Case study:

Predicting Customer Lifetime Value (LTV) for Fintech Users with H2O AutoML

---

1. Problem

Understanding and predicting the **Lifetime Value (LTV)** of users is critical for fintech platforms aiming to optimize acquisition strategies, personalize offerings, and allocate resources wisely. This project aims to forecast the LTV of a user based on early behavioral and transactional signals

> Business Impact: By identifying high- or low-value customers early, companies can focus retention and marketing budgets where they matter most.

---

2. Dataset

- **Source:** Synthetic dataset from Kaggle
- **Size:** \~10,000 rows (user-level records)
- **Features:**
- User demographics (e.g., age, region, gender)
- Behavior metrics (app usage, support tickets, logins)
- Financial indicators (transactions, credit score, loan amount)
- **Target Variable:** `LTV` (numeric, continuous)

---

3. Workflow Summary

1. Data Preparation

- Cleaned and encoded categorical variables (e.g., gender, region)
- Handled missing values (mode/mean imputation where relevant)
- Scaled numeric features

2. Feature Engineering

- Aggregated user activity features (e.g., avg. transactions per week)
- Created composite behavior scores
- Removed outliers in `LTV`

3. AutoML Modeling

- Tool used: **H2O AutoML** (regression mode)
- Ran with 10-fold CV and leaderboard evaluation
- Trained over 20 model types (GLM, GBM, XGBoost, Stacked Ensembles)
- Optimized for **RMSE** and **R^2**

4. Results

Metric	Value
RMSE	0.39 (without leakage, realistic)
R^2 Score	\~0.36
Best Model	H2O GBM / Stacked Ensemble

Note: Results indicated solid performance with good generalization on hold-out set. Top features included total transactions, app usage frequency, and initial credit rating.

5. Lessons

- Synthetic data has limits: woudn't reflect noise or distribution of real fintech users
- Regression problems require careful **outlier handling**, especially with long-tail distributions like LTV
- H2O AutoML made model selection easy but still required **manual feature review** to avoid leakage or redundancy

6. What I'd Do Next

- Test on real-world anonymized fintech data (if available, e.g. publicly traded companies)
- Add time-based features (e.g., recency/frequency trends)
- Add cost-benefit modeling: how much does mispredicting LTV cost?
- Build a simple dashboard for internal growth/marketing teams to simulate LTV predictions

7. Final Reflection

> "LTV prediction is not just about forecasting numbers — it's about **mapping future value to early behavior**. Even with synthetic data, AutoML helped uncover the core signals of long-term retention and revenue."

🔗 GitHub Repository