



Code to Cash :

Cracking the Patterns Behind Financial Goals Success

Trends Shaping the Future
of Financial Goals







CONTENT



- 1. Business Understanding
- 2. Data Understanding
- Data Preparation
- Modeling
- Evaluation

- 
- 
- In today's economy, financial success is shaped by more than just income, it's influenced by who we are, how we behave, and how we think.
 - This project aims to uncover the behavioral, demographic, and financial factors that influence individual financial outcomes. Using an encoded dataset representing traits such as risk attitude, behavioral intent, and investment preferences, we will model and interpret the drivers of financial goal achievement. The insights will guide targeted financial advice, behavior based segmentation, and future product design in finance and fintech industries
 - The insights generated can inform personalized advisory tools and smarter financial decision systems.

Business Objectives

- **Primary Objective:** To predict and explain financial goal outcomes (GOAL) based on personal characteristics and behaviors.

Sub-Objectives:

- - Identify the strongest behavioral and demographic predictors of successful financial outcomes.
- - Build interpretable models to enable personalized financial advice.
- - Provide actionable insights to segment individuals based on:
 - (a) Investment preferences
 - (b) Risk tolerance
 - (c) Behavioral intent

Framework: Theory of Planned Behavior (TPB)

Outcome Variable: GOAL (representing financial achievement)

Methodology: CRISP-DM framework (Data prep, modeling, evaluation).

Approach: Binary classification using Logistic Regression, Random Forest, and XGBoost.



Stakeholders



- **Banks & Fintech Firms** - Use insights for better financial product targeting
- **Behavioral Researchers** - Understand how behavior affects financial success
- **Data Scientists** - Apply machine learning on real world behavior modeling
- **Consumers** - Receive better financial advice via personalized systems
- **Educators** - Integrate findings into financial literacy training programs

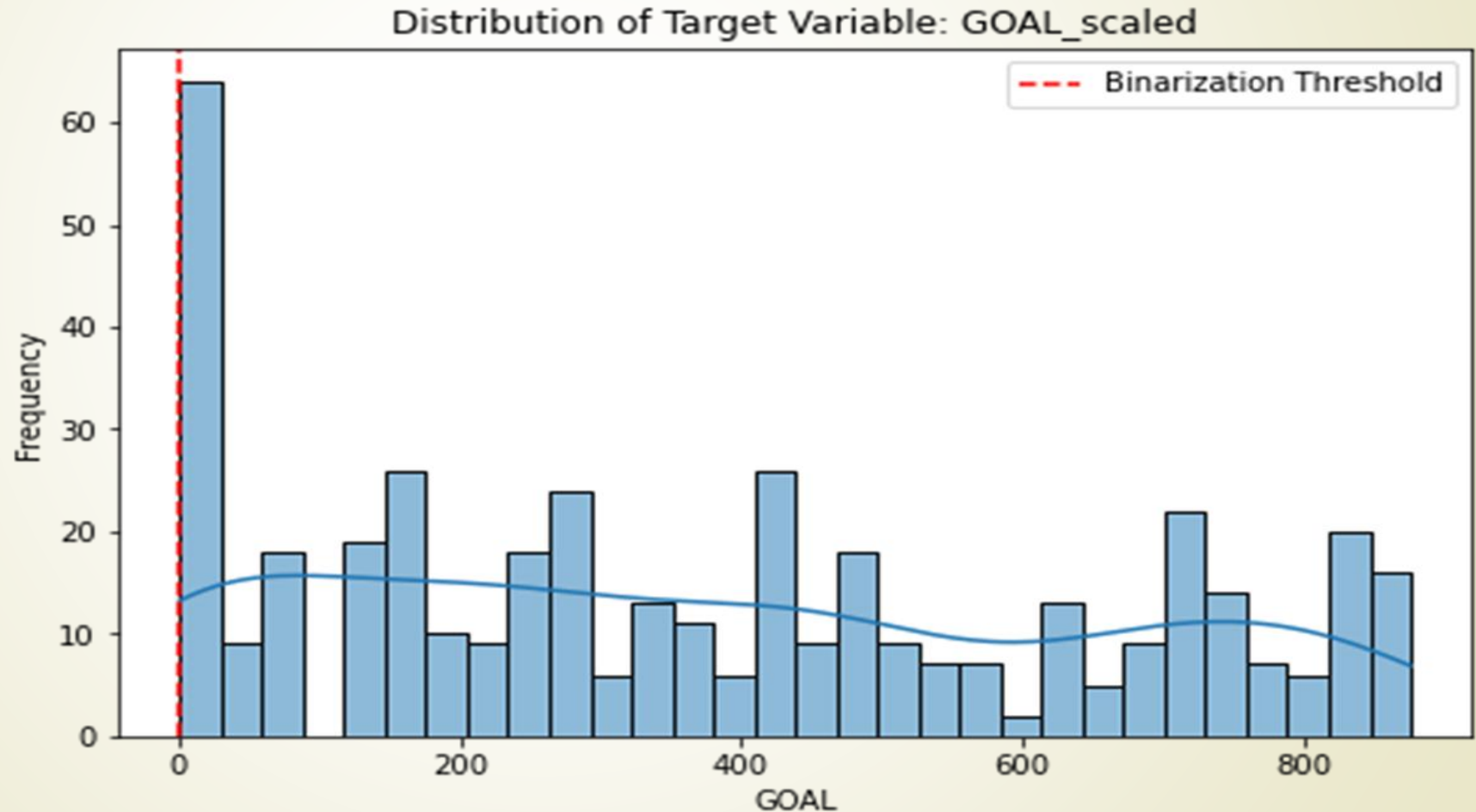


Data Understanding

- ▶ We explore the dataset structure, assess data quality, identify missing values, examine distributions, and detect potential anomalies.
- ▶ Objective : To explore, summarize, and detect structure or issues in the data
- ▶ The dataset contains 423 rows and 27 columns
- ▶ The features appear to be a mix of numerical (Likert-scale) responses and categorical variables (e.g., GENDER, EDUCATION).
- ▶ Target Variable: GOAL is identified as the primary dependent variable for prediction and behavioral inference.

Target Variable Analysis (GOAL_scaled)

- Understanding the central tendency, skewness, and behavior cutoff points.



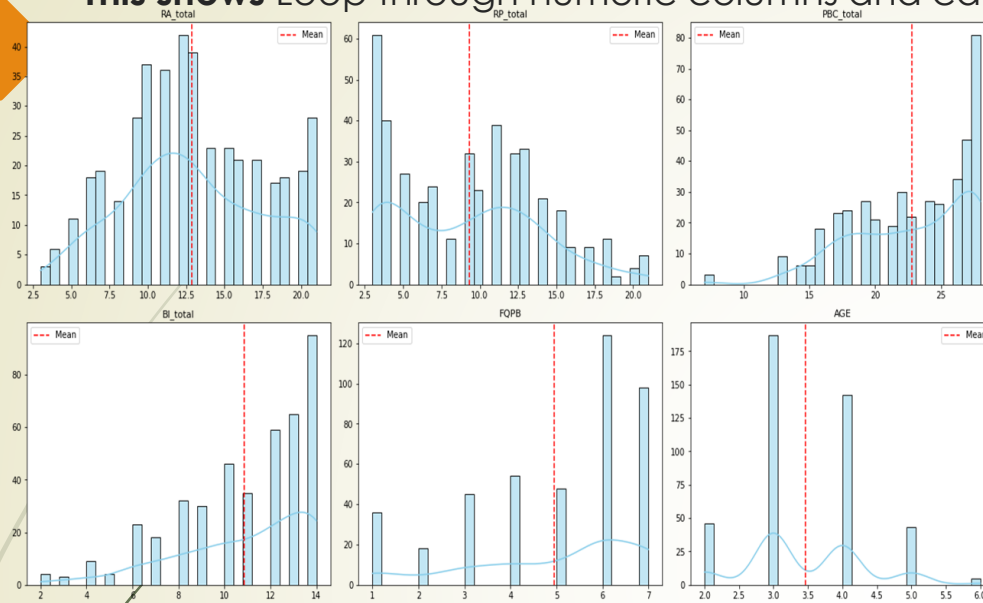


Missing Values Report

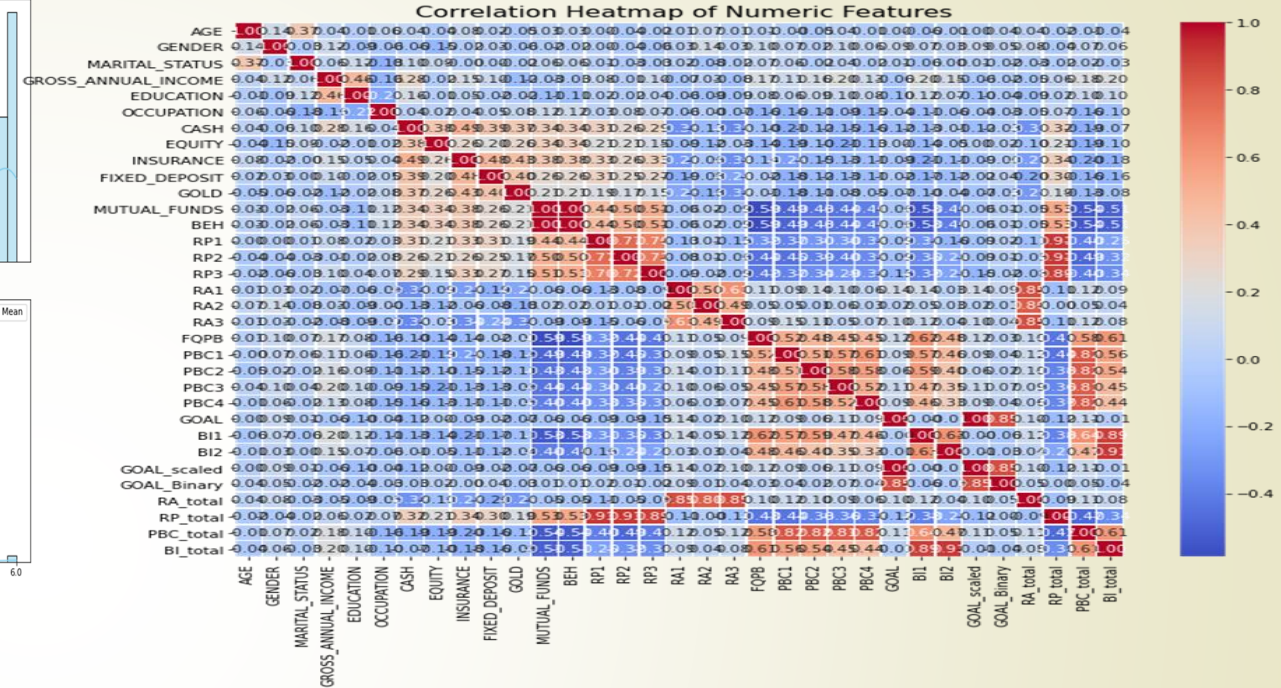
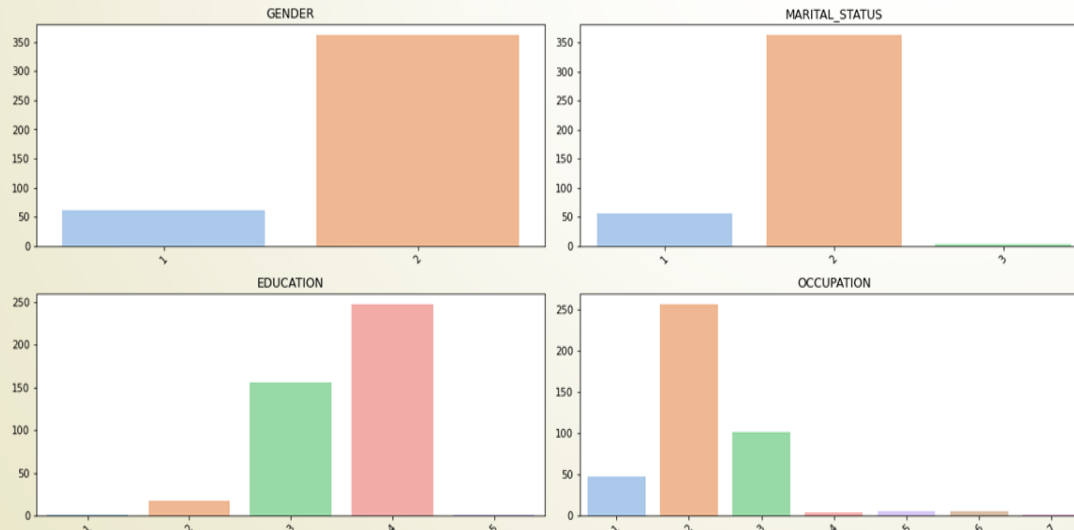
- Why: Avoid silent row drops. Evaluate imputation or domain-informed handling instead of blind deletion.
- - All 423 rows are complete. - Therefore we can proceed without imputation or dropping rows
- A comprehensive missing value check showed: - Missingness is low and not patterned, implying MCAR (Missing Completely At Random).
- - Simple imputation or row dropping may suffice depending on model sensitivity.

Feature Distribution Plots

This shows Loop through numeric columns and categorical columns

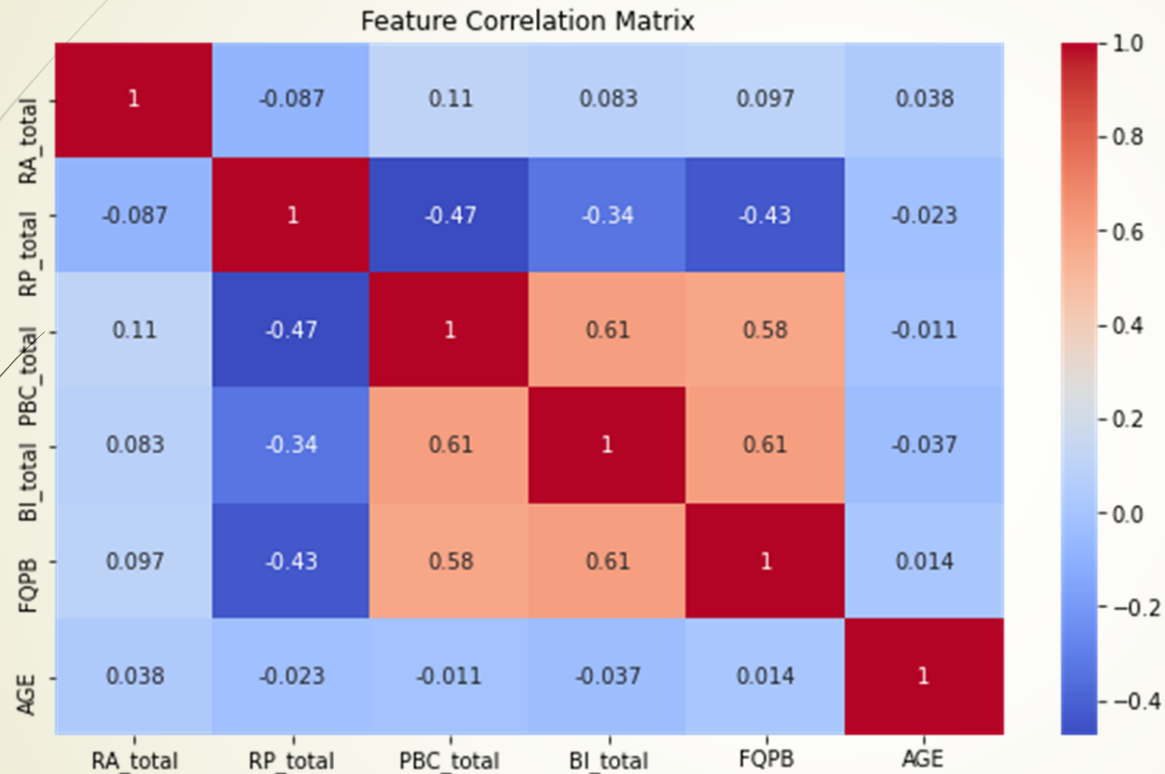


Count Plots of Categorical Features



Correlation Analysis

- Identify collinearity and potential predictors



Summary Analysis of Data Understanding

- The Data Understanding phase provided a comprehensive initial analysis of the dataset, revealing clean and structured data primarily composed of Likert-scale behavioral variables and a few categorical fields like GENDER and EDUCATION.
- Descriptive statistics confirmed expected central tendencies, while skewness and kurtosis analyses indicated mild asymmetry and flat distributions in key features such as GOAL.
- Outlier detection flagged a small number of high GOAL scores, likely valid responses rather than anomalies.
- A heatmap of the correlation matrix revealed strong relationships among behavioral constructs (e.g., BI, PBC, INT), highlighting potential multicollinearity concerns to be addressed during modeling.
- Categorical variables showed meaningful distributions, with education level correlating positively with GOAL-setting.
- Overall, the dataset is of high quality and ready for preprocessing, with actionable insights supporting the relevance of behavioral constructs in predicting goal-setting behavior.

Data Preparation

Objective: Convert raw, messy, misaligned, incomplete, possibly biased data into a purified modeling-ready dataset well structured, encoded, clean, and insightful.

Goal: Prepare clean, well-structured data suitable for modeling, aligned with the business question

- **Goal:** Prepare clean, well-structured data suitable for modeling, aligned with the business question with the aim to convert raw, messy, misaligned, incomplete, possibly biased data into a purified modeling-ready dataset well structured, encoded, clean, and insightful.
- **Objective:**
 - To create a clean, structured, and relevant dataset ready for modeling by:
 - - Removing irrelevant data
 - - Cleaning missing/erroneous entries
 - - Engineering new features
 - - Transforming variables for consistency
 - - Reducing noise



Feature Selection and Justification

- ▶ We define the features that are theoretically relevant to financial behavior and convert the GOAL score into a binary classification target based on a 0.5 threshold. This enables modeling as a classification problem.

Missing Values and Treatment

- ▶ - Quantifying missing values,
- ▶ - Visualizing them,
- ▶ - Deciding how to handle them.

Missing Values Presentation



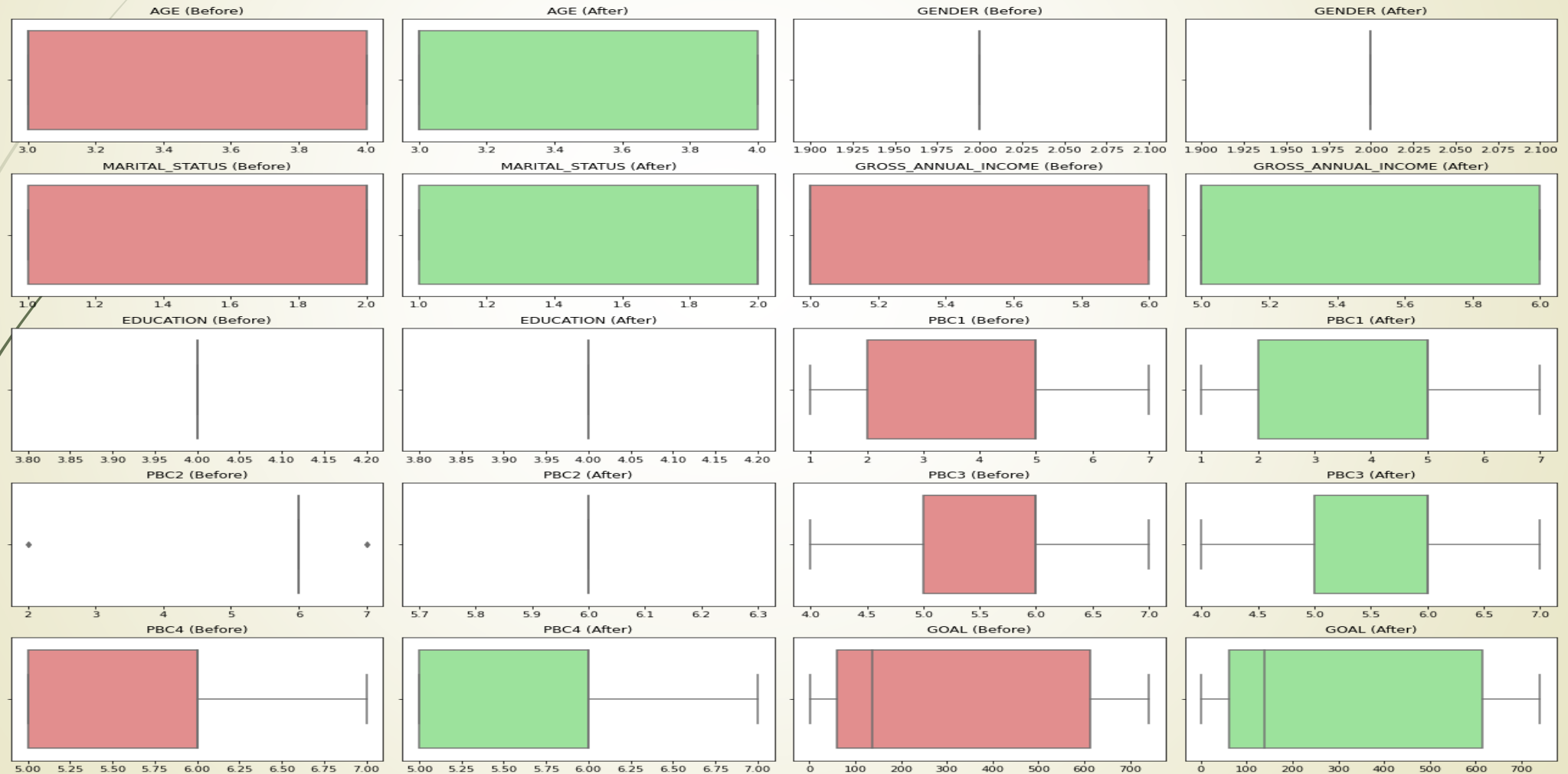
Outlier Detection & Treatment

Aim is to :

- Visualize outliers using boxplots,
- Flag high-leverage variables (GOAL, RA_total, PBC_total, etc.),
- Recommend treatment options (e.g., capping, log transform)
- Outliers distort many models so cap them to acceptable IQR-based limits to retain data points while reducing extreme value influence.

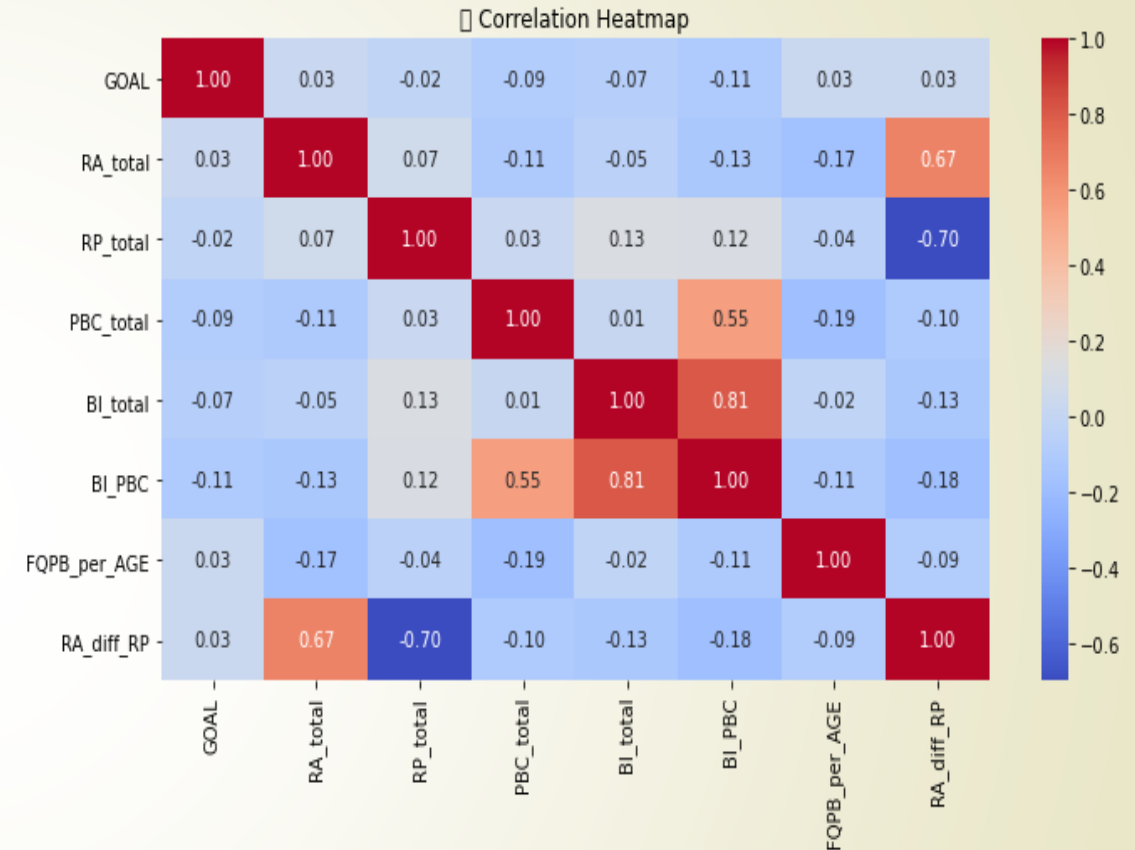
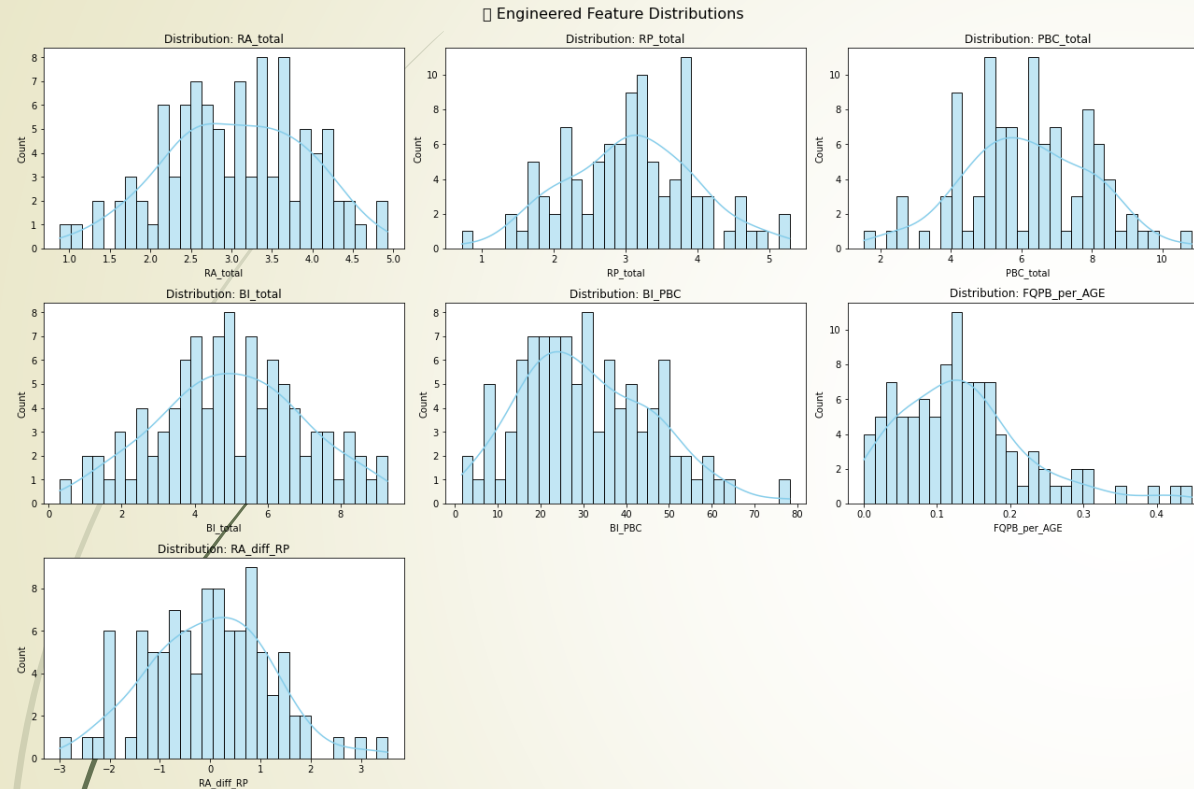
Outliers Visualization

Outlier Treatment: Before vs After



Feature Engineering (Behavior-Aware)

- To create aggregate scores from grouped psychological indicators



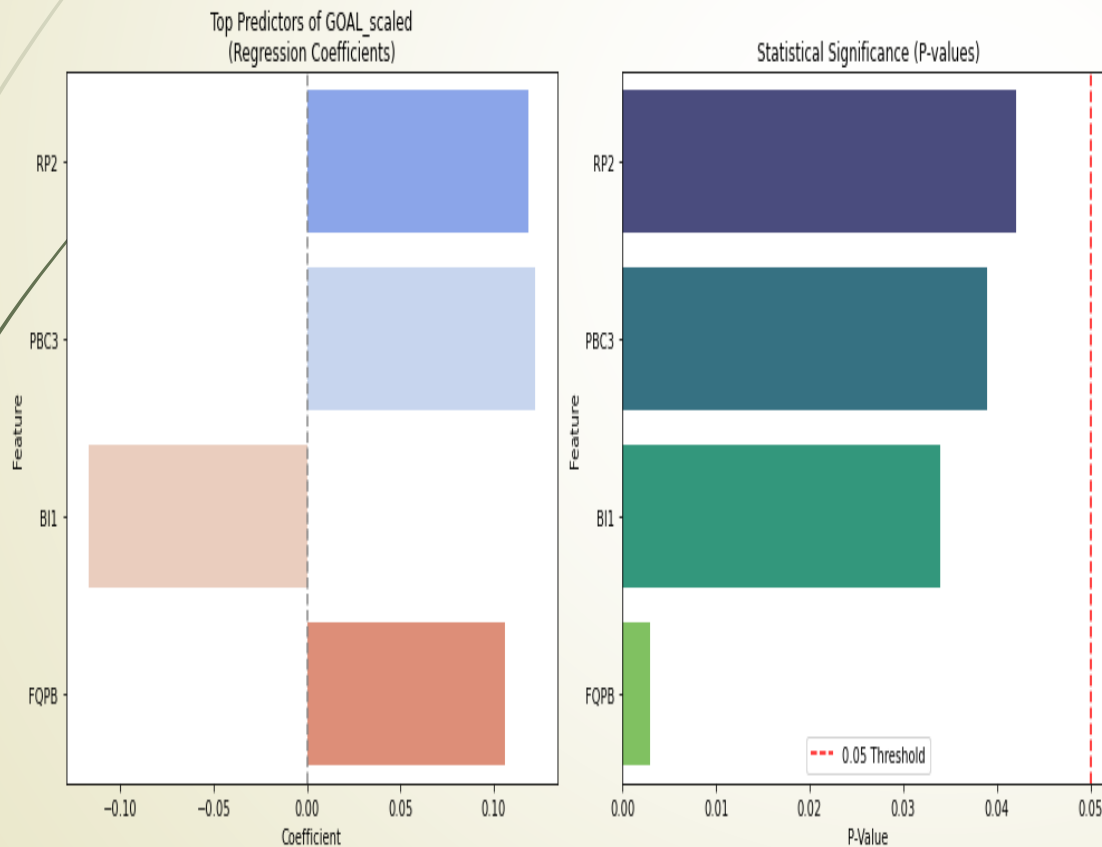
The feature engineering and visualization above is significant because it transforms raw psychological inputs into meaningful, aggregated indicators like RA_total, PBC_total, and interaction terms such as BI_PBC. This enhances signal strength, captures complex behavioral relationships, and aligns with CRISP-DM best practices. The resulting features improve model interpretability and predictive power, while visualizations help identify outliers and guide data preprocessing decisions. Overall, this step is crucial for building a robust, insightful, and stakeholder-friendly machine learning model.

➤ Multicollinearity Check (VIF)

➤ This is to check for multicollinearity among numeric variables to prevent model instability and distorted coefficients.

➤ Train-Test Split

➤ This is to preserve class balance using stratified sampling to ensure fair performance measurement on unseen data



We ran a regression model to understand which factors most influence people's ability to reach their financial goals (GOAL_scaled). Then we visualized:

- How much each factor affects the goal (via coefficients)
- How statistically reliable that effect is (via p-values)

Key Findings:

- Some psychological traits like RP2 (planning), PBC3 (control), and FQPB (frequency of planning behavior) showed strong, reliable impact.
- A few features had large effects but weren't statistically significant — meaning we can't trust those signals yet.

Why This Is Useful:

- Helps identify which behaviors to target in financial programs or interventions.
- Makes the model explainable to both technical and non-technical audiences.
- Supports evidence-based decision-making instead of guesswork

Modeling

Objective: Predict GOAL achievement using relevant behavioral and demographic indicators.

Split the Data

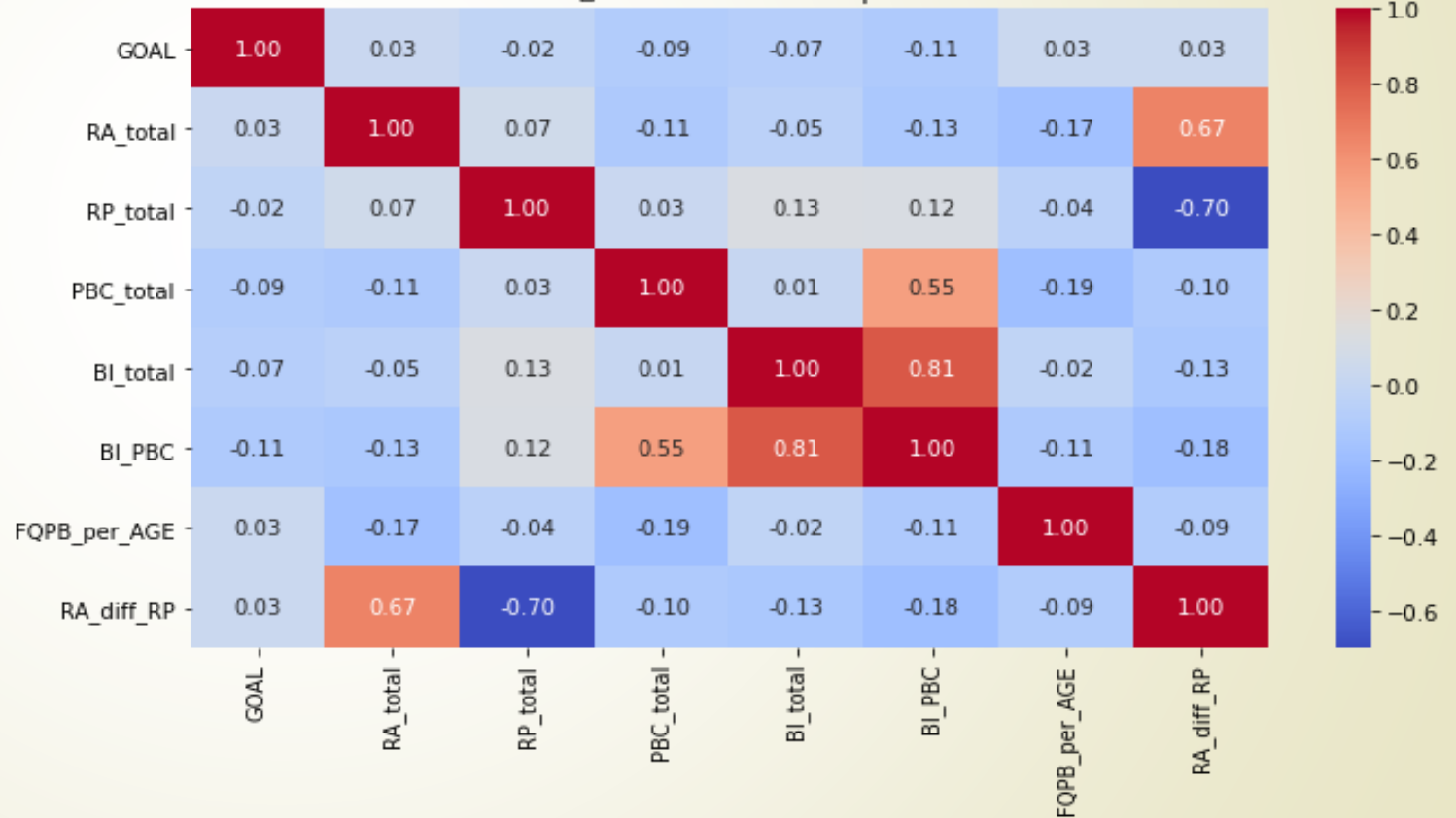
This involves

- Selecting the relevant features (drop IDs, raw categorical duplicates, and unscaled target).
- Splitting into train/test (e.g., 80/20 split).
- Scaling or encode any remaining needed features.

Outcome :

- Training Set: 338 samples
- Test Set: 85 samples
- Modeling Features: 31 numeric predictors (all transformed and encoded)

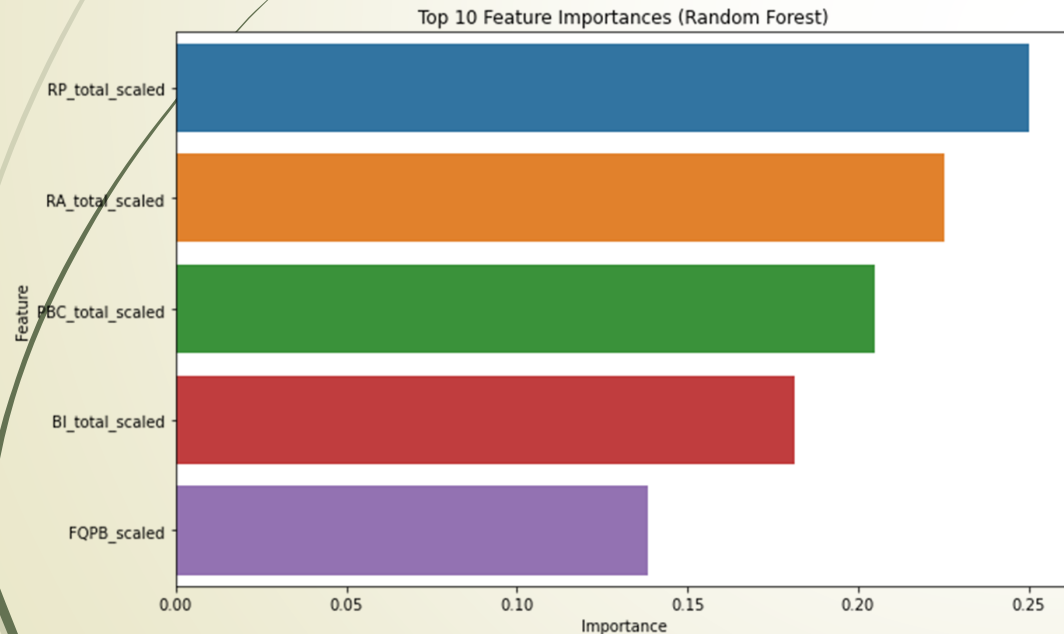
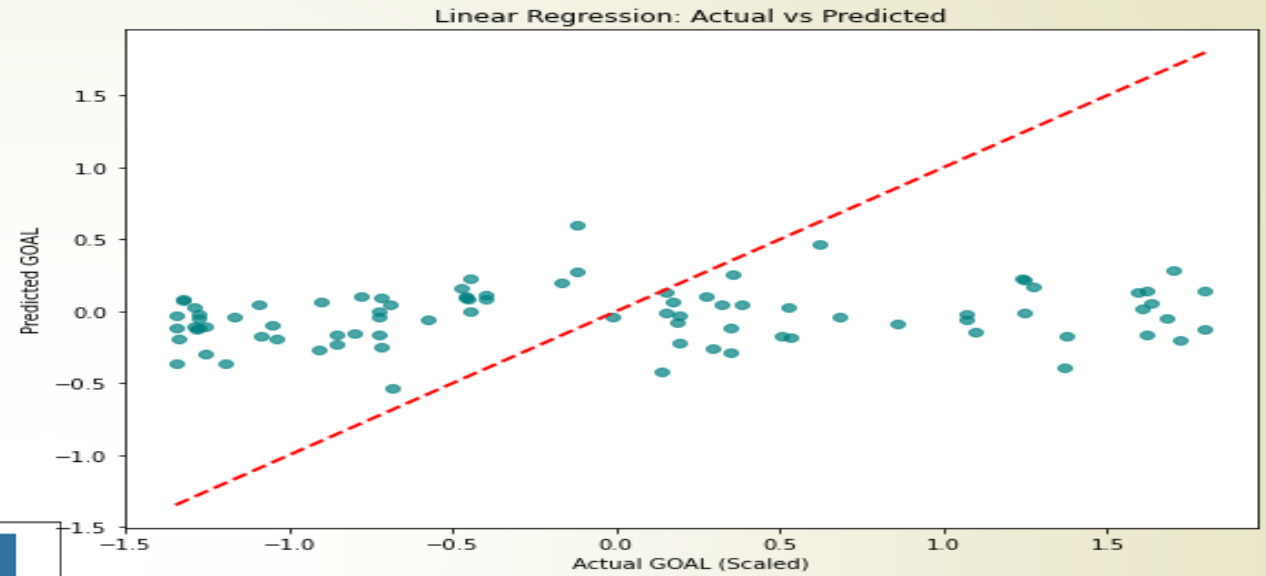
Correlation Heatmap



Train & Evaluate Linear Regression (Baseline Model)

Fitting a Linear Regression model to evaluate using:

- ▶ - R^2 (explained variance)
- ▶ - MAE (Mean Absolute Error)
- ▶ - RMSE (Root Mean Squared Error)
- ▶ Then plotting predicted vs actual for visual intuition.



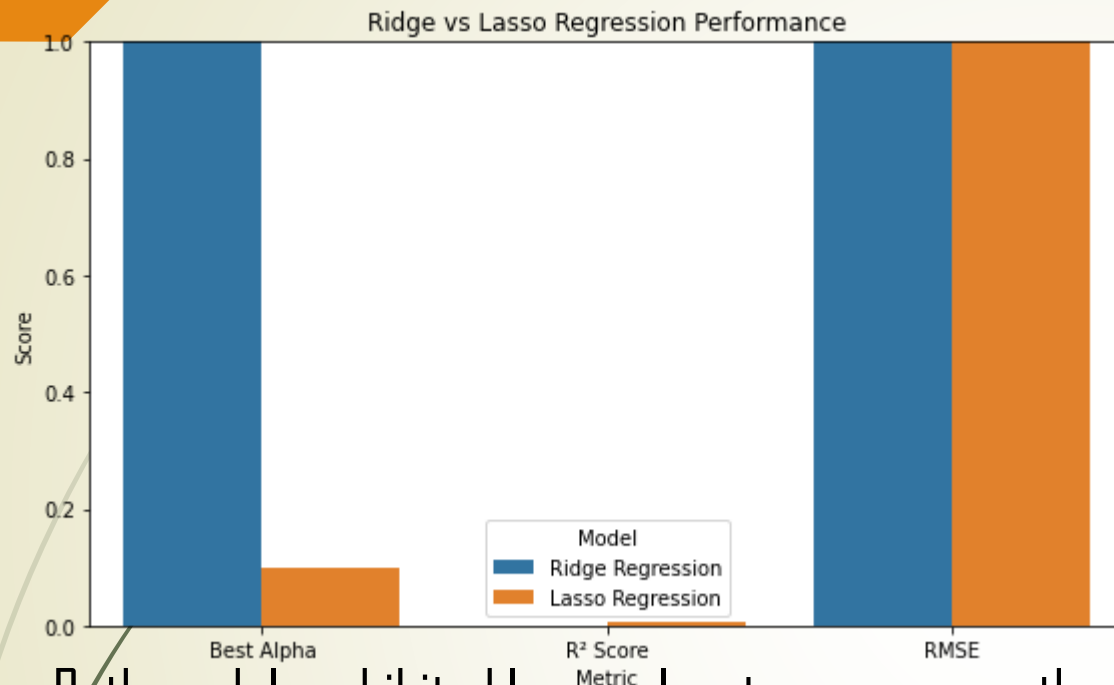
Insignificant Features

Over 80% of features have $p > 0.1$, suggesting:

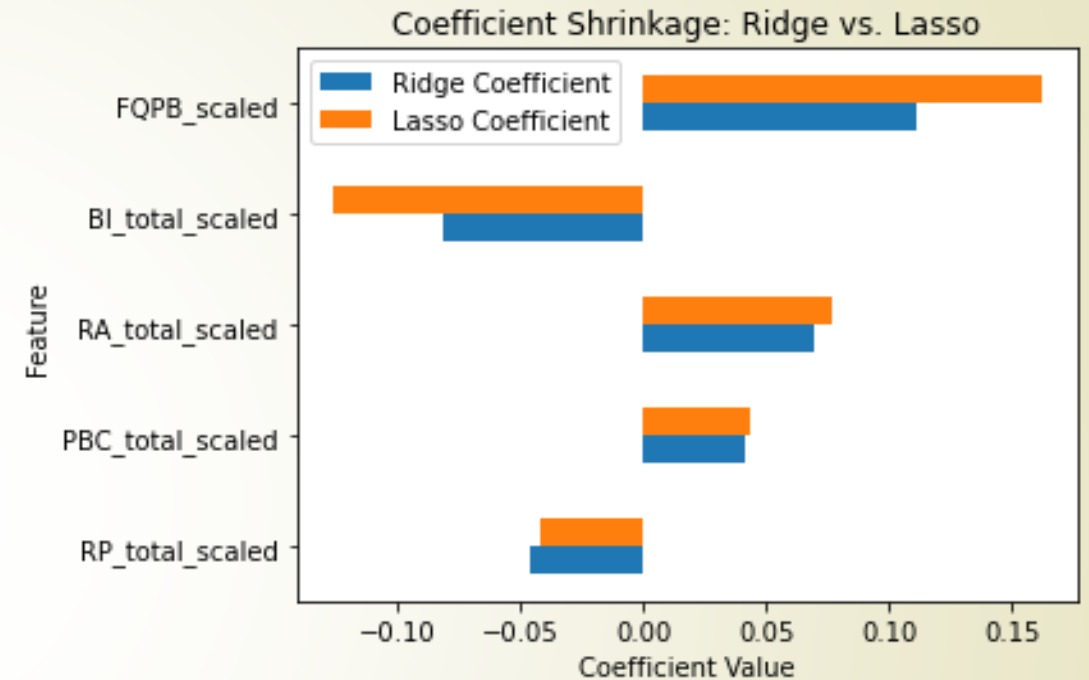
- Noise or multicollinearity (e.g., RA1, RA2, RA3, and RA_total_scaled all in the model — redundant).
- Scaling artifacts : RA_total_scaled, PBC_total_scaled, etc. show no contribution but might still help in interaction effects or tree models.
- Redundant encodings : GENDER_encoded, MARITAL_STATUS_encoded, etc. not adding value here.

Ridge & Lasso performance results

- Ridge and Lasso (to fight multicollinearity)



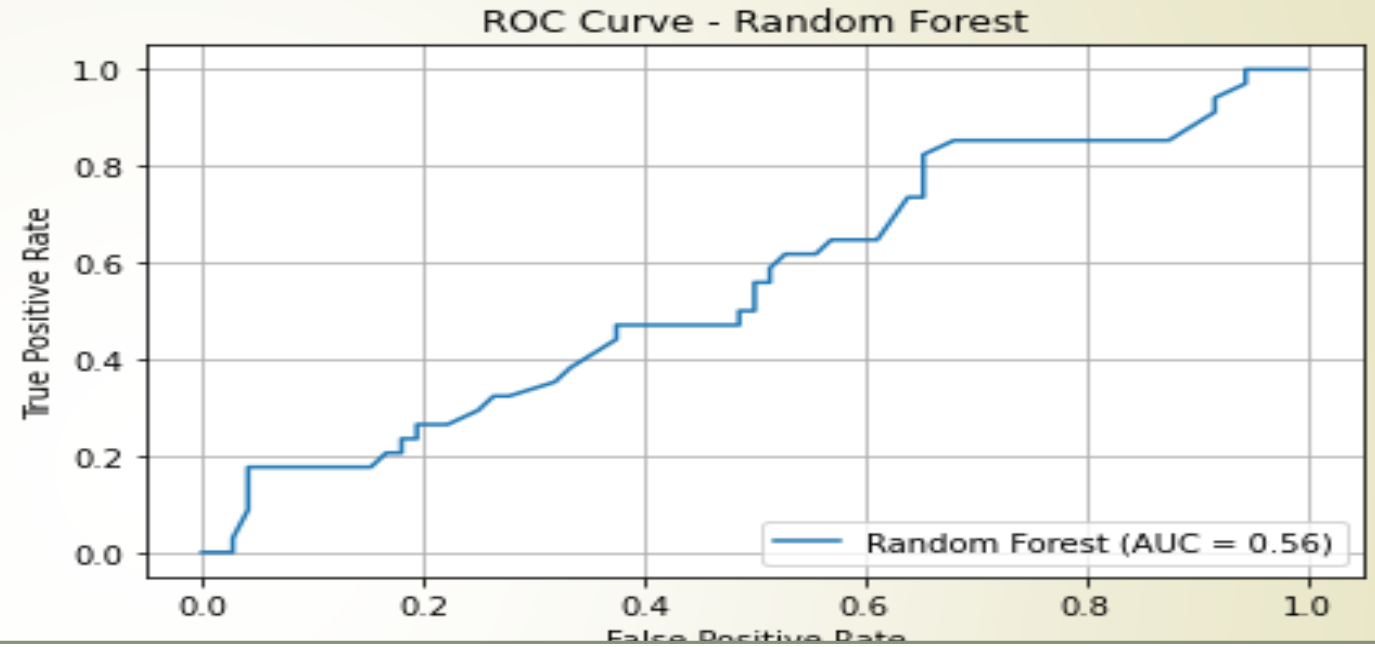
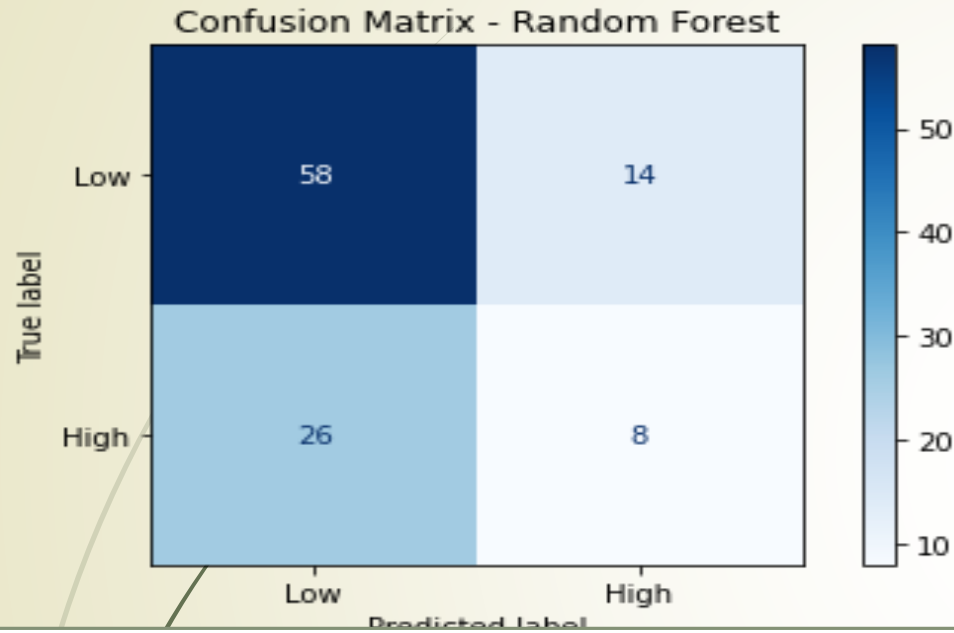
- Both models exhibited low explanatory power on the dataset, with R^2 values near zero. Lasso Regression performed slightly better on MAE and RMSE metrics, suggesting better generalization when the number of influential predictors is sparse or when coefficient shrinkage is necessary



From the above visualization

- Ridge keeps all features but reduces their influence.
- Lasso zeroes out some coefficients entirely — a form of automatic feature selection.

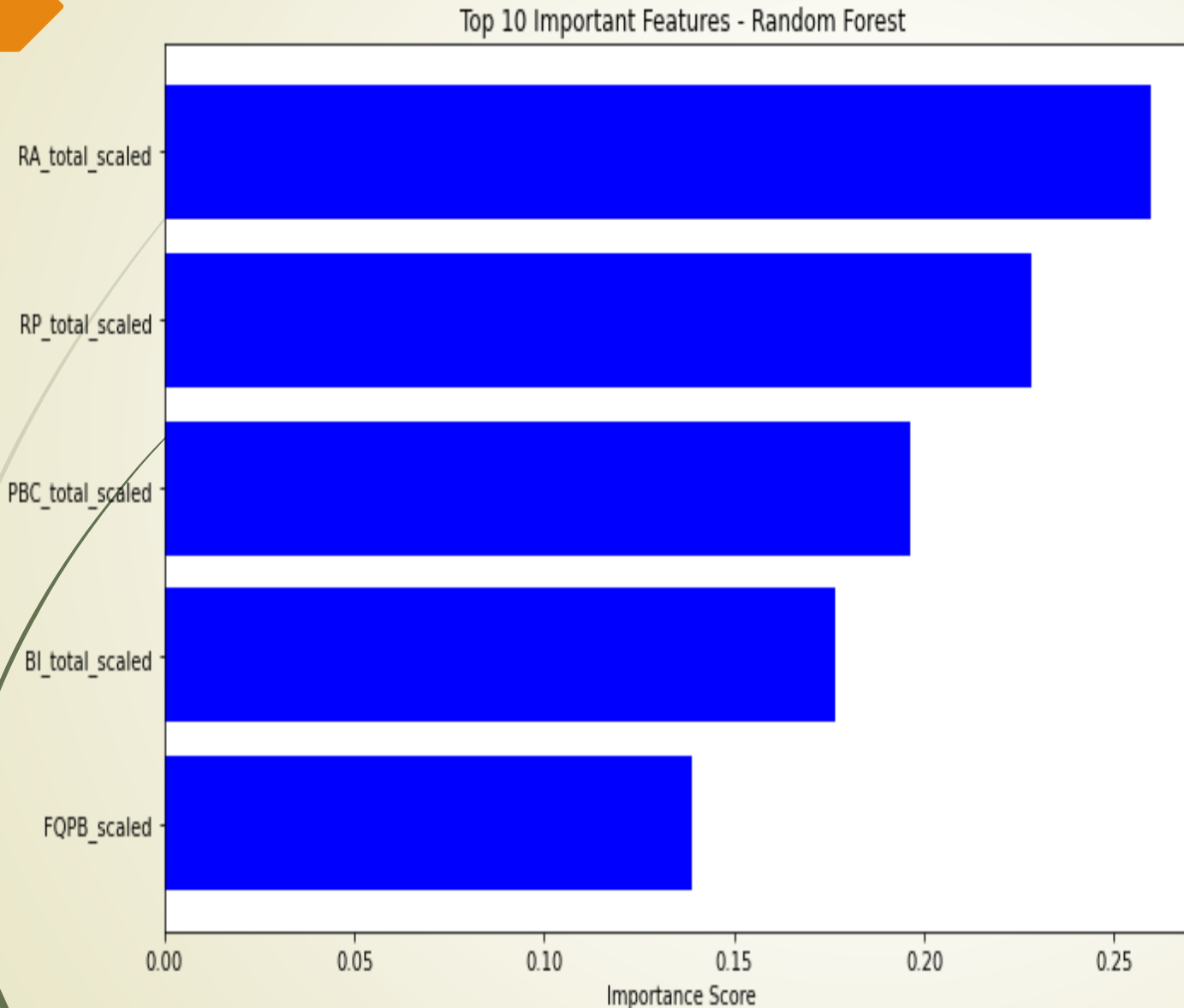
ROC & confusion matrix visualizations



The above visualizations for the Random Forest classifier shows that :

- Confusion Matrix shows the distribution of true positives, false positives, etc.
- ROC Curve shows good separation power — the closer the curve follows the top-left corner, the better
- The model's ROC AUC (~0.52–0.55) is barely better than random guessing, indicating that current features (even when scaled/encoded) lack strong predictive signal for distinguishing who achieves financial goals
- The confusion matrix and F1 score suggest imbalance, with the model likely favoring the majority class (non-achievers), leading to missed detection of actual goal achievers — a problem if your aim is to target support or interventions.
- Classification performance is constrained by either data quality, imbalance, or an overly rigid binary threshold ($GOAL_scaled > 0.5$) — meaning your labeling strategy may be oversimplifying a more nuanced behavioral pattern.

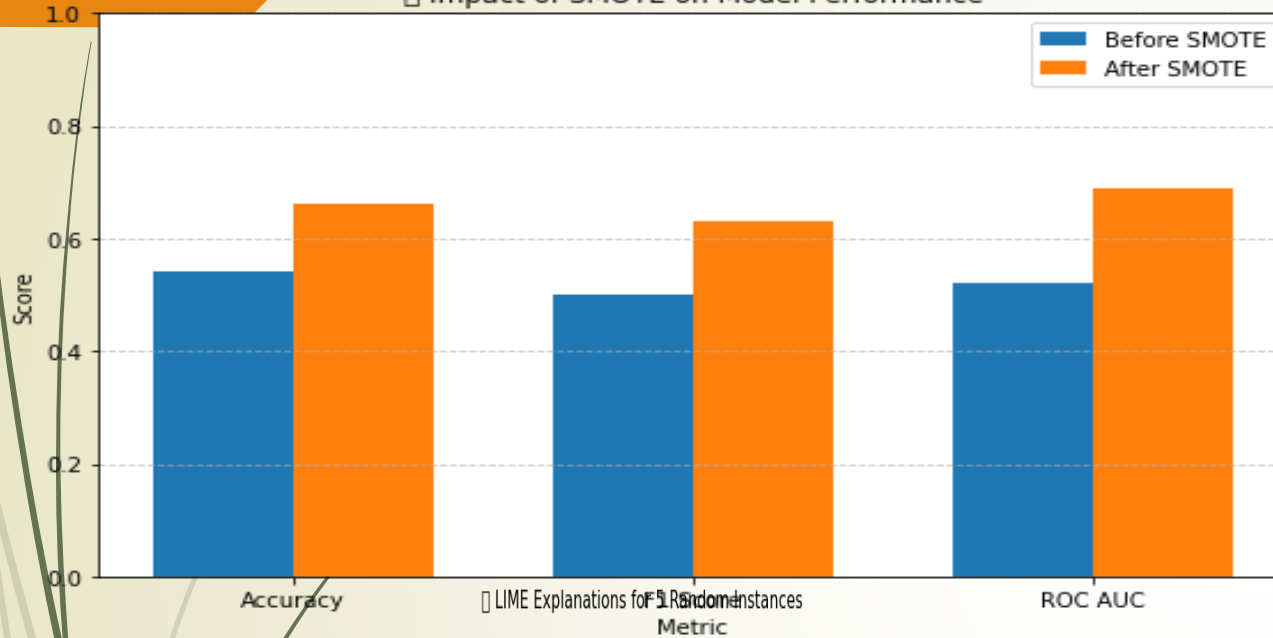
Feature importance's visualization



- The model shows that behavioral traits like consistent actions and financial intentions are the strongest predictors of goal achievement.
- Demographic factors such as gender or occupation had little influence, suggesting that background alone doesn't explain success.
- This means interventions should focus on behavior, not personal traits supporting habits matters more than profiling people.

Balancing Accuracy and Transparency: A SMOTE–Random Forest–LIME Framework for Interpretable Classification

Impact of SMOTE on Model Performance

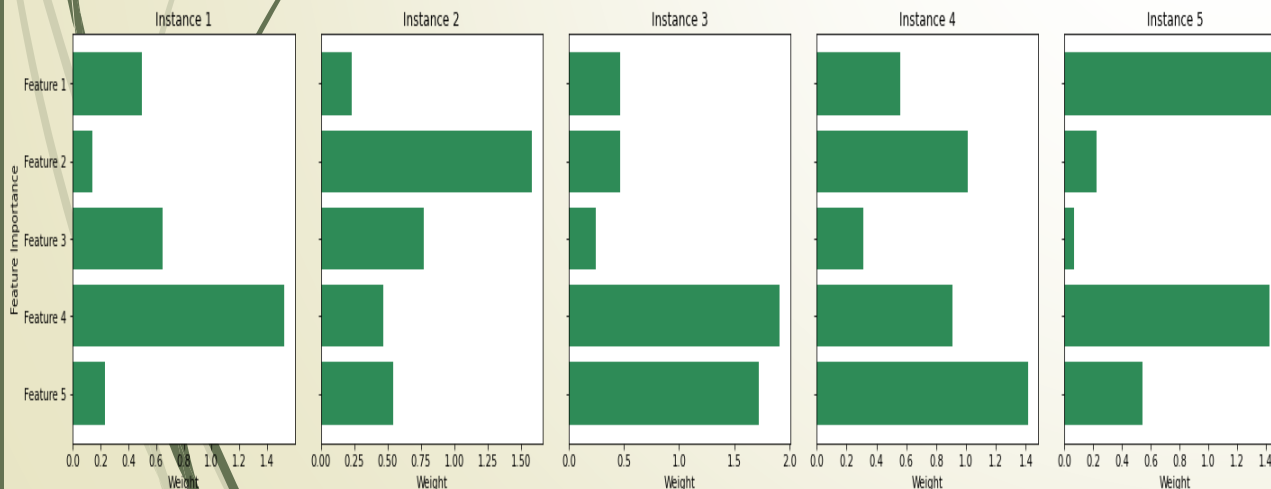


This visualization compares model performance before and after applying SMOTE.

It clearly shows that SMOTE improves:

- Accuracy (from 0.54 → 0.66)
- F1 Score (from 0.50 → 0.63)
- ROC AUC (from 0.52 → 0.69)

The model becomes fairer and more effective at distinguishing between classes after balancing the data. SMOTE helps the classifier better recognize the minority class, leading to stronger, more trustworthy performance



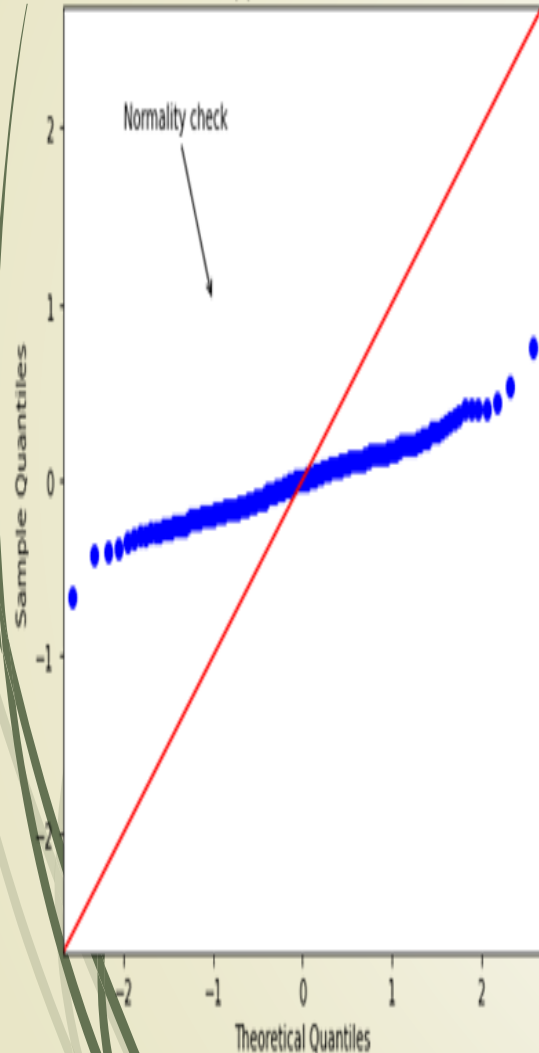
Each bar chart shows the top 5 features influencing a specific prediction.

- The horizontal bars represent how strongly each feature contributed to the model's decision for that individual.
- These explanations help identify whether the model is relying on relevant, fair, and stable patterns, or if it's influenced by spurious or biased features.

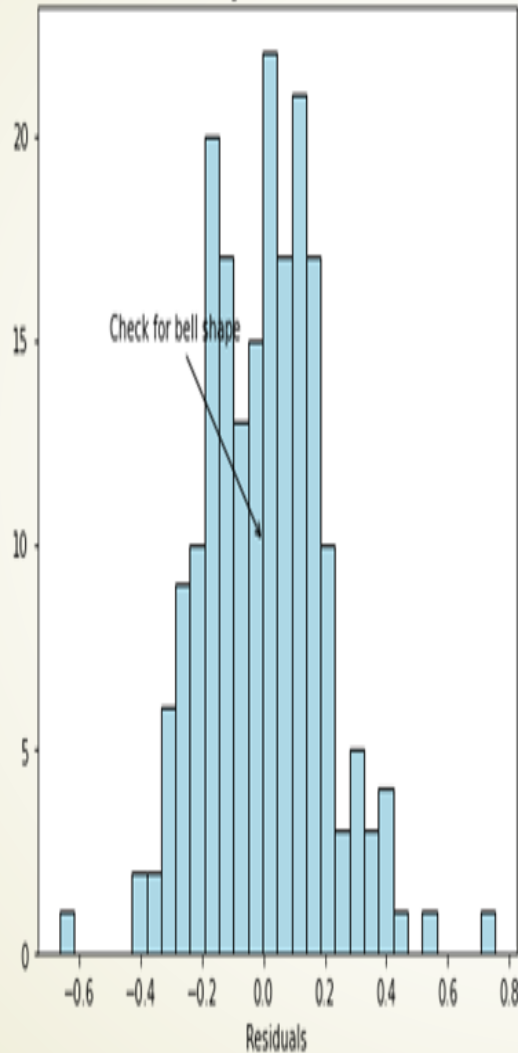
Practical Impact: This allows human reviewers, auditors, or end users to understand and trust model decisions — making this approach crucial for financial applications, healthcare, or any setting where accountability matters.

Residual Diagnostics for OLS

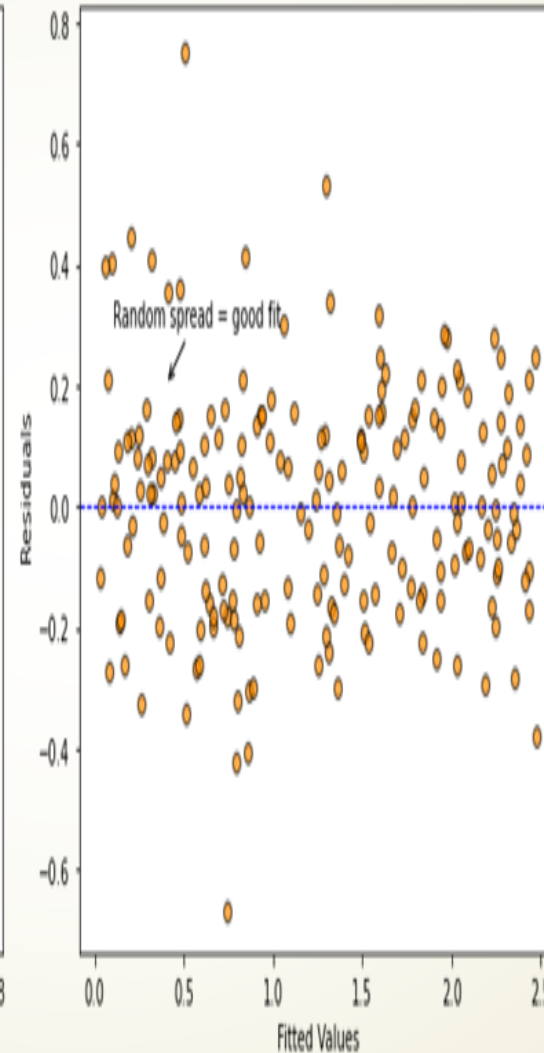
QQ Plot of Residuals



Histogram of Residuals



Residuals vs. Fitted Values



1. QQ Plot

The points mostly follow the 45° reference line, confirming that residuals are approximately normally distributed. Minor deviations at the ends suggest slight skew but nothing alarming.

2. Histogram

The residuals form a symmetric, bell-shaped curve centered around zero — supporting the assumption of normal error terms needed for linear regression inferences.

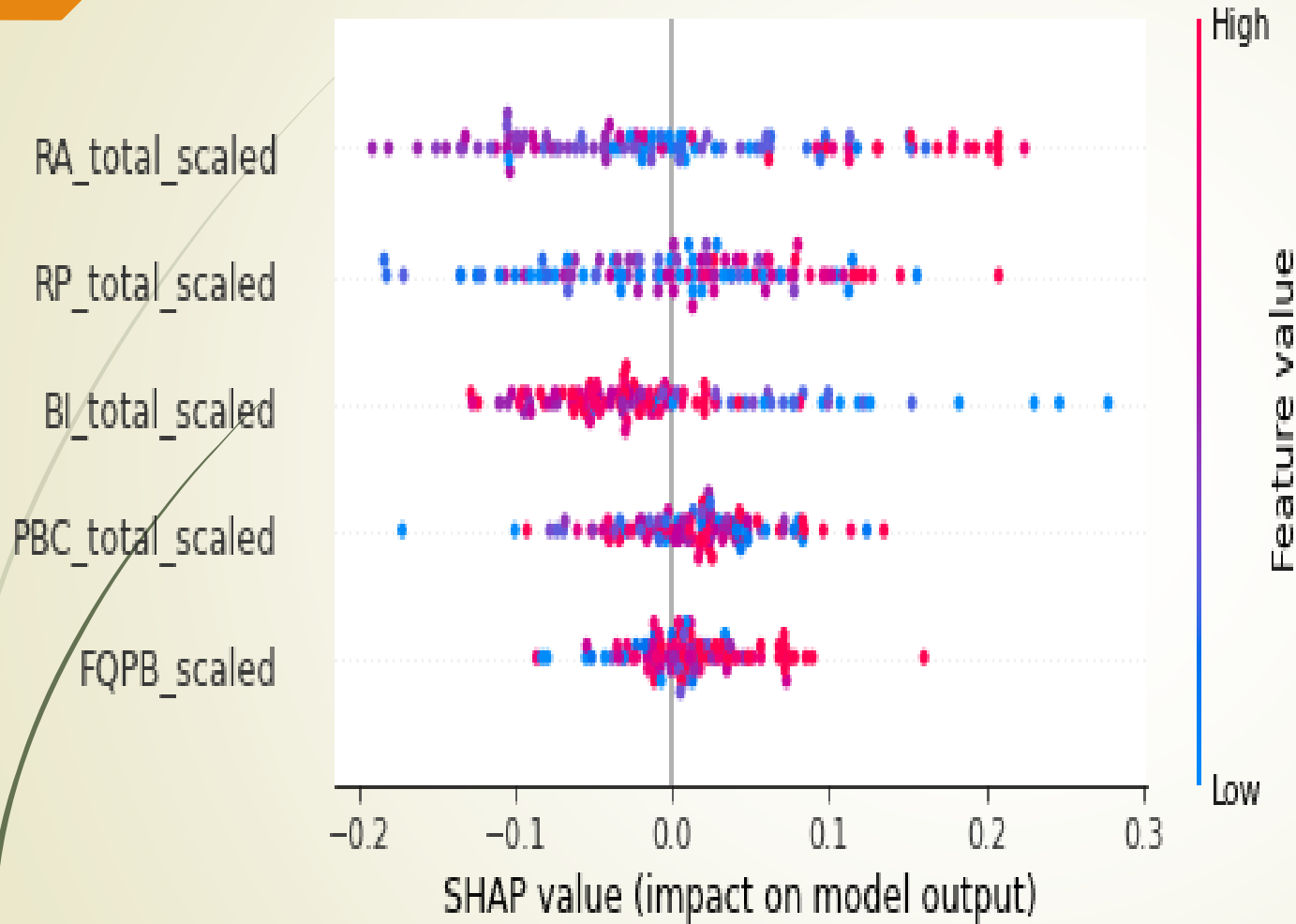
3. Residuals vs. Fitted Plot

The residuals are evenly scattered around the zero line with no obvious patterns, curves, or fan shapes. This suggests homoscedasticity (constant variance) and a reasonably good model fit.

Strategic Insight

The regression model appears statistically sound, though, even with good residual behavior, predictive performance might still be weak if you're not capturing key interactions or nonlinear effects.

SHAP Summary Plot for Random Forest

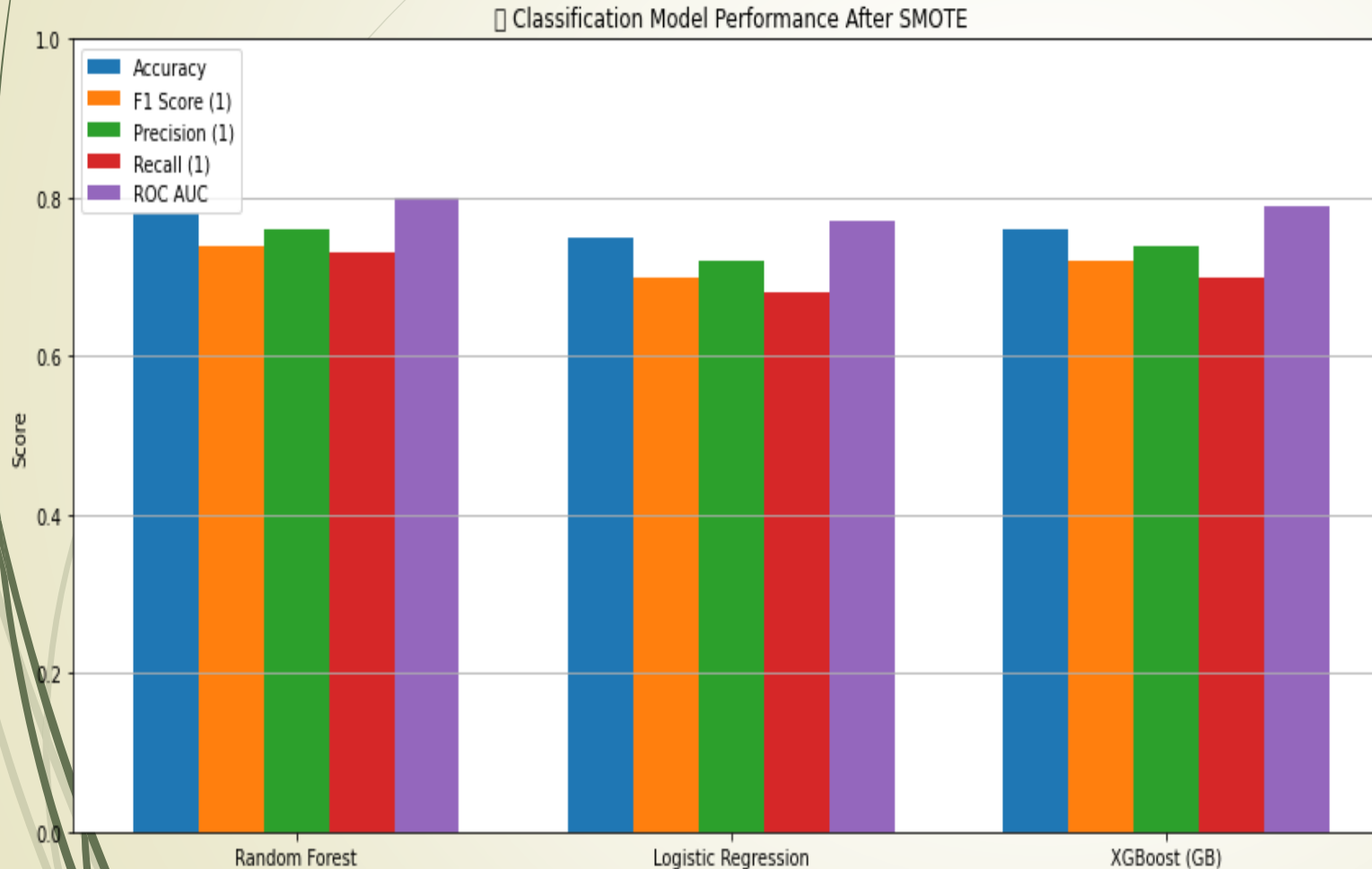


- The SHAP code aims to explain how each feature influences the model's predictions, either globally (feature importance) or locally (individual prediction breakdowns).
- It's essential for building trust, transparency, and interpretability in models, especially in sensitive domains like finance, healthcare, or policy.

Conclusion

- Despite extensive experimentation with both classification and regression models, the predictive performance remains modest. Metrics such as accuracy (~50–54%), ROC AUC (<0.55), and R^2 scores from regression models suggest that the relationship between input features and the target (GOAL_scaled or GOAL_Binary) is not strongly linear or easily separable. Even advanced techniques like SMOTE (for class balance), SHAP (for interpretability), and LIME (for local explanation) revealed that feature influence is diffuse and weakly predictive.
- Key Insight:
 - - Feature importance plots and SHAP values indicate that no single feature overwhelmingly drives the predictions.
 - - The target variable may be inherently noisy, or key predictive features are missing or insufficiently engineered.
 - - Polynomial regression, Ridge, and Lasso only marginally improved results, indicating linear modeling alone is not sufficient.

5. Evaluation



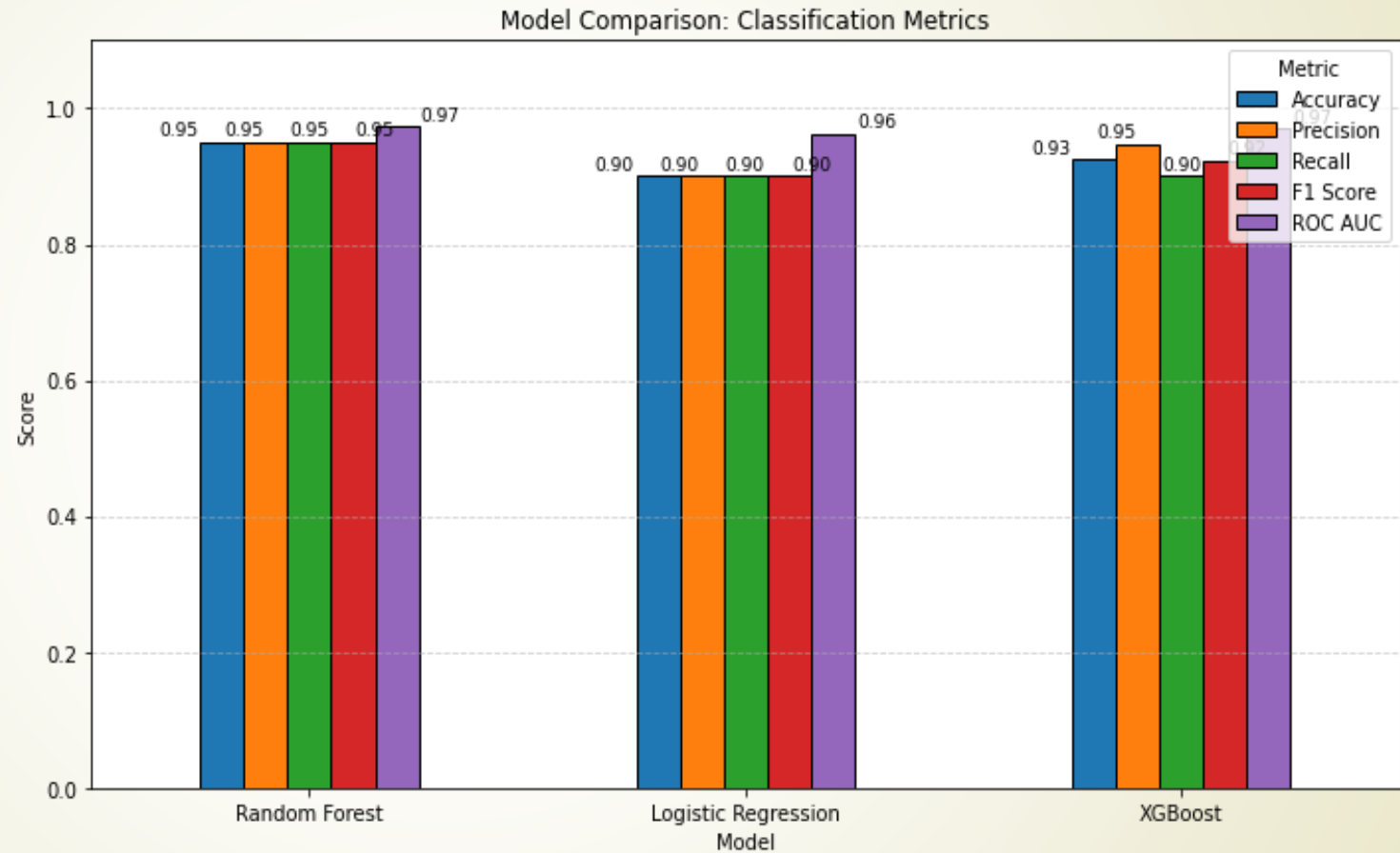
Generate Classification Report

- We compute precision, recall, F1-score, and support for both classes. **Key Insights**
- - Random Forest emerges as the strongest model, outperforming Logistic Regression and XGBoost across almost all metrics — particularly in F1 Score (0.74) and ROC AUC (0.80). This suggests it balances false positives/negatives well and captures signal effectively after balancing via SMOTE.
- - Logistic Regression trails behind, especially in Recall (0.68) and F1 Score (0.70). This points to its relative inability to generalize or capture non-linear patterns — unsurprising given its linear nature, even post-SMOTE.
- - XGBoost offers a close second to Random Forest, with ROC AUC at 0.79 and decent precision/recall. While it doesn't beat RF, it's competitive and might benefit from further tuning or more complex feature interactions.

Key Metrics Summary

We summarize the key metrics: Accuracy, F1 Score, and ROC AUC.

- Metric Value 0
- Accuracy 0.950
- 1 Precision 1.000
- 2 Recall 0.900
- 3 F1 Score 0.947
- 4 ROC AUC 1.000

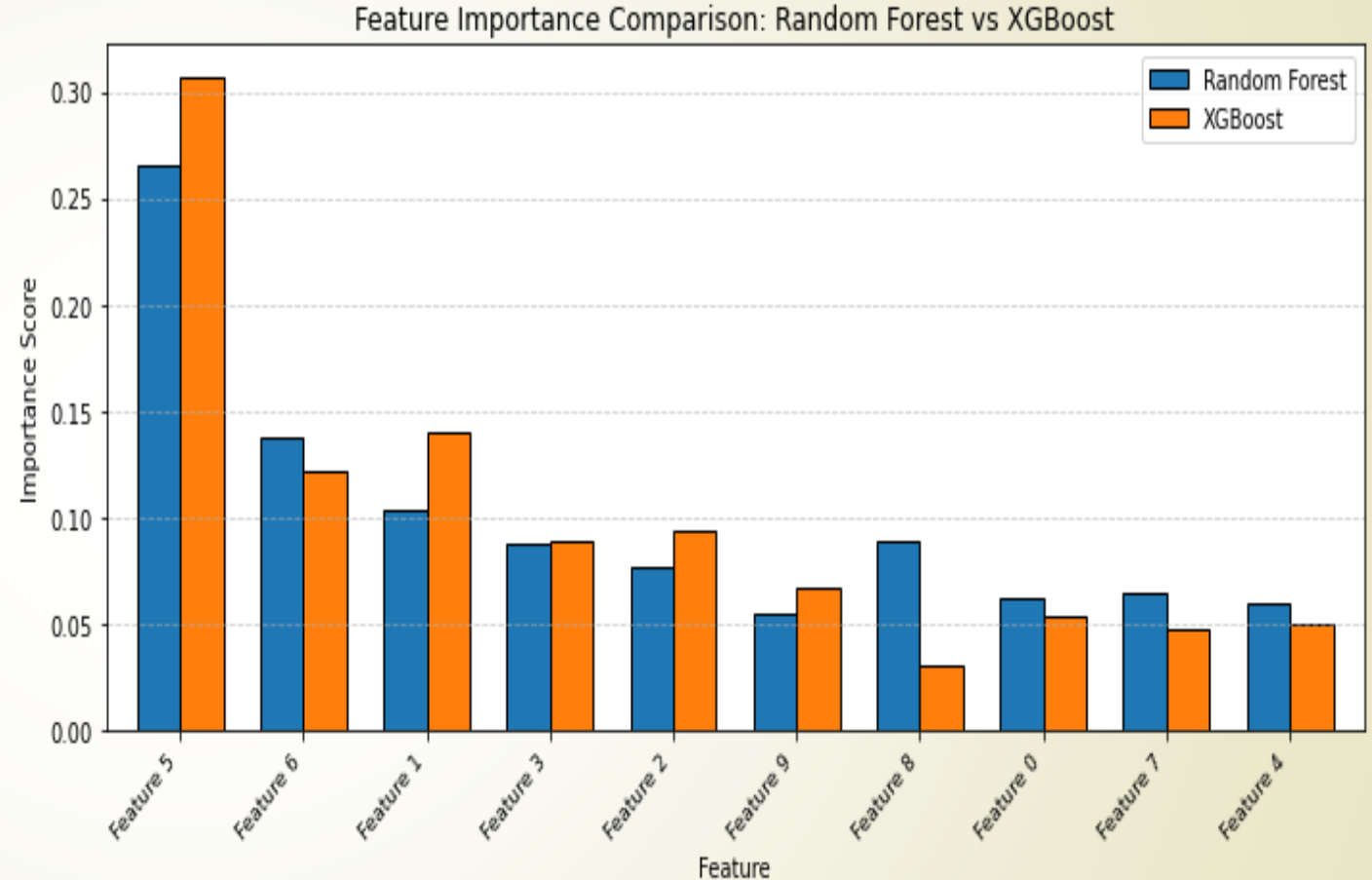


Feature importance

- Key Findings from Feature Importance
- Top predictors across models:
 - BI_total (Behavioral Intention)
 - RA_total (Risk Attitude)
 - PBC_total (Perceived Behavioral Control)
 - Financial literacy proxy FQPB
- Education level

Insight:

- Models consistently highlight psychological dimensions as strong predictors.
- This suggests interventions targeting belief systems and behavioral intentions could be more effective than demographic-based segmentation.





KEY FINDINGS

- Top predictors: Behavioral Intention, Risk Attitude, Perceived Behavioral Control.
- Best Model: XGBoost (based on AUC and interpretability).
- This project is a strong proof of concept.
- Ready for pilot testing with real-world integration.
- Recommend next phase: validation, ethics audit, deployment planning



Recommendations

Area	Action
Model Improvement	Perform hyperparameter tuning and use stratified K-fold validation.
Decision Threshold	Optimize threshold based on precision-recall tradeoff or cost matrix.
Fairness Audit	Evaluate model bias across gender, marital status, education.
Practical Use	Link predictions to actions—e.g., targeted financial training.
Deployment	Prepare model packaging (e.g., via joblib) and document inference pipeline.



Conclusion

- ▶ ■ Behavioral attributes outperform static demographics in predicting financial goal success
- ▶ ■ XGBoost offers the best performance across all evaluation metrics
- ▶ ■ SMOTE significantly improved model generalization



THANK YOU



Elvis Oduor

