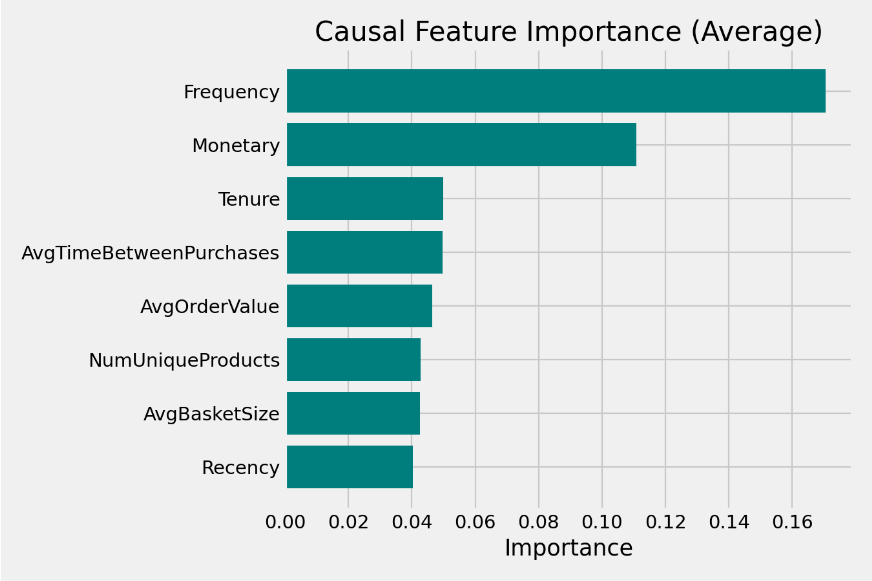
CAUSAL EXPLAINABILITY

Uplift Distribution and Causal Importance

- Treatment effects vary substantially by customer profile (heterogeneous CATE).
- Causal importance supports policy targeting beyond raw CLV ranking.



CATE distribution



Causal feature importance

VALIDITY

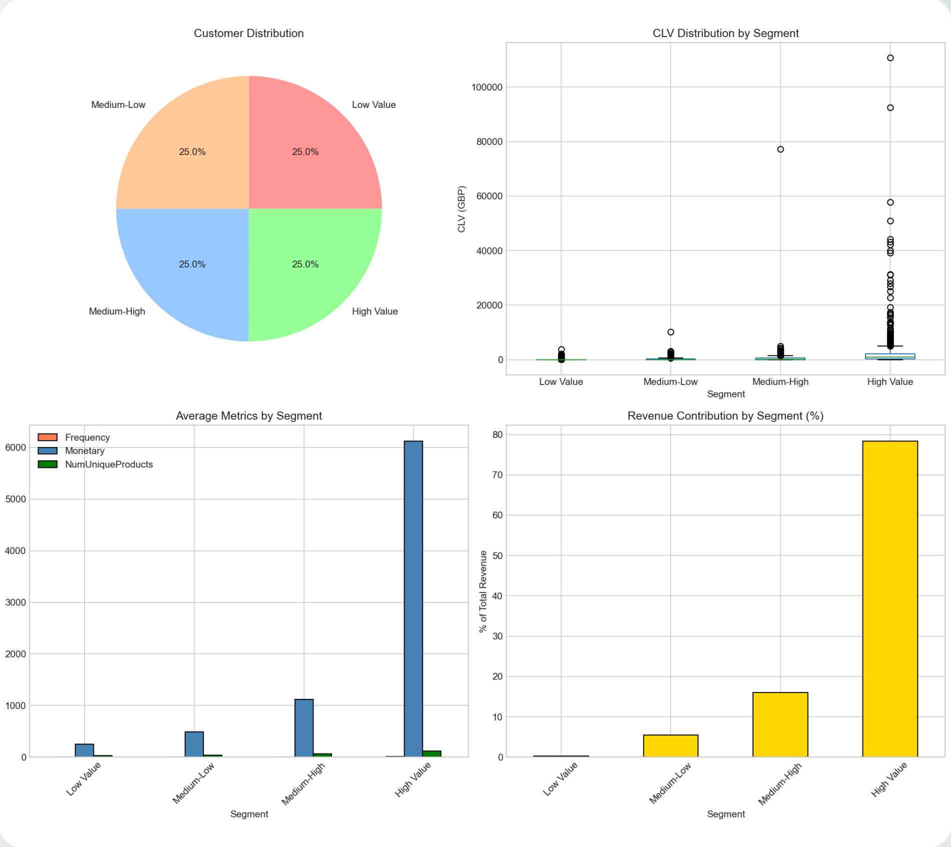
Threats to Validity (5.9.7)

- Treatment assignment is observational/simulated, not randomized.
- Potential unobserved confounding and temporal effects.
- Outliers and historical policy effects can bias relationships.
- External validity may differ across geographies and verticals.

MLOPS

Launching, Monitoring, Maintenance (5.10)

- Production checks: schema validation, feature quality checks, artifact versioning.
- Monitoring: RMSE drift, CLV capture@K, feature drift via PSI.
- Alerting: warning at PSI >= 0.10, critical at PSI >= 0.25.
- Observed max PSI in validation run: 0.036 (stable).



CONCLUSION

Conclusions, Lessons, and Next Steps (5.9.8, 5.9.9)

- An end-to-end enterprise DS pipeline was implemented and validated.
- Predictive + explainable + causal layers improve decision quality.
- Next step: validate uplift on real campaign logs / A/B tests.
- This repository is submission-ready with reproducible artifacts.

THANK YOU

Questions and Discussion

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

Othmane Zizi (261255341) | Fares Joni (261254593) | Tanmay Giri (261272443)