TEAM 34: XIN'AI HAO; DOAN LE; YANZE LIU; YANG XIAO; JIAYI YU; NINA ZHUANG

DATA SCIENCE FOR BUSINESS

BANKRUPTCY PREDICTION

14 OCTOBER, 2025

BUSINESS UNDERSTANDING

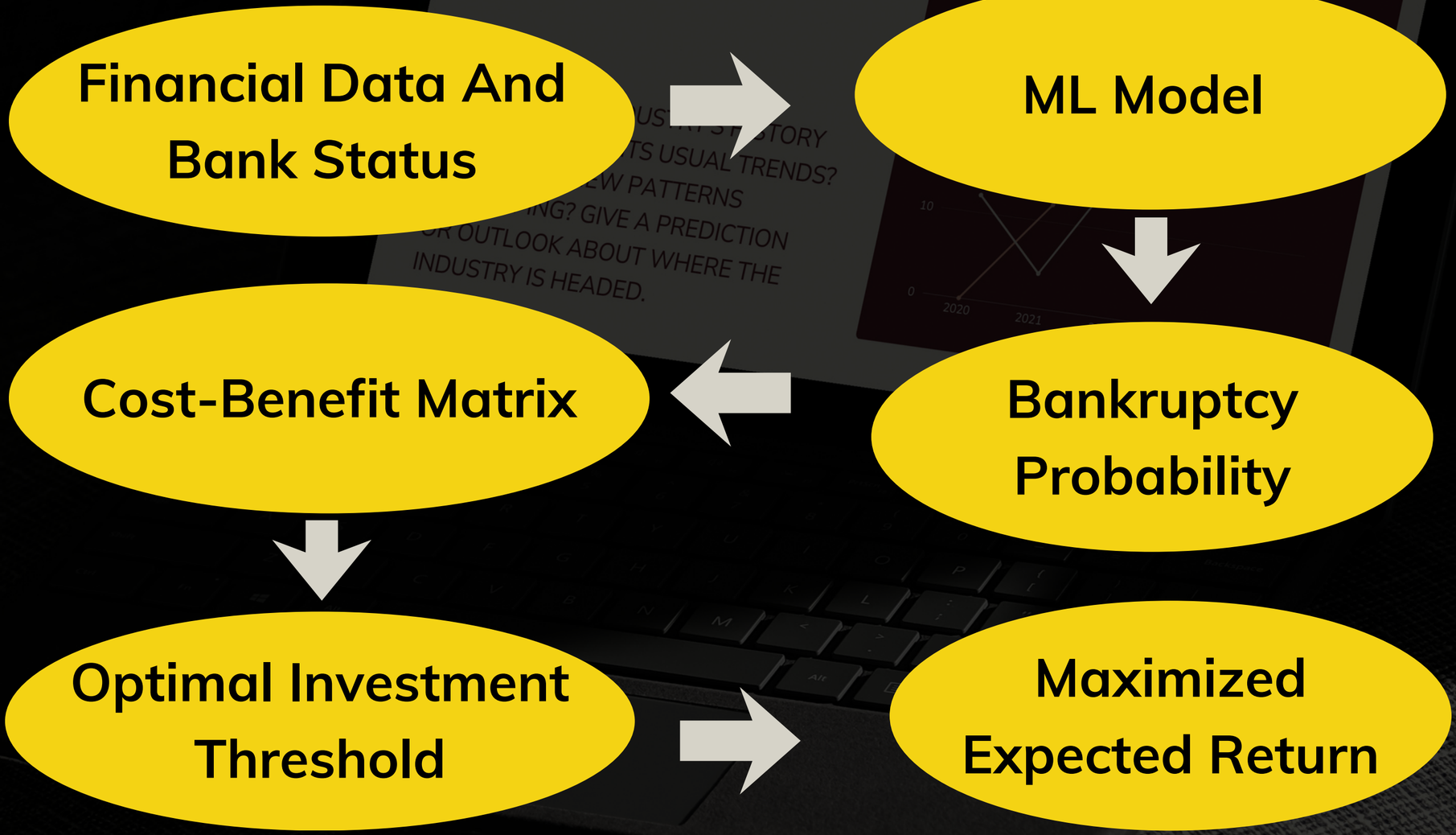
BANKRUPTCY → PORTFOLIO LOSS → RISK-ADJUSTED DECISION MAKING.

Institutional investors face significant losses from unexpected corporate bankruptcies.
Our project aims to predict which firms are likely to fail and translate predictions into portfolio optimization under different risk preferences.

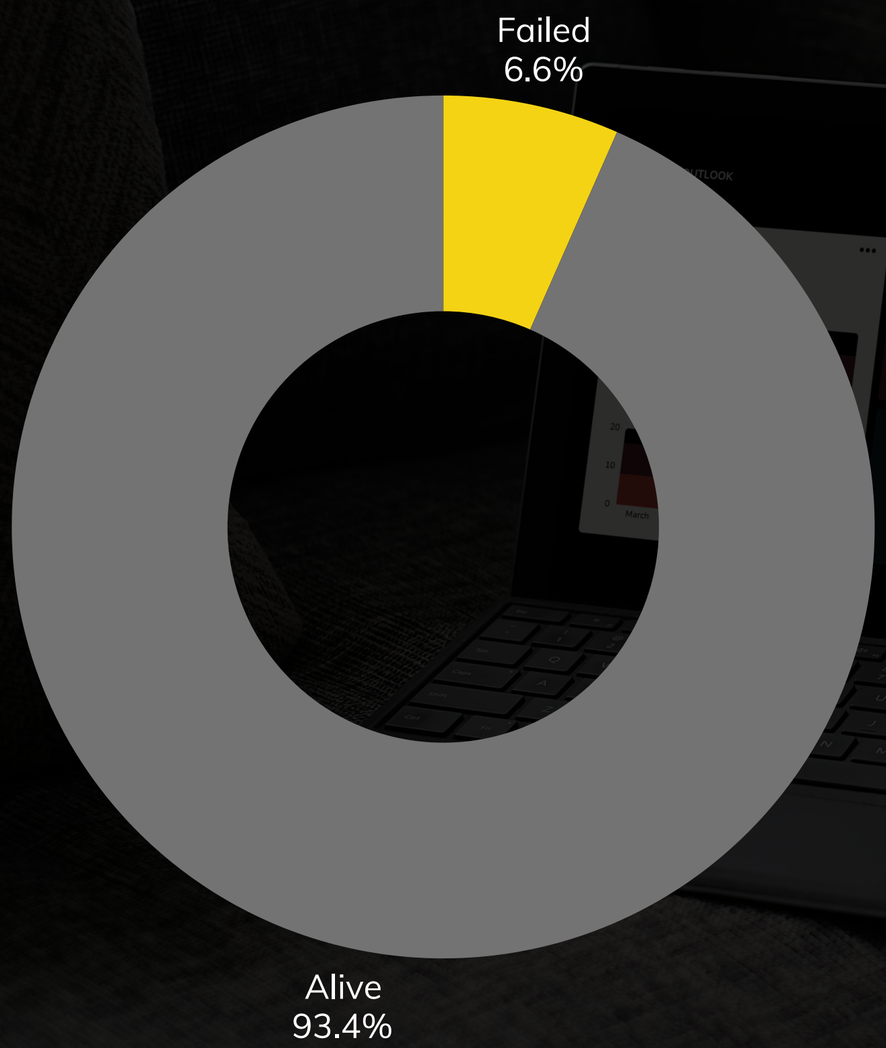
BUSINESS UNDERSTANDING

how will a data mining solution address the problem?

DATA MINING SOLUTION



DATA UNDERSTANDING



A novel dataset for U.S. public companies listed on the NYSE and NASDAQ is used for bankruptcy prediction. It contains accounting **financial data** from 8,262 firms between 1999–2018, totaling 78,682 firm-year observations with no missing or synthetic values.

According to the SEC, bankruptcy in the U.S. occurs under two legal codes:

- Chapter 11 – reorganization while continuing operations.
- Chapter 7 – complete liquidation and business termination.

For modeling, the fiscal year before bankruptcy filing (under either Chapter 11 or 7) is labeled as **Bankrupt (1)**, and all other firm-years are labeled as **Alive (0)**.

Source: https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset/data

DATA UNDERSTANDING

Conservative Investor

Investment Decision	Bankrupt (1)	Not Bankrupt (0)
Do Not Invest (ŷ = 1)	**0 (TP)**	**−0.05 (FP)**
Invest (ŷ = 0)	**−1.50 (FN)**	**+0.10 (TN)**

Neutral Investor

Investment Decision	Bankrupt (1)	Not Bankrupt (0)
Do Not Invest (ŷ = 1)	**0 (TP)**	**−0.10 (FP)**
Invest (ŷ = 0)	**−1.00 (FN)**	**+0.10 (TN)**

Aggressive Investor

Investment Decision	Bankrupt (1)	Not Bankrupt (0)
Do Not Invest (ŷ = 1)	**0 (TP)**	**−0.20 (FP)**
Invest (ŷ = 0)	**−0.80 (FN)**	**+0.10 (TN)**

Speculative Investor

Investment Decision	Bankrupt (1)	Not Bankrupt (0)
Do Not Invest (ŷ = 1)	**0 (TP)**	**−0.30 (FP)**
Invest (ŷ = 0)	**−0.50 (FN)**	**+0.10 (TN)**

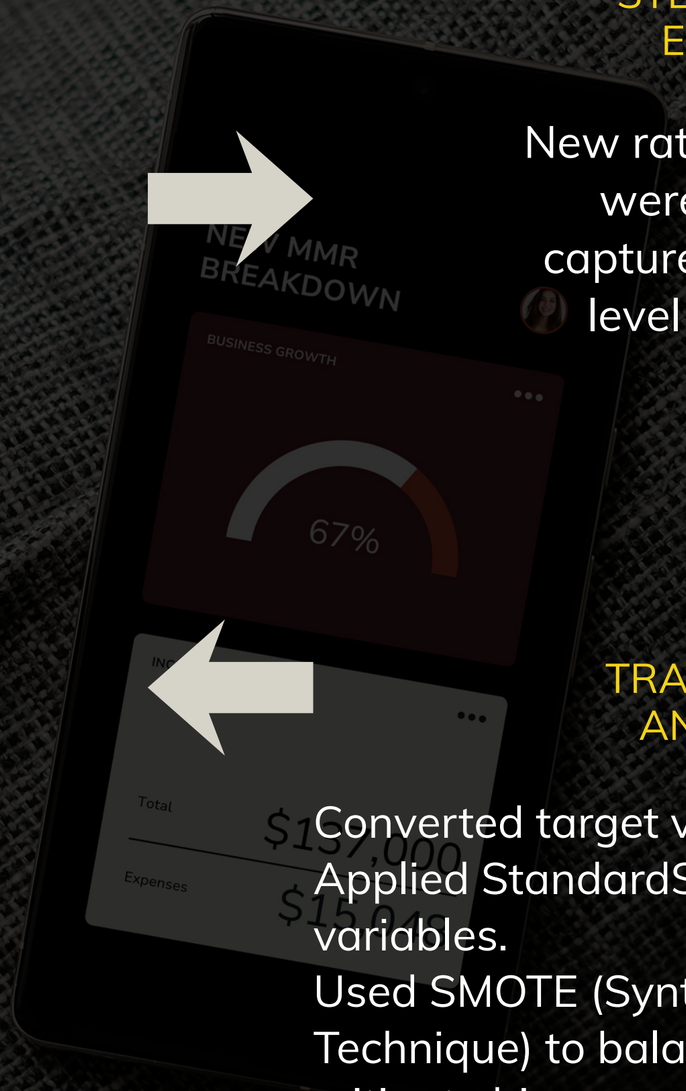
DATA PREPARATION

STEP 1: FEATURE CLEANING AND REDUCTION

Removed redundant or collinear features (X9, X12, X13, X18) after correlation analysis

STEP 2: FEATURE ENGINEERING

New ratio-based variables were engineered to capture normalized firm-level financial health



STEP 3: TRANSFORMATION AND ENCODING

Converted target variable status_label to binary.
Applied StandardScaler to normalize all numeric variables.
Used SMOTE (Synthetic Minority Oversampling Technique) to balance class distribution and mitigate bias.

STEP 4: DATASET PARTITIONING

specify how these data are integrated to produce the format required

MODELING

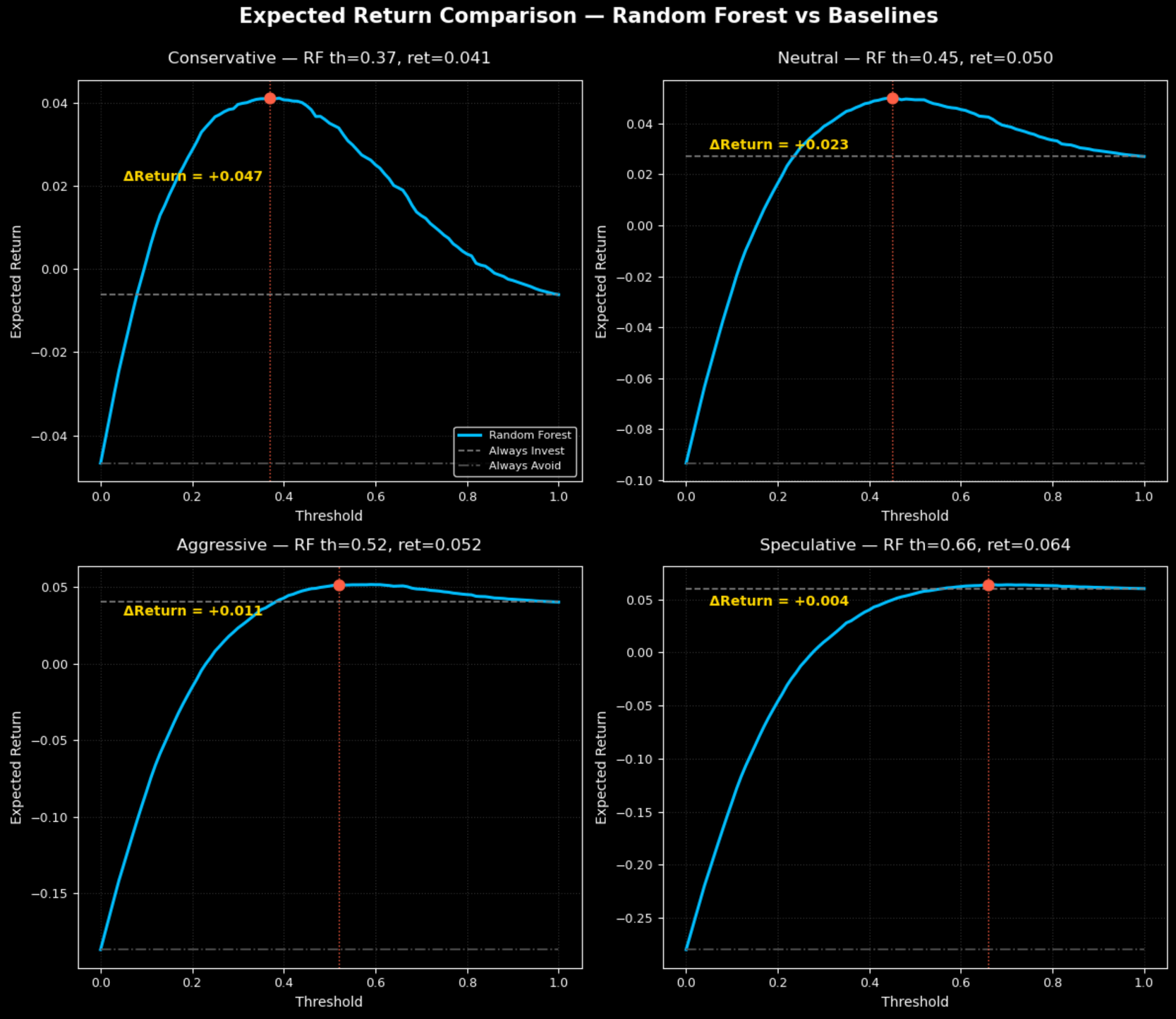
Model	Advantages	Disadvantages
Random Forest	Captures nonlinear relationships; highly robust and stable	Limited interpretability; can be slower on large datasets
XGBoost	High predictive performance; flexible tuning options	Complex parameter optimization; may risk overfitting
Logistic Regression	Simple and interpretable; good baseline for comparison	Limited expressive power; struggles with nonlinear patterns
SVM	Effective in high-dimensional spaces; performs well with clear margins	Computationally expensive; less interpretable and sensitive to parameter choice

We ultimately selected **Random Forest** as the core model because it provides the best balance between predictive accuracy, stability, and interpretability in bankruptcy risk estimation.

MODEL EVALUATION

Model	Conservative	Neutral	Aggressive	Speculative	Overall Performance	Remarks
Random Forest	th=0.37, ret=0.041	th=0.45, ret=0.050	th=0.52, ret=0.052	th=0.66, ret=0.064	🥇 **Best practical model**	Stable, accurate, interpretable.
XGBoost	th=0.35, ret=0.039	th=0.54, ret=0.047	th=0.79, ret=0.050	th=0.91, ret=0.064	⭐ **High returns**	High gains but unrealistic thresholds.
Logistic Regression (L2)	th=0.55, ret=0.007	th=0.57, ret=0.030	th=0.93, ret=0.040	th=1.00, ret=0.060	⚙️ **Moderate baseline**	Simple and interpretable, but weak on nonlinearity.
SVM	th=0.56, ret=0.005	th=0.58, ret=0.030	th=0.86, ret=0.041	th=1.00, ret=0.060	⚙️ **Moderate baseline**	Similar to Logit; poor probability calibration.

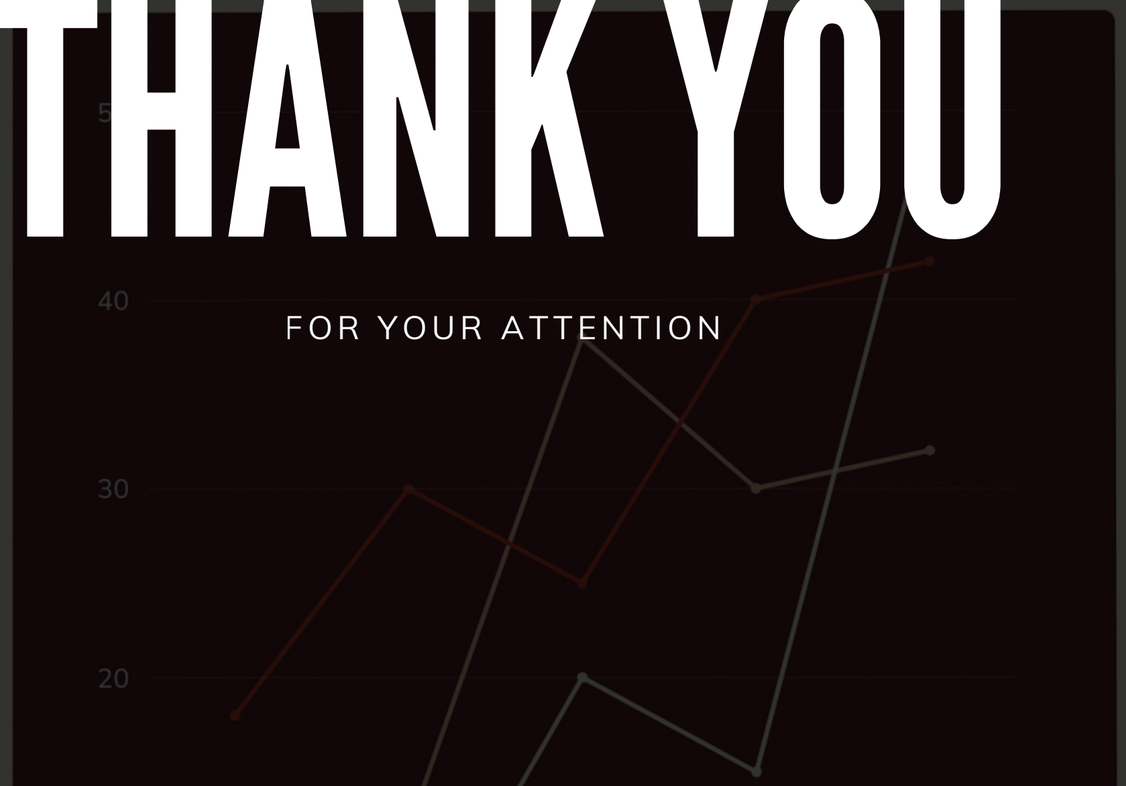
DEPLOYMENT


Expected Return Comparison — Random Forest vs Baselines
Conservative — RF th=0.37, ret=0.041
ΔReturn = +0.047
Random Forest
Always Invest
Always Avoid
Neutral — RF th=0.45, ret=0.050
ΔReturn = +0.023
Aggressive — RF th=0.52, ret=0.052
ΔReturn = +0.011
Speculative — RF th=0.66, ret=0.064
ΔReturn = +0.004
Expected Return
Threshold

WE WANT TO SAY

THANK YOU

FOR YOUR ATTENTION



GROUND

HE INDUSTRY'S HISTORY