International School of
Economics at TSU

THESIS PRESENTATION

Income Estimation Models Analysis: Balancing Accuracy with Regulatory Requirements in the Georgian Banking Sector

Sandro Gogaladze

July 12, 2025

INTRODUCTION

Banks have to **verify a borrower's income** before issuing a loan.

They typically rely on data from the revenue service, employer-provided salary documents, payslips, tax records, etc.

INCOME VERIFICATION

NO RECORDS	**HAVE RECORDS**
Paid in cash, no payslips	Official income often underreported
Income can’t be verified	Inaccurate assessment
↓	↓
Rejected	Misjudged risk

Traditional income verification often leads to incomplete or inaccurate assessments — and **both sides suffer**.

BORROWERS

Face **delays**, **uncertainty** — often labeled **high-risk** and **rejected** despite being creditworthy.

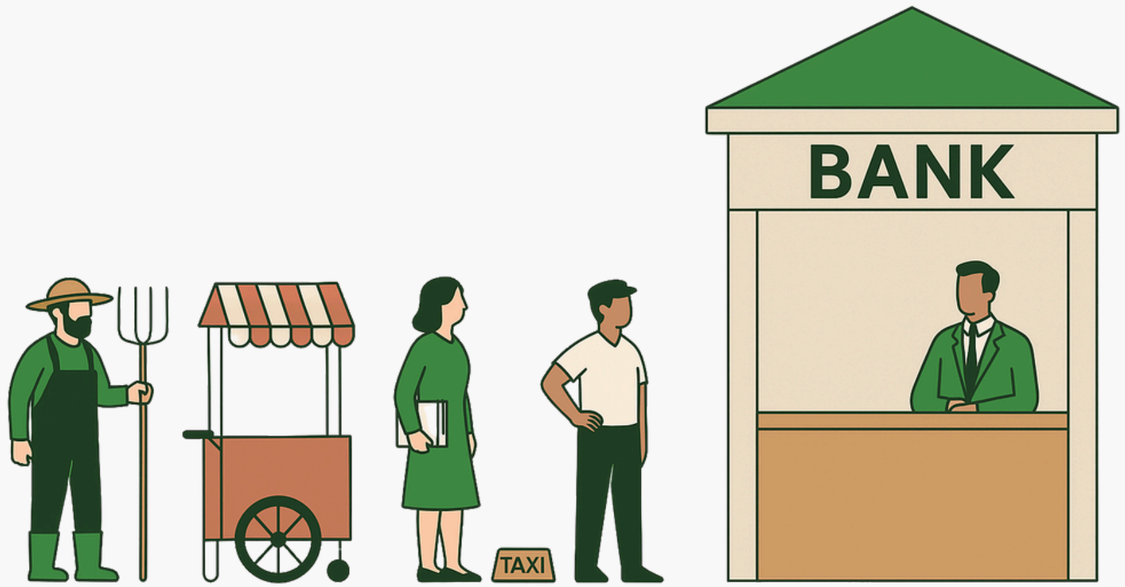
BANKS

Deal with **slow**, **costly**, **manual** verification processes — making it harder to serve borrowers efficiently.

THE RESULT

System that’s inefficient, costly, and fundamentally exclusionary.

Creditworthy — but invisible to traditional lending.



GEORGIAN BANKING SECTOR

Since 2020, the **National Bank of Georgia (NBG)** has **allowed** financial institutions to use statistical, machine learning, and AI-based models for income verification, within a regulated supervisory framework.

These are supervisory models **regulated by the NBG**, subject to formal review, validation, and approval prior to production use.

A few banks have already developed income estimation models, while **most are still researching and working** on it.

Legal Basis for Data-Driven Income Estimation

Regulation "On Approving the Regulation on Lending to Individuals", Approved by Order №44/04 of the President of the National Bank of Georgia Date: March 13, 2020

"... it is permissible to use a data-driven model, provided that the model allows for an adequate assessment of income. When using such a model, the requirements established by the National Bank of Georgia for data-driven models must be fully met."

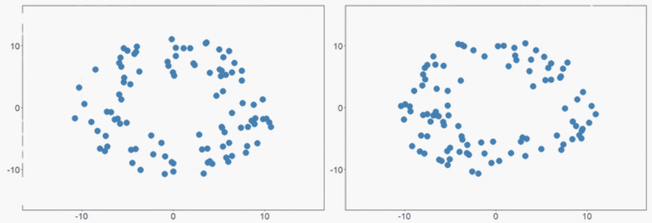
Building accurate and reliable income estimation models is challenging — there is **need for reserach**.

This thesis compares different modeling approaches, evaluating their predictive accuracy and alignment with regulatory requirements in the Georgian financial system.

DATA

This thesis is conducted in **collaboration** with the **Bank of Georgia (BOG).**

BOG provided a **synthetic dataset** designed to mirror the structure and statistical properties of real customer data.



Original data Synthetic data

DATASET DETAILS

22 expert-selected features from credit registry, transactions, balances, and internal sources.

Income-Related Features

- **Target**: Estimated stable monthly income — reflects long-term earning capacity, excluding temporary or irregular inflows
- **Inc_Past**, **Inc_Past_avg**, **Inc_Past_max**: Historical income levels used to capture consistency and trends
- **Inc_6M** : Total income inflow over the past 6 months
- **Salary**, **Inc_in** : Salary deposits and other credited income

Account Activity & Usage

- **Transfers_in**, **Transfers_out**, **Min_transfer_In**, **Tot_in**: Cash flow patterns
- **Acct_Trns**, **Transactions**, **Turnover** : Account usage behavior
- **Trn_max** : Largest single outgoing transaction

Balance Information

- **Balance**, **Bal_Cur** : Current and historical account balances

Loans & Liabilities

- **Liab_Tot** : Total outstanding liabilities
- **Loan** , **Loan_Cnt** : Active loan amounts and loan count
- **Payments**, **Payments_L**: Loan repayment history and monthly obligations

OBJECTIVES

TRADE - OFF

COMMERCIAL

Banks aim to **minimizing prediction error**.

Accuracy is the main priority.

VS

REGULATORY

Supervisors (e.g., NBG) emphasize **preventing overestimation** to avoid irresponsible lending.

Stability and conservatism matter more than precision.

Explained in detail →

This trade-off is the reason behind the thesis title:
“Balancing Accuracy with Regulatory Requirements”

REGULATORY REQUIREMENT

NBG PRACTICE

The National Bank of Georgia (NBG) recognizes that income estimation models are not perfectly accurate.

NBG allows some overestimation — typically up to 10% — as long as it occurs in no more than 10% of cases.

NBG

Thesis

SIMILAR BUT MORE FLEXIBLE APPROACH:

IMPORTANT

Overestimation is **allowed up to 20% or 200 GEL** — whichever is higher

However, threshold **exceedance** must occur in **less than 10%** of cases.

A model is compliant if no more than 10% of its predictions exceed the greater of 20% or 200 GEL over the actual income.

EXAMPLES

Actual income: **1,000** GEL Prediction: **1,180** GEL	→ 180 GEL (18%) over — acceptable (within 20%)
Actual income: **800** GEL Prediction: **980** GEL	→ 180 GEL (22.5%) over — acceptable (below 200 GEL)
Actual income: **500** GEL Prediction: **800** GEL	→ 300 GEL (60%) over — violation (exceeds both)
Actual income: **1500** GEL Prediction: **1850** GEL	→ 350 GEL (23%) over — violation (exceeds 20%)

MODELING STRATEGIES

Default models focus solely on minimizing prediction error, often ignoring regulatory constraints.

To address this, two categories of correction strategies are explored: **POST-HOC** **PRE-HOC**

POST-HOC CALIBRATION

- Applied after training
- Adds compliant correction layer
- Useful when retraining is not feasible

Method used:
- **Quantile Calibration**

PRE-HOC ADJUSTMENT

- Applied before training
- Integrates regulatory constraints into training
- Encourages the model to avoid violations by design

Methods used:
- **Huber + Threshold Loss**
- **Segment-Aware Huber + Threshold Loss**

BASELINE MODEL

Objective: Maximize predictive accuracy

Approach:

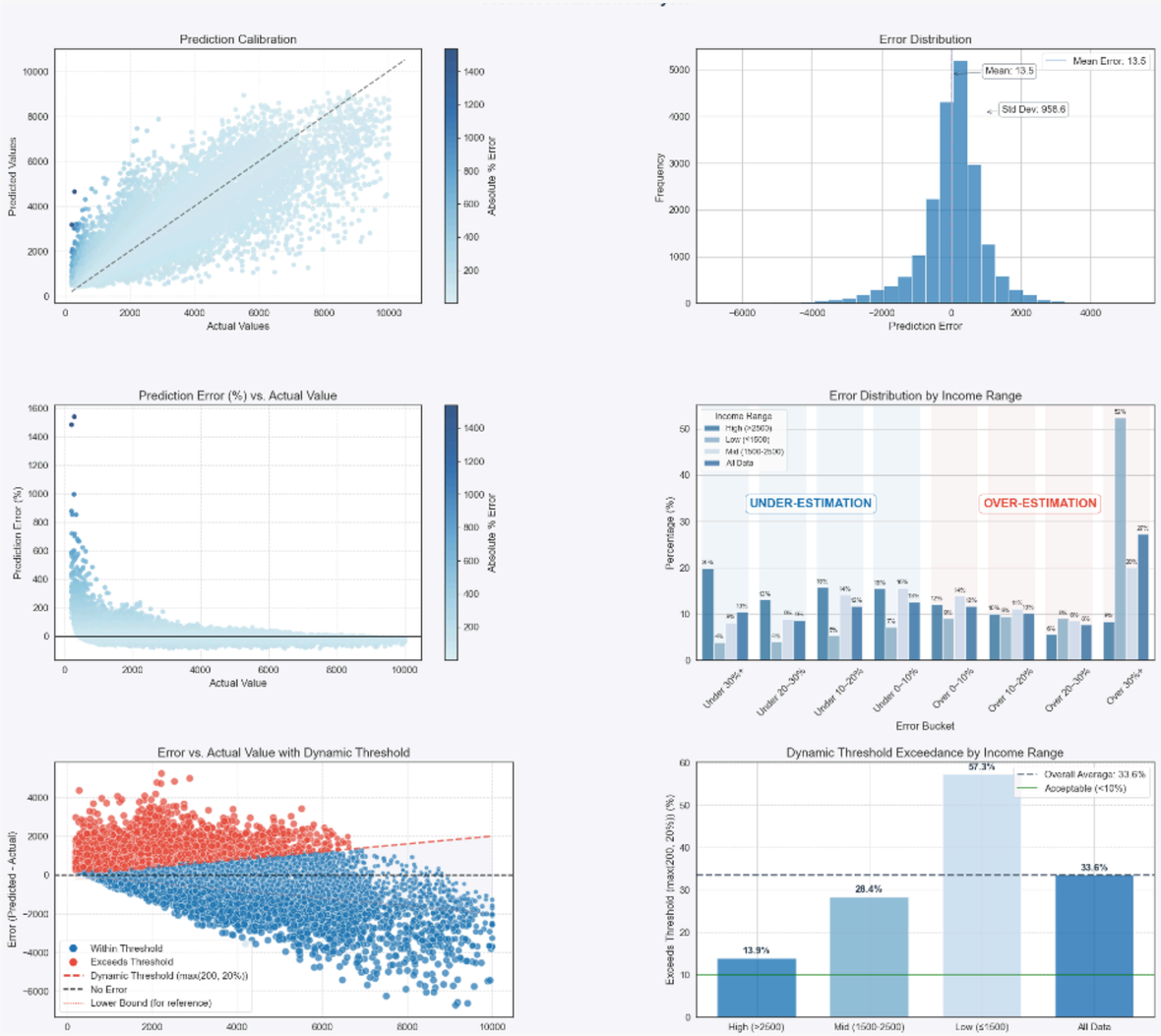
- Evaluated multiple tree-based models
- All performed similarly — selected XGBoost
- Performed data preprocessing
- Did hyperparameter tuning

Model Performance - Test set metrics:

- **MAE** ≈ 653, **R^2** ≈ 0.75
- **33.6%** of predictions exceeded regulatory thresholds (>10%)
  - · **High-income**: 13.9%
  - · **Mid-income**: 28.4%
  - · **Low-income**: 57.3%

Good accuracy, poor regulatory fit

→ **Needs adjustment: POST-HOC or PRE-HOC**



ISET

POST-HOC

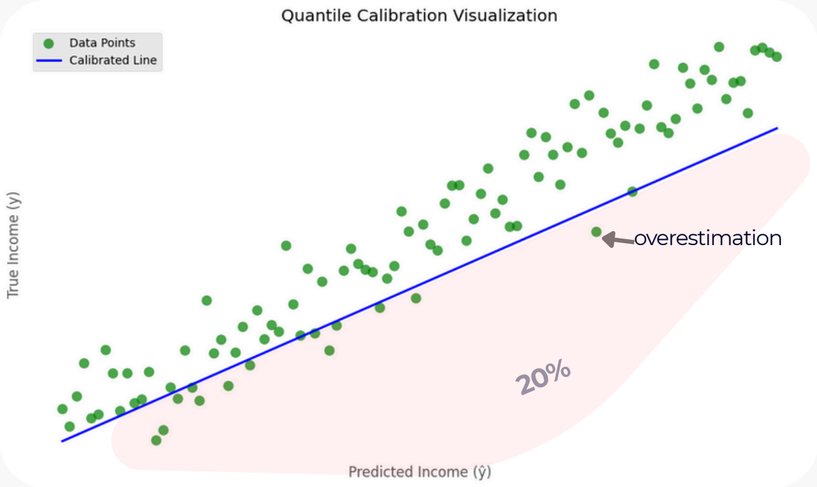
QUANTILE CALIBRATION

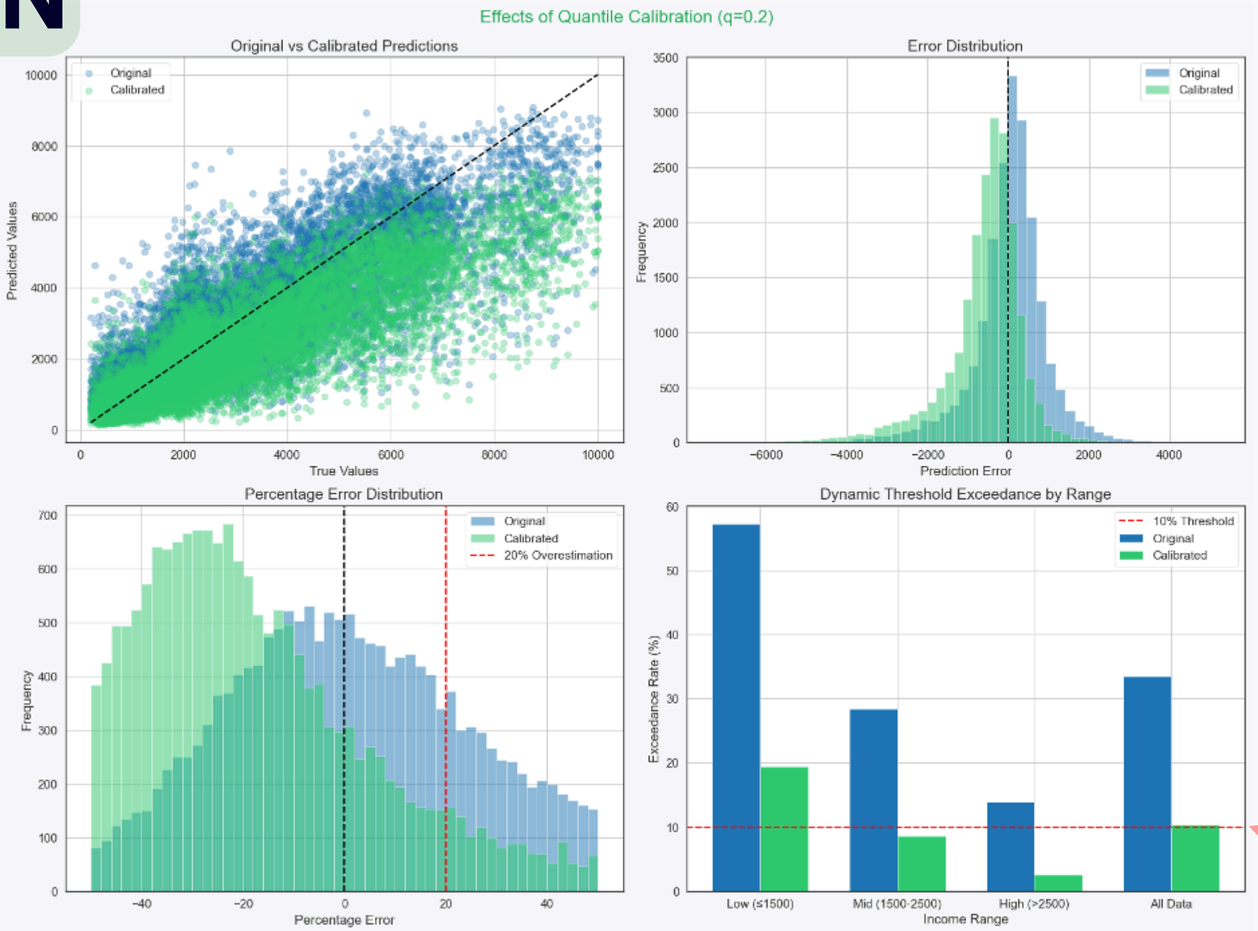
HOW IT WORKS:

- Trains a simple linear adjustment layer:
- Adjusted = $\beta_0 + \beta_1 \times$ Original Prediction

20^{th} quantile

- β_0 = - 192.92
- β_1 = 0.83



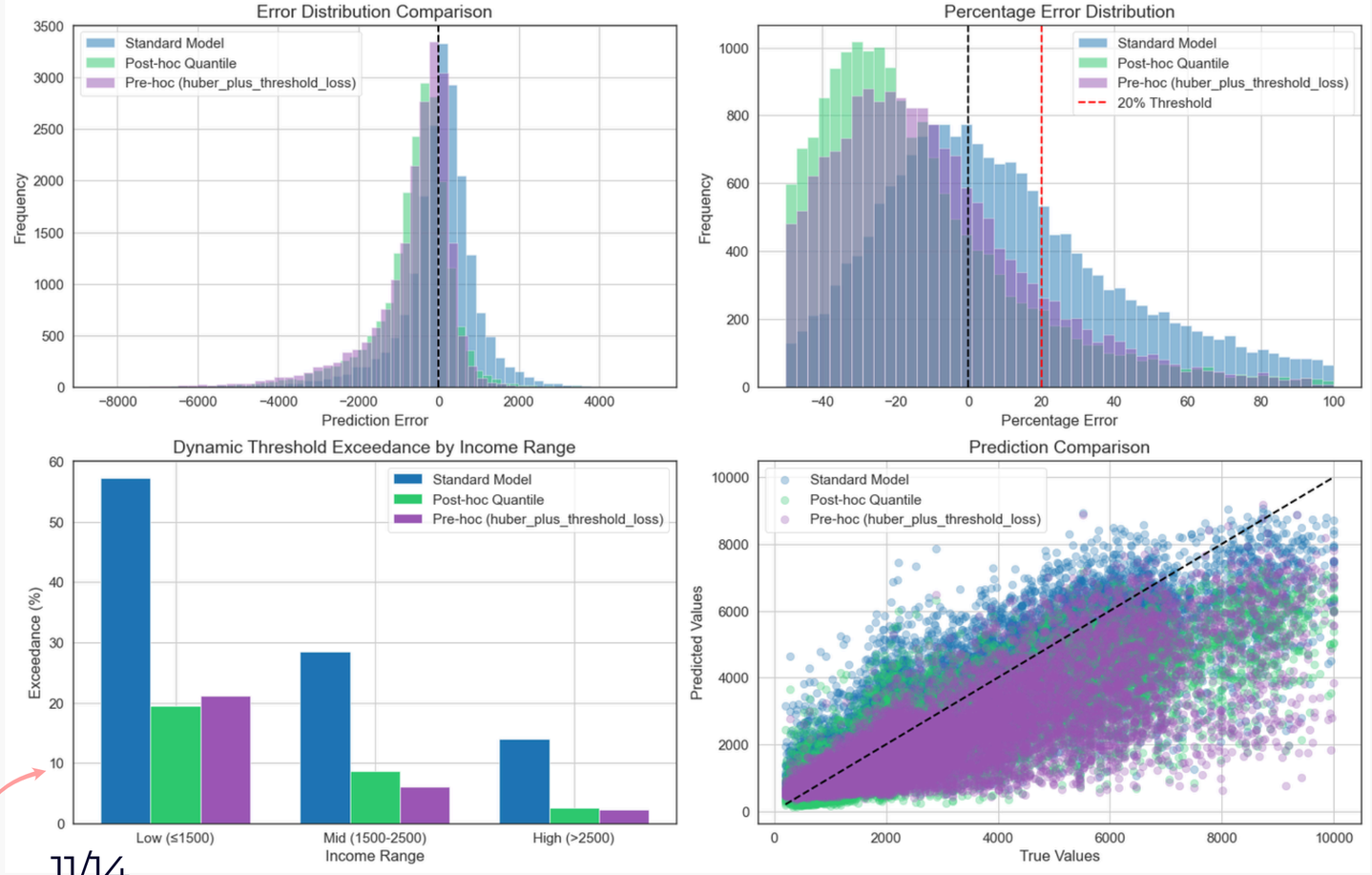


RESULTS:

- Accuracy drops (MAE ↑, R^2 ↓) → but compliance improves
- Violation rate reduced: **33.6% → 10.3%**
- Segment-Level Instability

PRE-HOC

HUBER + THRESHOLD LOSS



WHY HUBER LOSS?

- Less sensitive to outliers
- Behaves like MSE for small errors
- Switches to MAE for large errors
- Ideal for noisy financial data

ADDED PENALTY FOR THRESHOLD VIOLATIONS

- A custom **penalty** was added **for** predictions that **exceed the regulatory threshold**:

$$Loss = Huber\_Loss + \lambda \cdot Penalty\_for\_Threshhold\_Violation$$

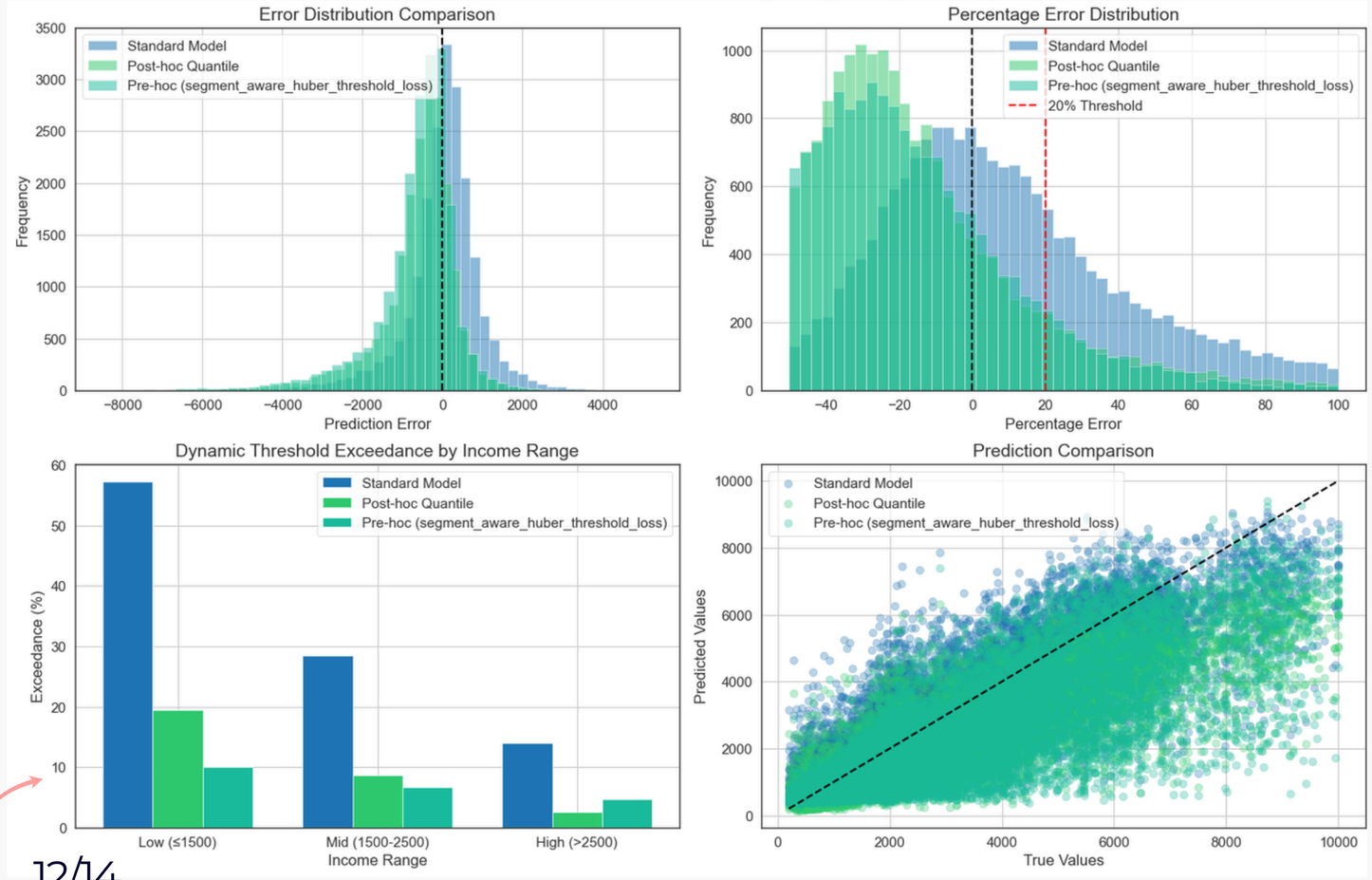
Incorporates both accuracy and compliance into training by design

RESULTS:

- **Less dramatic** shift than post-hoc — preserves more accuracy
- Regulatory Violation Rate: ↓ from **33.6% → 10.0%**
- Violation rate in low-income group remained high (21.2%)
- **Shows need for segment-specific adjustments**

PRE-HOC

SEGMENT-AWARE HUBER + THRESHOLD LOSS



WHY SEGMENT AWARE?

Previous models performed well on average but failed in the low-income group, where overestimation is most harmful.

PER SEGMENT SETTINGS

Low-income borrowers:

- Strong penalty → avoids overestimating vulnerable groups
- Smaller Huber delta → sensitive to small errors

Mid-income borrowers:

- Balanced delta and penalty → moderate conservatism

High-income borrowers:

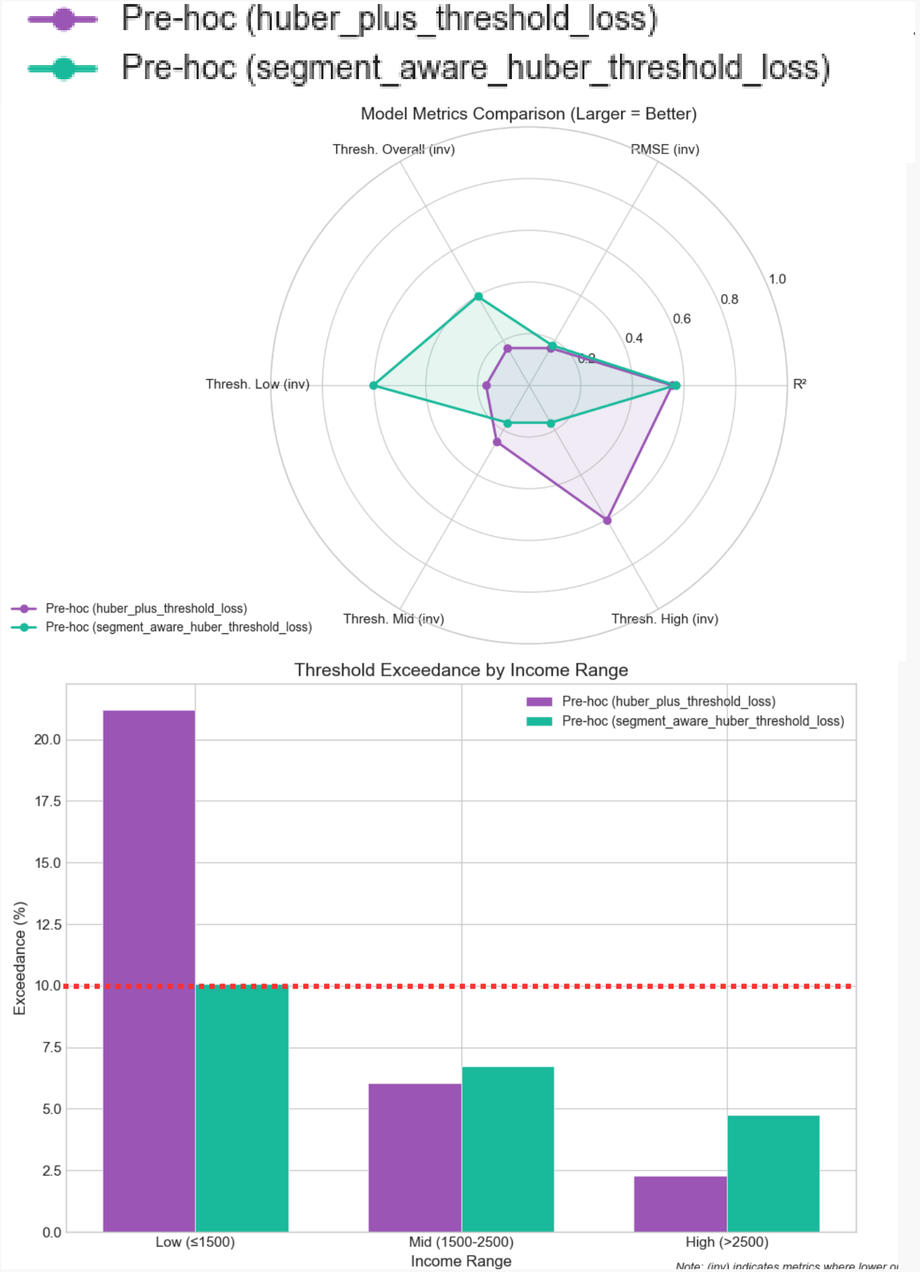
- Larger delta and weaker penalty → allows flexibility

RESULTS:

- Threshold violations drop below 10% overall
- More balanced compliance across segments

COMPARING PRE-HOC STRATEGIES

- **Segment-Aware Huber + Threshold** outperforms **Huber + Threshold** across all metrics
- Achieves better accuracy while meeting compliance
- Slightly higher violations in mid/high-income groups — but this reflects stability, not underperformance
- Ensures balanced and fair compliance across all segments
- **Conclusion**:
  - **"One-size-fits-all"** does not work
  - **Segment-aware** is the most effective and policy-aligned strategy overall
  - **Pre Hoc** Adjustment meets requirements with minimal accuracy loss
  - Incorporating **both accuracy and regulatory** objectives directly into training leads to better performance.



Thank You!

Sandro Gogaladze

DETAILED EXPLANATIONS

THESIS DOCUMENT

REPOSITORY

DATA PREPROCESSING

- Multiple preprocessing strategies were explored during development.
- The pipeline was fit only on the training set to prevent data leakage.
- The final dataset includes 25 features, 1 target, and 100,000 rows, split into 80,000 training and 20,000 test samples.

MISSING VALUE IMPUTATION

- No systematic missingness - assumed random.
- Applied K-Nearest Neighbors (KNN) imputation.

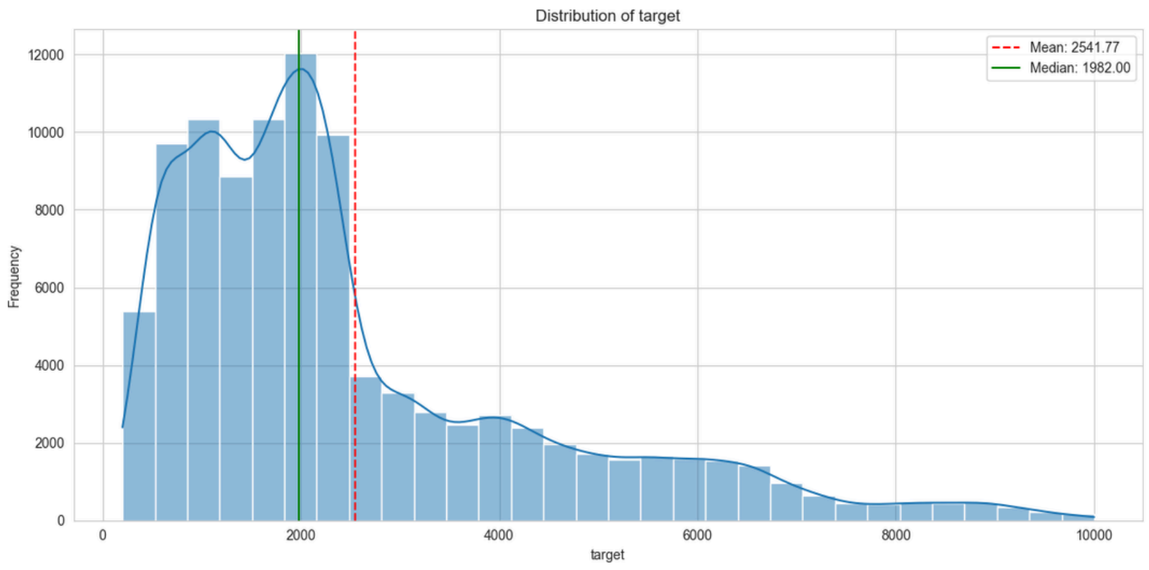
OUTLIER HANDLING

- Hybrid capping strategy based on:
  - Interquartile Range (IQR)
  - Robust Z-score thresholds
  - Percentile-based limits
- Methods chosen based on feature skewness and business relevance.

FEATURE ENGINEERING

- Created financial ratios to improve interpretability and relevance:
  - Loan-to-Income ratio
  - Balance-to-Liability ratio
  - Income Growth
  - Income-to-Liability ratio

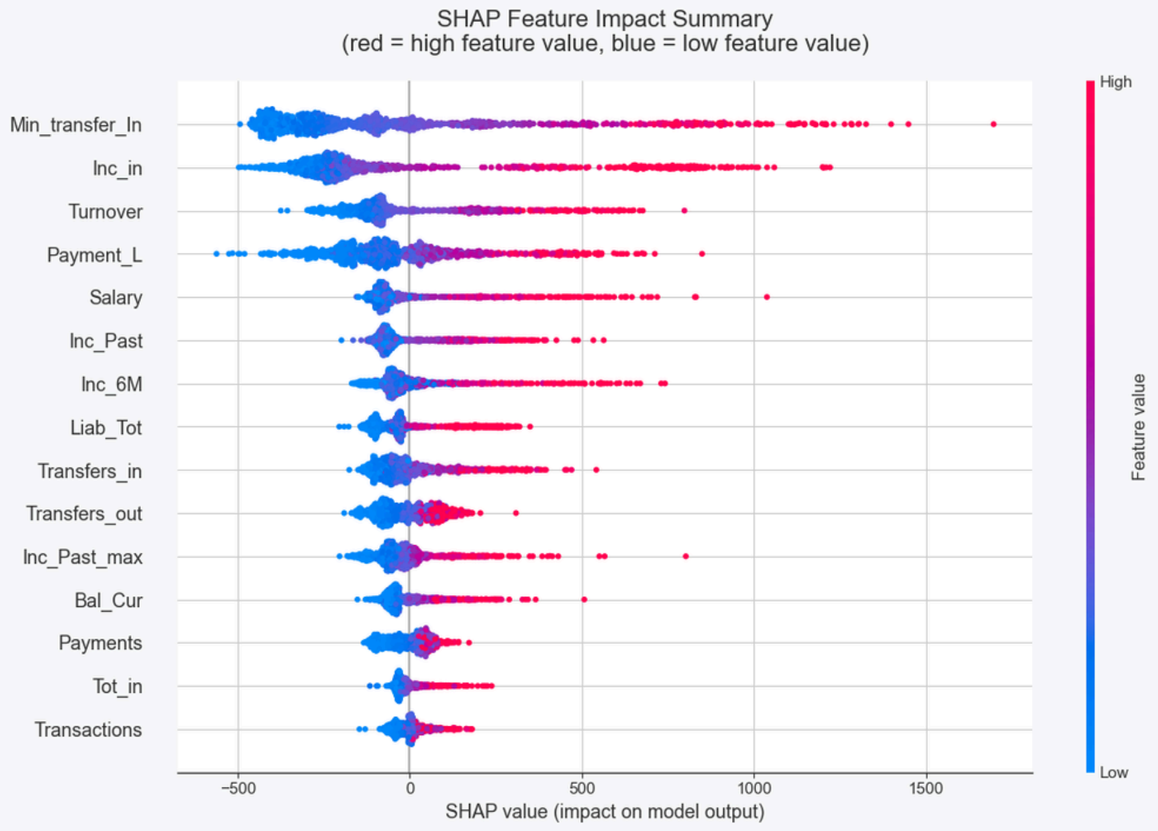
TARGET DISTRIBUTION & SEGEMTS

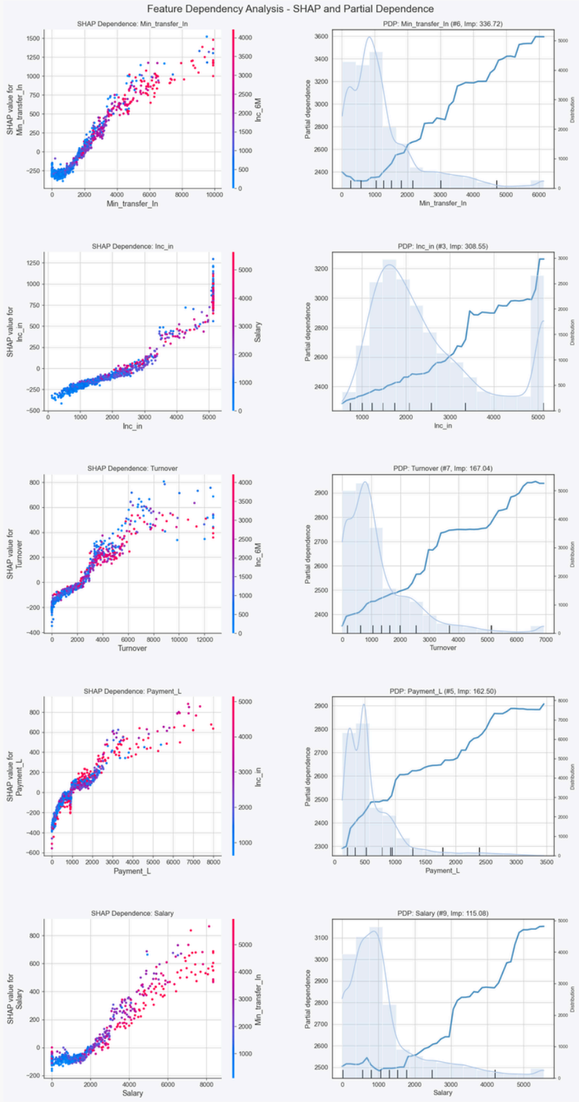


DATA FEATURES

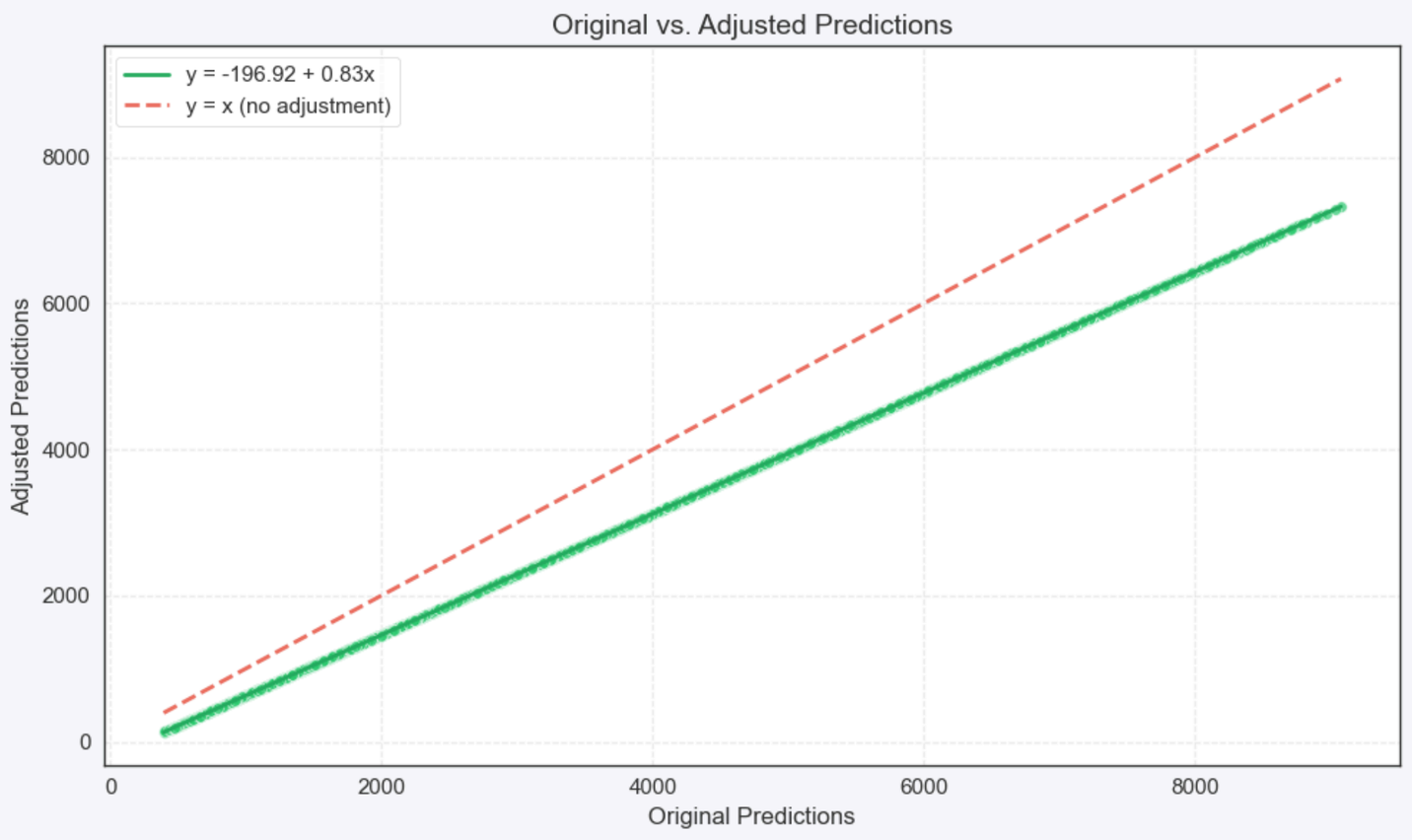
ცვლადი	აღწერა
target	სამიზნე (კლიენტის შემოსავალი)
Inc_Past	კლიენტის წარსული შემოსავალი
Liab_Tot	კლიენტის ვალდებულებები
Inc_in	კლიენტის ჩარიცხვები
Transfers_out	ანგარიშიდან ტრანსფერები
Payment_L	ყოველთვიური გადასახადი
Min_transfer_In	კლიენტის მინიმალური ჩარიცხვები
Turnover	ანგარიშზე ბრუნვა
Transactions	ანგარიშიდან ტრანზაქციები
Salary	ანაზღაურება
Inc_6M	6 თვის ჩარიცხვები
Acct_Trns	ანგარიშიდან გადარიცხვები
Transfers_in	ანგარიშზე შესული თანხები
Trn_max	ანგარიშიდან გადარიცხვების მაქსიმუმი
Balance	ანგარიშზე ნაშთი
Loan	სესხის თანხა
Payments	გადახდები
Inc_Past_avg	კლიენტის წარსული საშუალო შემოსავალი
Inc_Past_max	კლიენტის წარსული მაქსიმალური შემოსავალი
Tot_in	მთლიანი ჩარიცხვები ანგარიშზე
Bal_Cur	მიმდინარე ბალანსი
Loan_Cnt	სესხის რაოდენობა

BASELINE MODEL





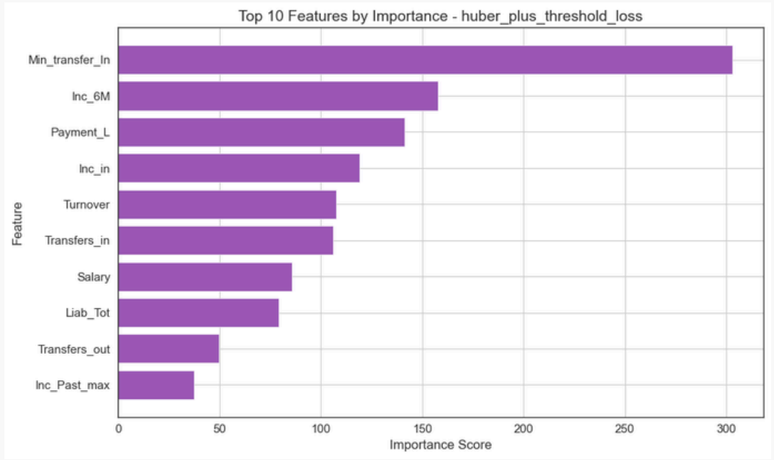
QUANTILE CALIBRATION

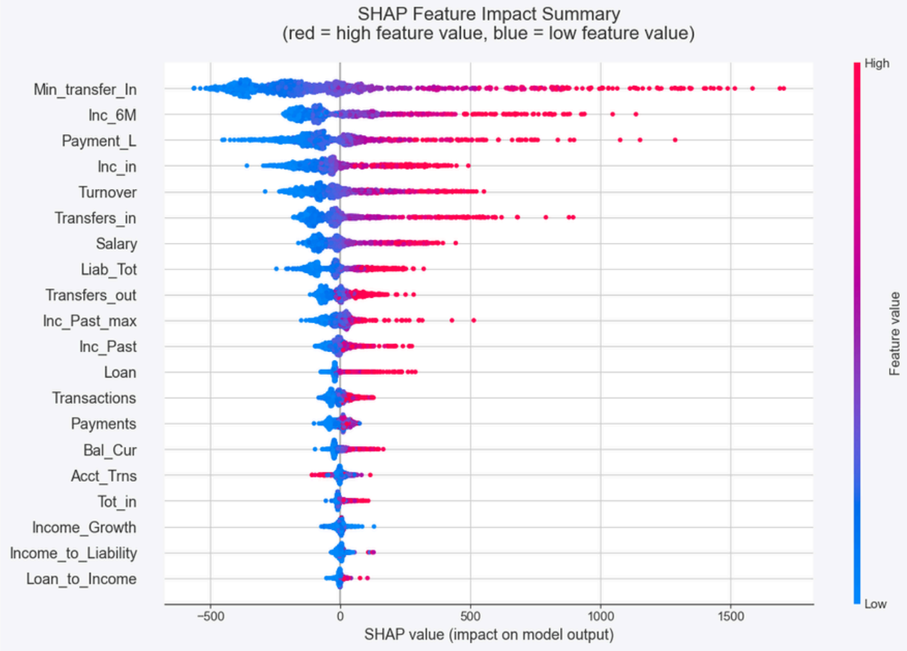


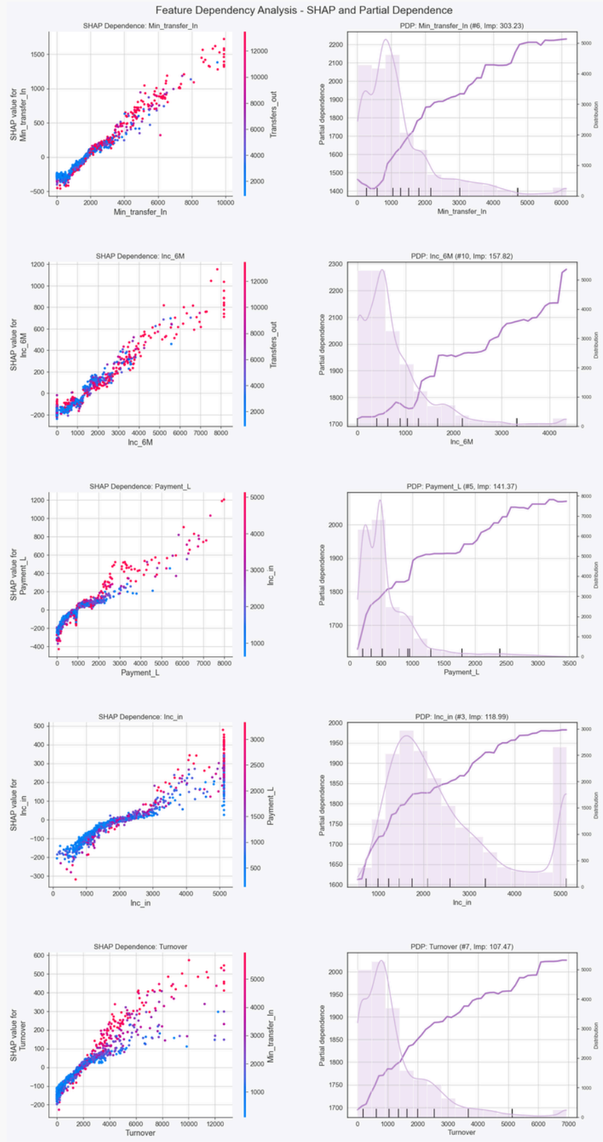
HUBER + THRESHOLD LOSS

$$\mathcal{L}_\delta(\hat{y}_i, y_i) = \begin{cases} (\hat{y}_i - y_i)^2, & \text{if } |\hat{y}_i - y_i| \le \delta \\ 2\delta \cdot |\hat{y}_i - y_i| - \delta^2, & \text{otherwise} \end{cases}$$

$$\mathcal{L}_i = \mathcal{L}_\delta(\hat{y}_i, y_i) + \lambda \cdot \mathbf{1}_{[\hat{y}_i > y_i + T_i]} \cdot (\hat{y}_i - y_i - T_i)^2$$







SEGMENT-AWARE HUBER + THRESHOLD LOSS

$$\mathcal{L}_{\text{Huber,s}} = \begin{cases} (\hat{y}_i - y_i)^2, & \text{if } |\hat{y}_i - y_i| \leq \delta_s \\ 2\delta_s \cdot |\hat{y}_i - y_i| - \delta_s{}^2, & \text{otherwise} \end{cases}$$

$$\mathcal{L}_{\text{Penalty,s}} = \lambda_s \cdot (\hat{y}_i - y_i - T_i)^2 \cdot \mathbf{1}_{[\hat{y}_i > y_i + T_i]}$$

$$\mathcal{L}_s = \mathcal{L}_{\text{Huber,s}} + \mathcal{L}_{\text{Penalty,s}}$$