Fintech

Final Project

Needs Based Recommendation Systems Collaborative Filtering and NBA

Team 15

Mehdi Ghiasipour - 10976638
Reza Ilzadi Jahromi - 10974659
Hirdesh Kumar - 10997383
Suraj Manickam Thirumalai Raja - 11076161
Christos Partasidis - 11045090

Date: 25 June 2025

POLITECNICO
MILANO 1863

Deliverables

Jupyter Notebook

The sections are:
- Ingestion
- EDA
- Feature Engineering
- CV + test
- NBA logic

acommpanied with comments for easy audit.

README

Provides instructions for use with regard to how files interact (Jupyter Notebook + Excel file with data)

PDF deck

Follows the course instructions structure:
- WHAT
- WHY
- HOW
- KEY RESULTS

WHAT

What have you done?

Project Overview

Objective

Estimate clients' need based on their personal and financial data and subsequently recommend a suitable financial product

Data

- ~ 5000 anonymised retail-bank client.
- 7 behavioral / financial inputs (Age, Gender, FamilyMember, FinancialEducation, RiskPropensity, Income, Wealth).
- 2 implicit labels.

Method

Two seperate **Decision Tree** models, one for income, and one for accumulation, both enhanced with engineered features.

Product catalogue

Eleven instruments tagged by **Type** and **Risk score**.

WHY

What problem is solved?

Business Context

Need first vs product first gap

Clients think in goals “protect my salary”, “grow my nest egg” while traditional advice flows from products to needs. This gap erodes trust and uptake.

Scaling and Bias Challenge

Advisors already spend significant time per client each year on meetings & prep. Manual needs assessment can’t keep up. [R2]
Historical product pushes reflect advisor incentives, not always genuine client goals thus driving mishmatches.

Hyper Personalisation And Regulation

Wealth management is shifting on always-on service that accurately spots needs. By 2030 it is estimated that 80% of new wealth manament client will be data driven advice. [R1]
Suitability rules also require evidence that advice maps to client needs.

Our Answer

Automate need detection and product matching to deliver explainable, always-on, bias-free recommendations at scale.

HOW

How you did it?

Data & Labels

Snapshot

- **Rows:** 5000 clients.
- **Raw features:** Age, Gender, FamilyMembers, FinancialEducation, RiskPropensity, Income (€ k), Wealth (€ k).
- **Engineered features:** Gender x Age, log (Income/Wealth).
- **Targets:** IncomeInvestment, AccumulationInvestment.

Data quality

- Missing values well below 2 % across all columns.
- Income & Wealth highly skewed → log transformed (Income_log, Wealth_log) before scaling.
- Extreme Wealth outliers trimmed via IQR rule

Class balance

- IncomeInvestment has ≈ **38** % positives, while AccumulationInvestment has ≈ **51** % positives (as shown in the two histograms). [A1] [A2]

* [A1] [A2] can be found in the appendix

Pipeline Diagram

1. Data ingestion

- Read Excel sheet
- 5000 clients
  7 raw features
  2 targets (income, accumulation)

2. Data prep & EDA

- Drop non-informative ID column
- < 2 % missing → median/mode imputation
- Histograms for class balance, box/violin plots for outliers
- Log transform Wealth and Income

3. Feature engineering

- Wealth_log, Income_log → de_skew
- Engineer Gender x Age
- Engineer log(Income/Wealth)
- Scaling with MinMaxScaler

4. Data segregation

- Split train (80%) and test (20%) sets
- 5-fold CV on the 80 % train set
- Same split reused for both need targets

Pipeline Diagram

5. Training & validation

Decision Tree (depth = 5) per need

Primary metric: F1

Secondary metrics: accuracy, precision, recall

CV mean ± std stored in models_results

6. Explainability

Gini feature-importance bar chart from the ‘Decision Trees’

Partial-Dependence Plots on top drivers

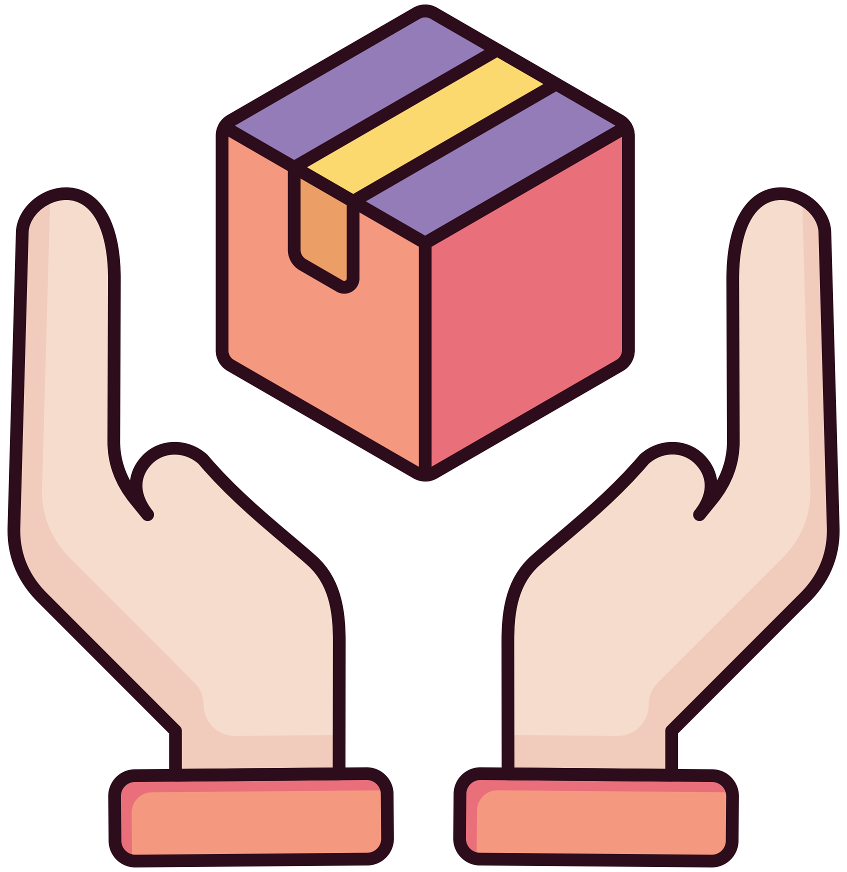
7. Next Best Action

Filter product list by Type
0 = Income
1 = Accumulation

Keep products with ProductRisk ≤ ClientRiskPropensity

Recommended highest-risk eligible product

Output:
{client_id, need_probs, recommended_product_id}



Feature Engineering

Base numeric set (after log + scaling)

Age, Gender, FamilyMembers, FinancialEducation, RiskPropensity, Wealth_log, Income_log.

Engineered features added in prepare_features()

Gender x Age captures life stage differences that vary by gender.

log(Income / Wealth) liquidity vs asset stock, normalised, avoids divide by zero.

Why they matter?

- Interaction term lets the tree split differently. For example, young males vs. older females.
- Liquidity ratio highlights clients who earn a lot relative to their assets (Accumulation focus) vs. high-net-worth retirees (Income focus).

Takeaway

Just two domain-informed features added to deliver sizable accuracy gains while keeping the model fully interpretable.

Need target	F1 baseline (raw features)	F1 + engineered	Δ F1
IncomeInvestment	0 .689	**0 .729**	+ 0 .04
AccumulationInvestment	0 .680	**0 .818**	+ 0 .14

Model & Validation

Model

- **One-vs-all Decision Tree** (depth = 5) per need
- Simple, fully explainable

Training setup

- 80 % train / 20 % hold-out test
- 5-fold cross-validation on the train split
- Primary metric = F1 (balances precision & recall under class-imbalance)

Results

Need target	CV F1 (mean ± std)	Test F1	Test Precision	Test Recall
IncomeInvestment	0.75 ± 0.02	**0 .729**	0.77	0.79
AccumulationInvestment	0.81 ± 0.03	**0 .818**	0.7	0.72

- Engineered features raise test F1 by +0.04 (Income) and +0.14 (Accum) over the raw-feature baseline.

Key takeaways

- Consistent CV vs test scores → model generalises well
- Decision Trees remain interpretable
- Feature importance and PDPs make reasoning transparent for compliance terms

Explainability & NBA

Global feature importance (Decision Tree Gini)

- Wealth_log - largest information gain
- RiskPropensity - client-stated risk appetite
- Income_log - liquidity indicator
- Gender x Age - life-stage interaction
- log(Income / Wealth) - relative earning power

Refer to [A9] [A10] for quantitative analysis

Partial Dependence highlights

Target	Critical/ Knee Points	Interpretation
AccumulationInvestment	Wealth_log > 0.65 or Income_Wealth_Ratio_log > 0.12	Affluent clients or those earning far more than they currently invest gravitate to growth product
IncomeInvestment	Age > 0.60 (≈ 60 yrs) with average liquidity	Retirees prioritise stable income streams

[A3] [A4]

Next Best Action (NBA) rule

1. **Pick need:** If predicted probability (>0.5) for either need Income or Accumulation, need is set to 1 (both of them can be 1 too).
2. **Filter catalogue**: Type 0 list for Income, Type 1 list for Accumulation
3. **Recommend:** highest-risk product within eligible set

Suitability Check

Scatter of ClientRiskPropensity vs ProductRisk shows all points below the line which means that every recommendation respects the MiFID risk cap. [A5] [A6]

- **Coverage:** 241 Income-need clients → 227 matched (94 %)
- **Risk cap breaches:** 0 (all recommended dots below the diagonal)

* [AX] can be found in the appendix

KEY ASPECTS

Any key aspects to underline, both of business and method

Methodological Edge

Explainability first

- Interpretable Decision Trees, feature importance bars & PDPs make every split auditable.
- Compliance reviewers can trace a recommendation from raw input → leaf node.

Big lift, tiny change

- Just two engineered features increase decently the results
- No heavy ensembles or hyper-param sweeps required.

MiFID & IDD alignment

- Captures and stores the evidence trail required by Art. 54/55 (client objectives, risk tolerance, rationale) by putting transparency and explanability first.

Tried streaming

- Evaluated Hoeffding Tree, EFDT, GNB, KNN, Perceptron, Passive-Aggressive (concept-drift ready).
- Static snapshot + higher batch accuracy → Decision Tree the pragmatic choice.

Risk logic decoupled

- NBA rule enforces ProductRisk ≤ ClientRiskPropensity after model scoring, separating statistical error from regulatory breach.

Net result

- robust uplift, full transparency, and deployment within tight compute budgets without sacrificing regulatory comfort.

KEY RESULTS

What results you got?

Accuracy Snapshot

Need target	Baseline F1*	F1 + engineered	Δ F1	Precision	Recall
IncomeInvestment	0 .689	0 .729	+ 0 .040	0 .77	0 .79
AccumulationInvestment	0 .680	0 .818	+ 0 .138	0 .70	0 .72

- Engineered features deliver the bulk of the uplift, no ensemble needed.
- Low Cross-validation std. → The model captures the true signal from the given data
- Precision/Recall balance is healthy for both targets, confirming F1 gains are not due to skew in a single metric.

Recommendation Coverage

Need target	Need client predicted	Clients matched	Coverage
IncomeInvestment	241	227	**94%**
AccumulationInvestment	475	357	**75%**

Accumulation Need (Top3) [A7]

1. Balanced Mutual fund (Risk 0.41) - 84 recs
2. Cautious Allocation Segregated Account - 98 recs
3. Balanced Mutual fund (Risk 0.55) - 62 recs

Income Need (Top 3) [A8]

1. Balanced High Dividend Mutual fund - 79 recs
2. Fixed Income Segregated Account - 74 recs
3. Income Conservative Unit Linked - 71 recs

Takeaway

The engine supplies a suitable product for 9 clients out of 10 with Income needs and 3 out of 4 with Accumulation needs, automatically respecting each client's risk threshold.

Take-aways

Personalised Advice

- Predicts each client's Income vs Accumulation need and surfaces a product that fits both the goal and the client's risk appetite.
- **Impact** → higher relevance, better client satisfaction.

Regulatory Compliance

- Suitability rule baked in: ProductRisk ≤ ClientRiskPropensity.
- Suitability scatter plots confirm 0 breaches across 5000 clients.
- **Impact** → MiFID II & IDD evidence generated automatically.

Operational Efficiency

- Fast Need detection + matching run in; 94 % Income-need and 75 % Accumulation need coverage.
- **Impact** → frees advisors from manual checks, saving precious time.

Lightweight Deployment

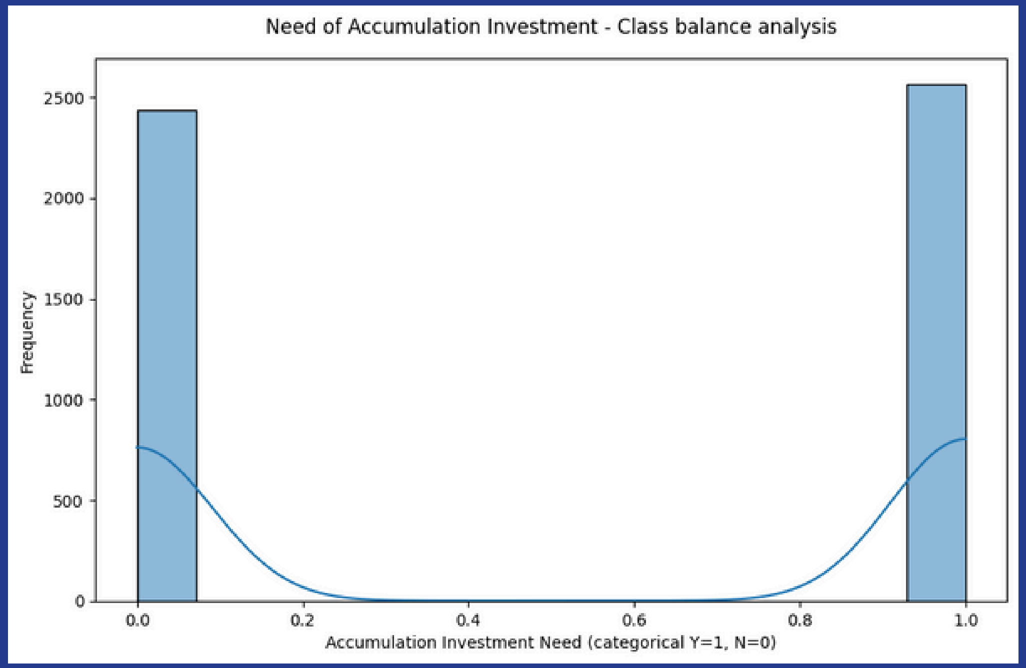
- Lightweight single Decision-Tree per need + two engineered features lift F1 by up to +14 pp.
- No complex ensembles or cloud GPUs required.
- **Impact** → easy to integrate in CRM or serverless function.

References

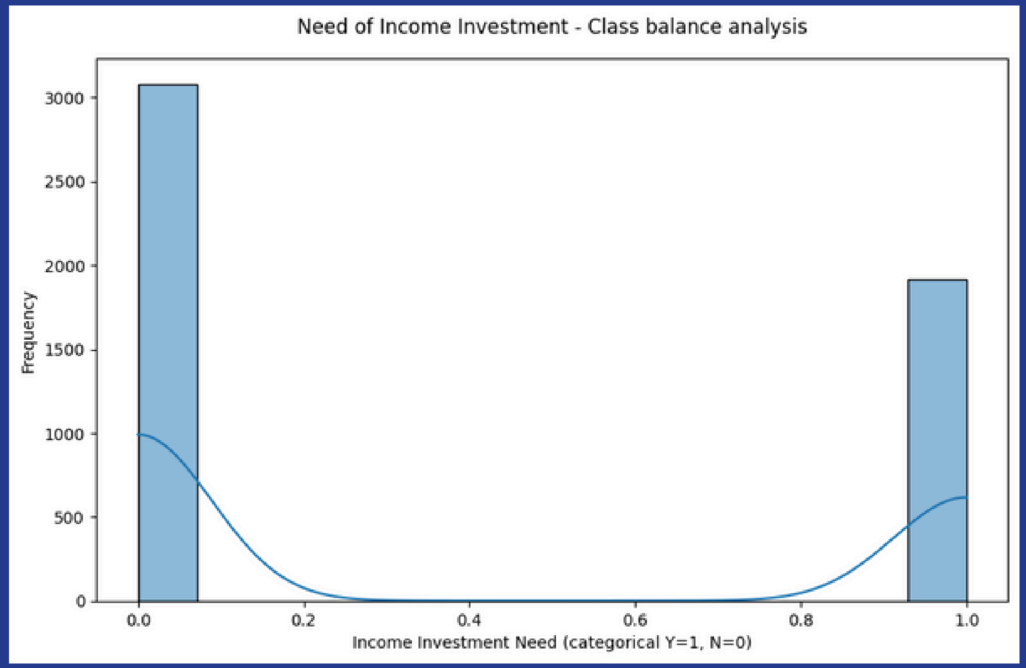
[R1] P. Baghai, A. D'Amico, R. de la Roche-Zhu, O. Erzan, V. Golyk, and J. Zucker, "On the cusp of change: North American wealth management in 2030," McKinsey & Company, Jan. 22, 2020. [Online]. Available: https://www.mckinsey.com/industries/financial-services/our-insights/on-the-cusp-of-change-north-american-wealth-management-in-2030.

[R2] M. Kitces, "How do financial advisors actually spend their time and the limitations of productivity?," Nerd's Eye View, Kitces.com, Mar. 18, 2019. [Online]. Available: https://www.kitces.com/blog/how-do-financial-advisors-spend-time-research-study-productivity-capacity-efficiency/

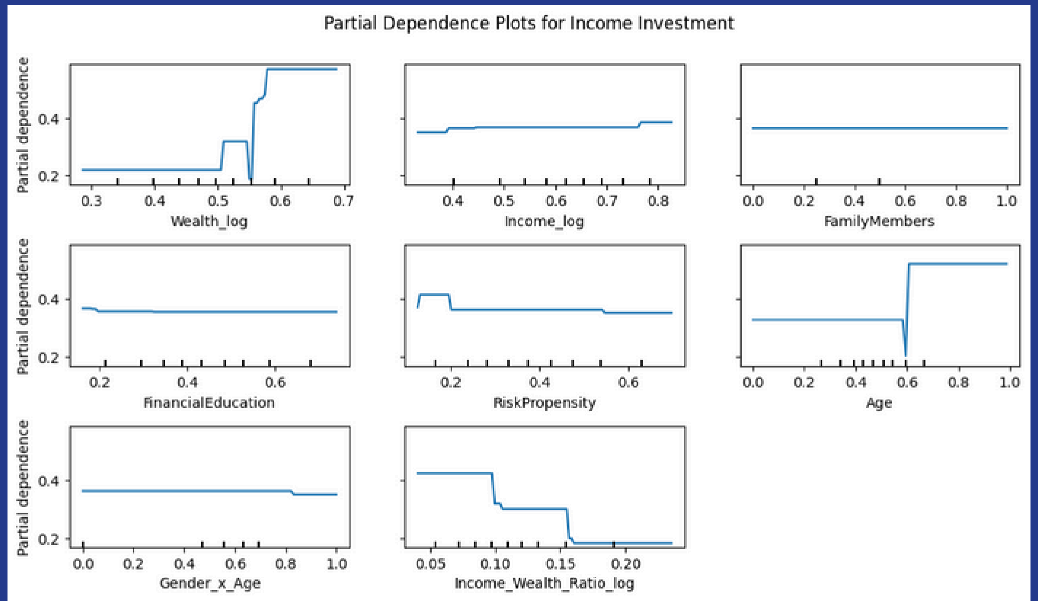
Appendix



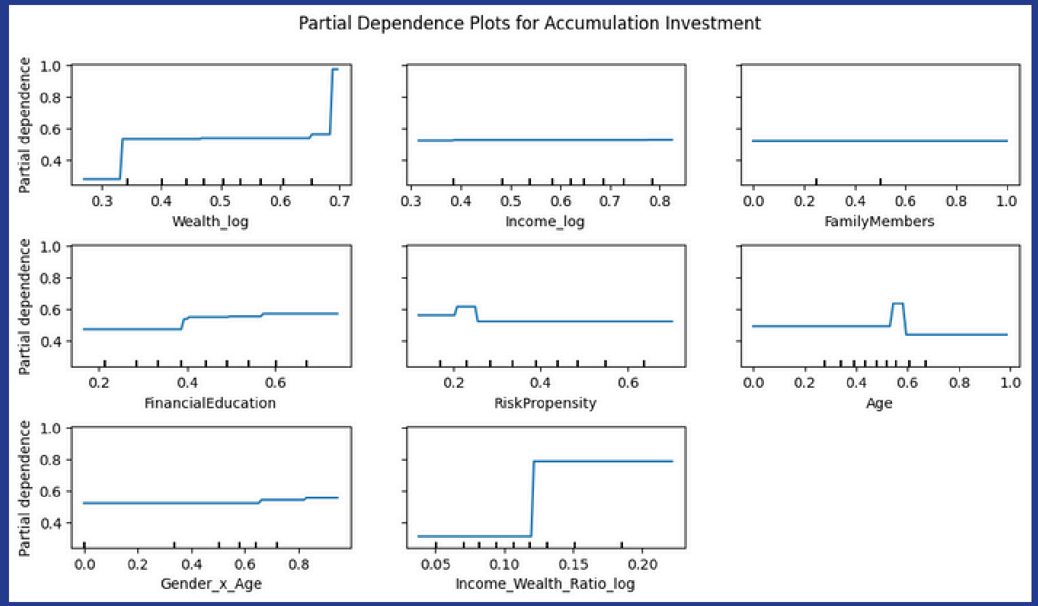
[A1]



[A2]

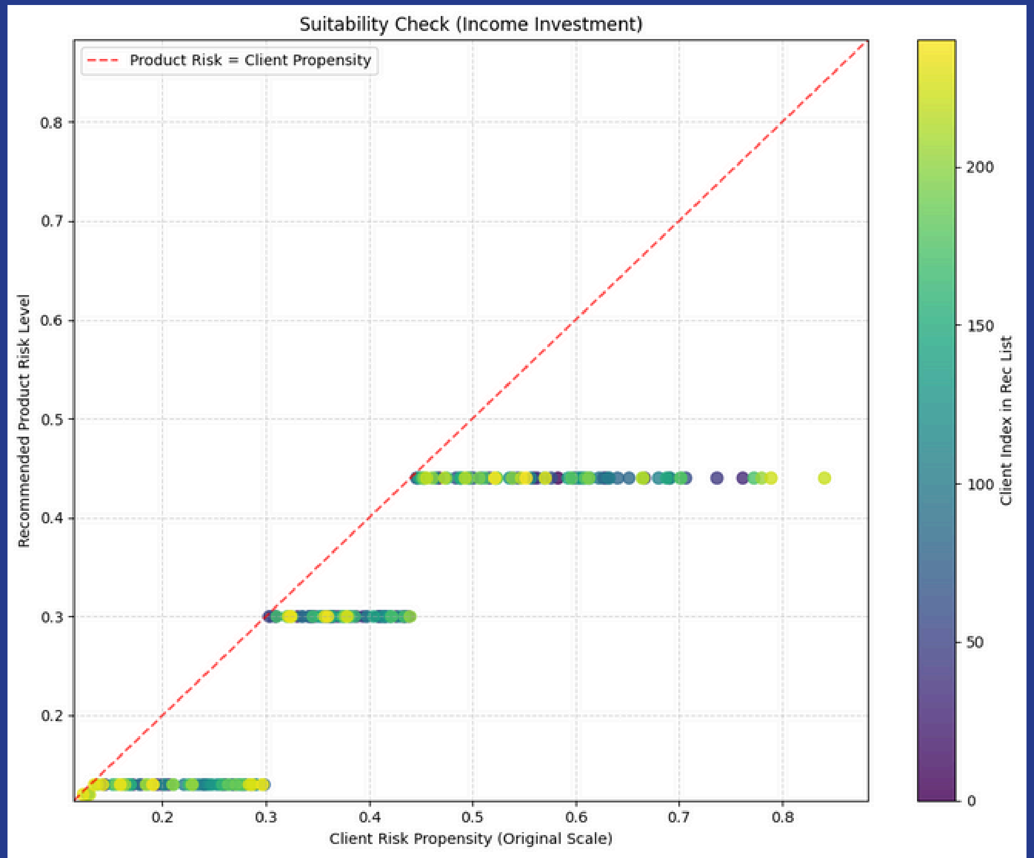


[A3]

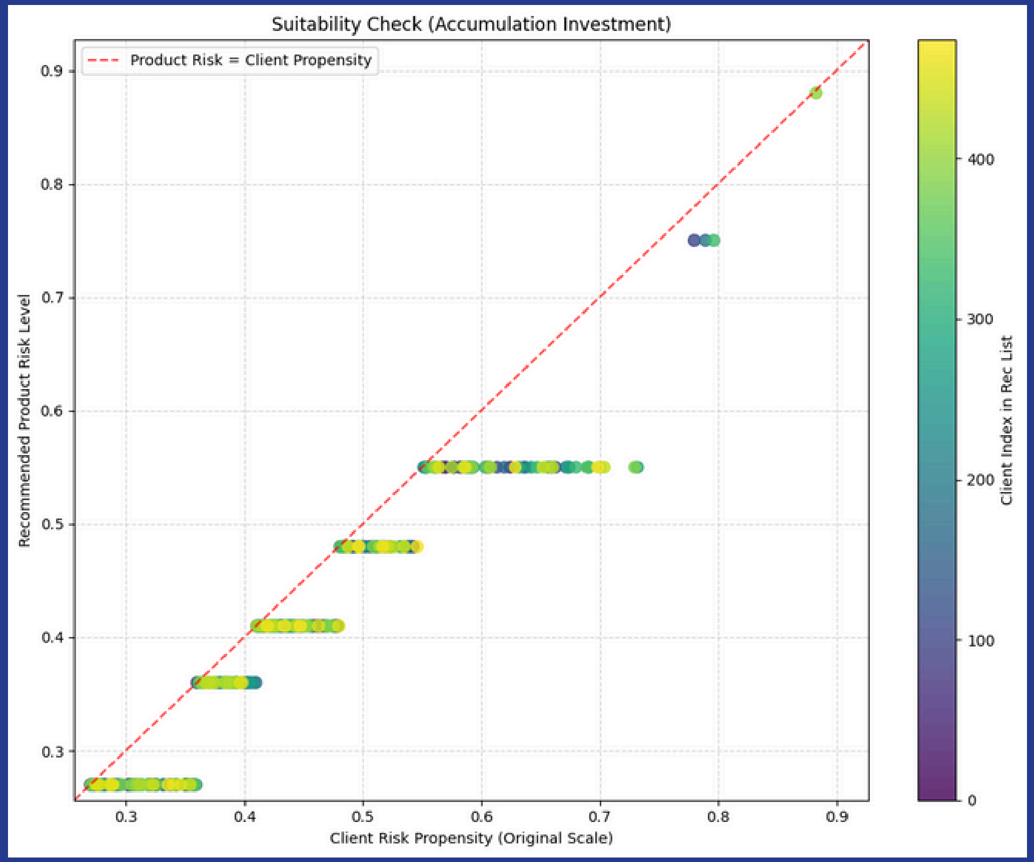


[A4]

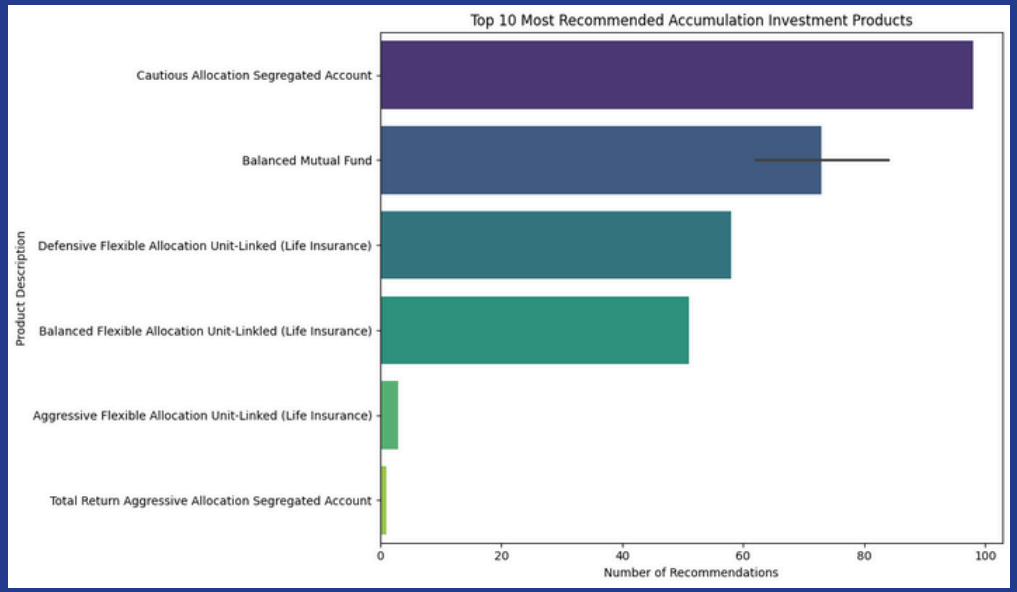
Appendix



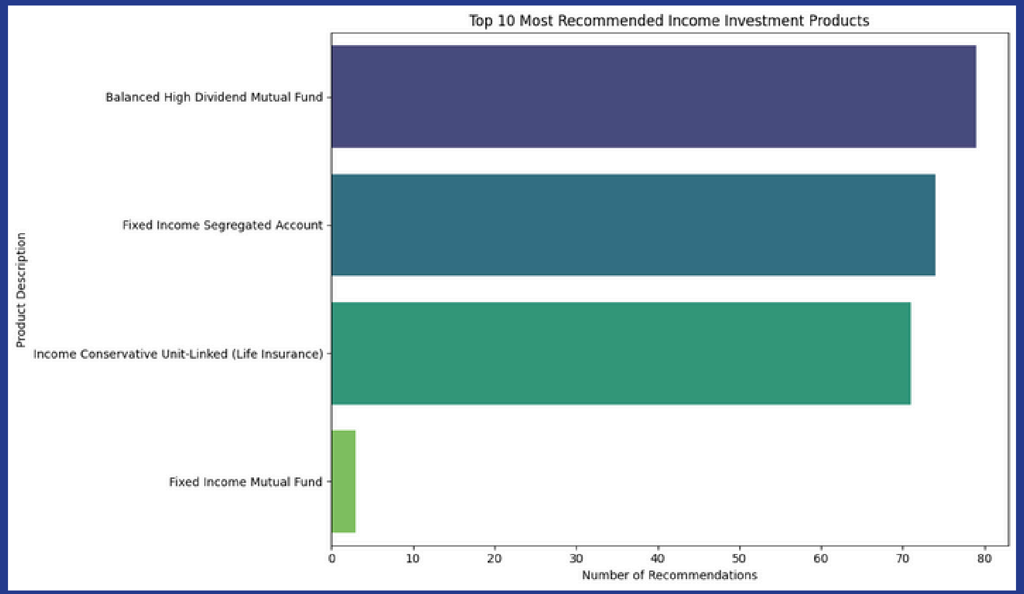
[A5]



[A6]

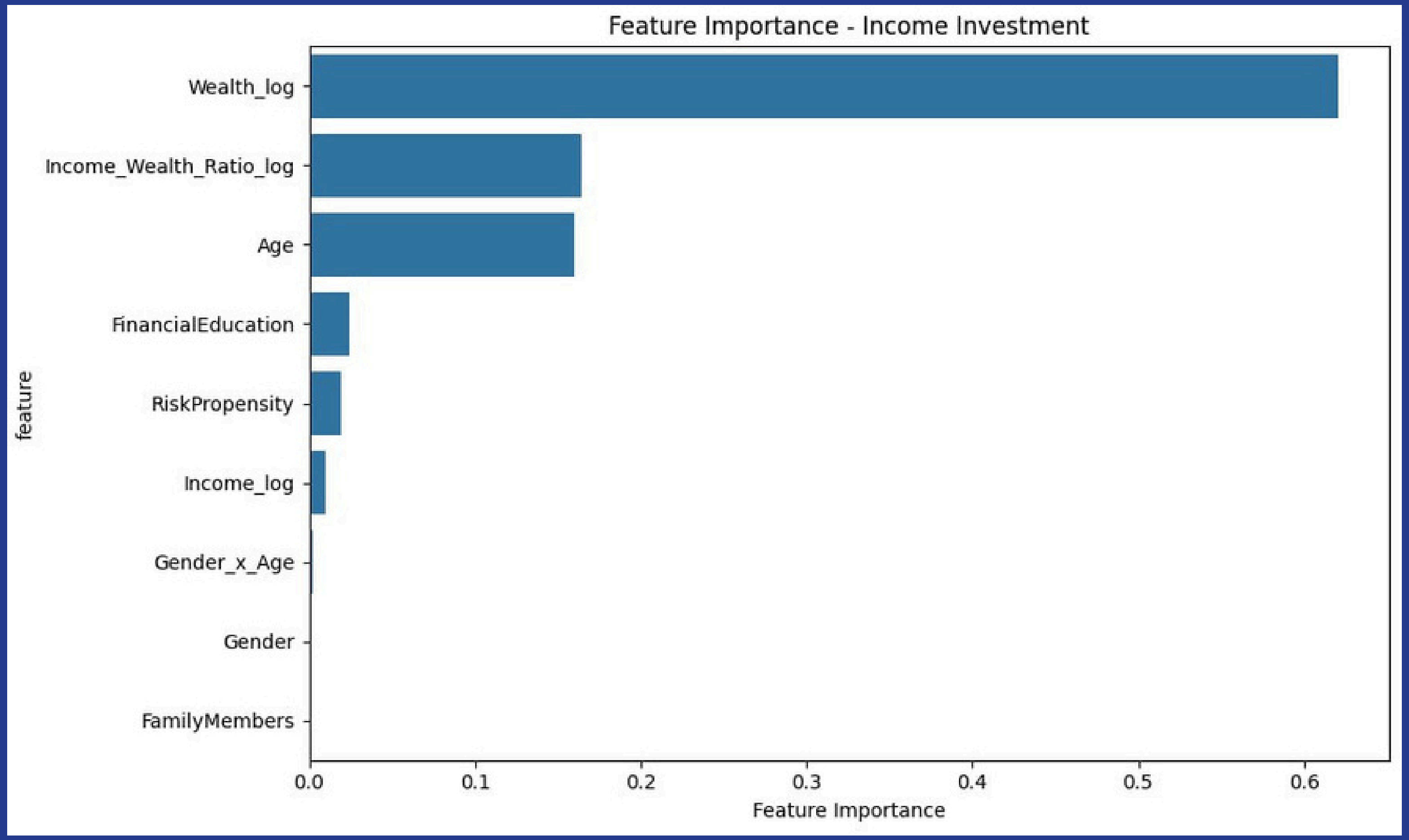


[A7]

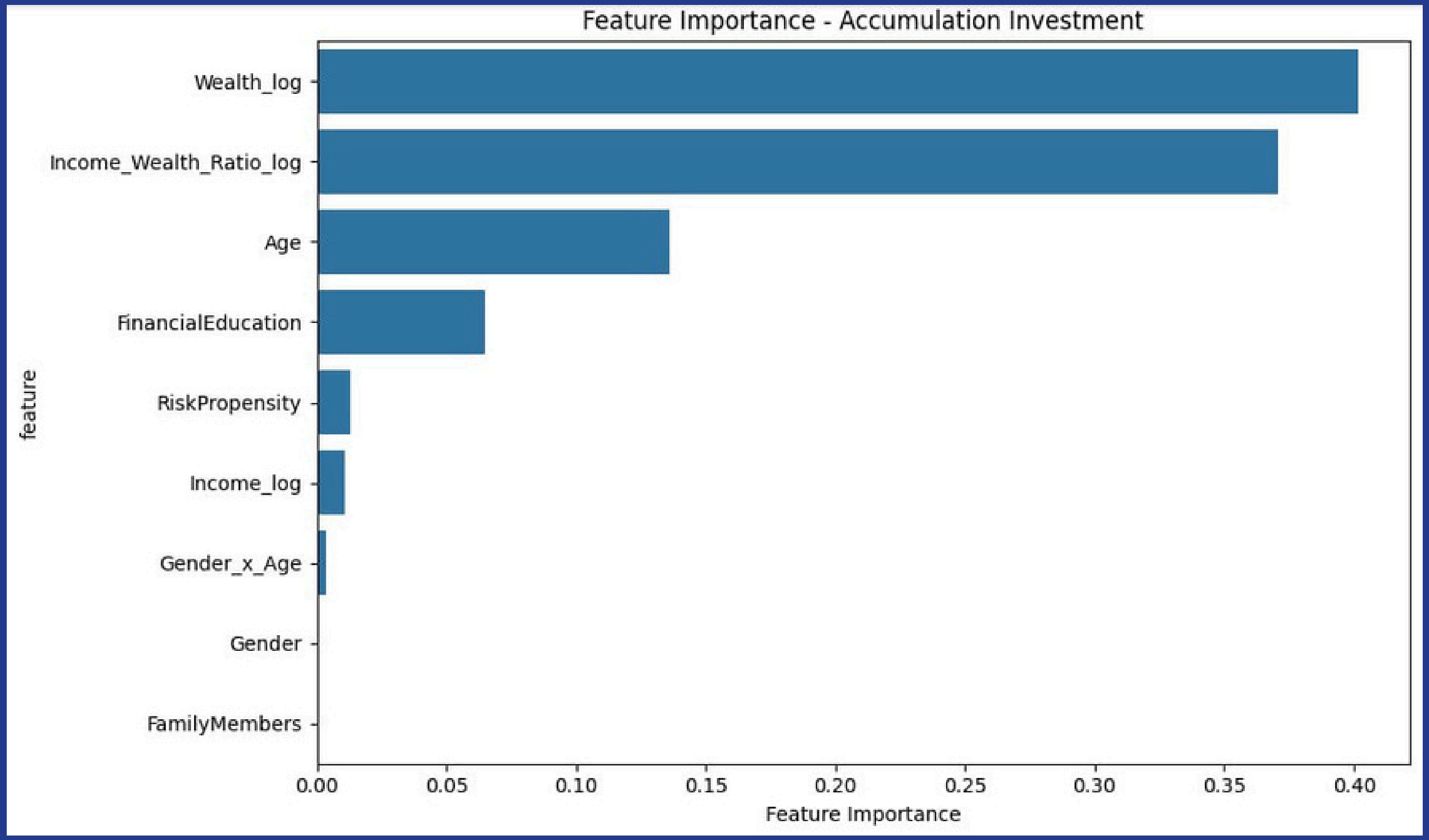


[A8]

Appendix



[A9]



[A10]