SinoPac AI GO 2025 - High Return Stock Prediction Report

This report summarizes the latest progress in the binary classification project for predicting high-return stocks (“飆股”), developed for the SinoPac AI GO 2025 competition. For official details, visit the competition page. It provides a comparison between the initial stage (XGBoost on a reduced original dataset) and the most recent update (XGBoost for feature selection + KNN imputation + initial MLP).

Dataset Balancing Process

- Original dataset: 200,864 rows × 10,214 columns (~12GB).
- Target feature ('飆股') was highly imbalanced: Class 0 → 199,394 samples, Class 1 → 1,470 samples.
- To address imbalance, rows were reduced to 2,940 (balanced 1,470 vs 1,470).
- Selection criteria: keep majority class rows with fewer missing values.

Null Value Distribution

Even after balancing, the dataset contained a large number of missing values.

- Columns with null values: 9,862 out of 10,214.
- Columns with >30% nulls: 20 (discardable).
- Discarding them only removes ~0.20% of the dataset.

Key insight: heavy sparsity across features; imputation is crucial.

```
Columns with more than 30.0% of null values:
日外資_外資自營商買張                    2940
日外資_外資自營商賣張                    2940
日外資_外資自營商買賣超                  2940
日外資_與前日異動原因                    2916
日自營_自營商買均價                      1524
日自營_自營商賣均價                      1504
日投信_投信買均價                        2432
日投信_投信賣均價                        2590
日投信_投信持股成本                      1140
月營收_預估年營收(千)                    2940
月營收_累計營收達成率(%)                 2940
月營收_重要子公司本月營業收入淨額(千)      2940
月營收_重要子公司本年累計營收淨額(千)      2940
季IFRS財報_固定資產對長期負債比率(%)      1142
季IFRS財報_權益對長期負債比率(%)          1142
季IFRS財報_長期負債對淨值比率(%)           919
季IFRS財報_營業利益成長率(%)               898
```

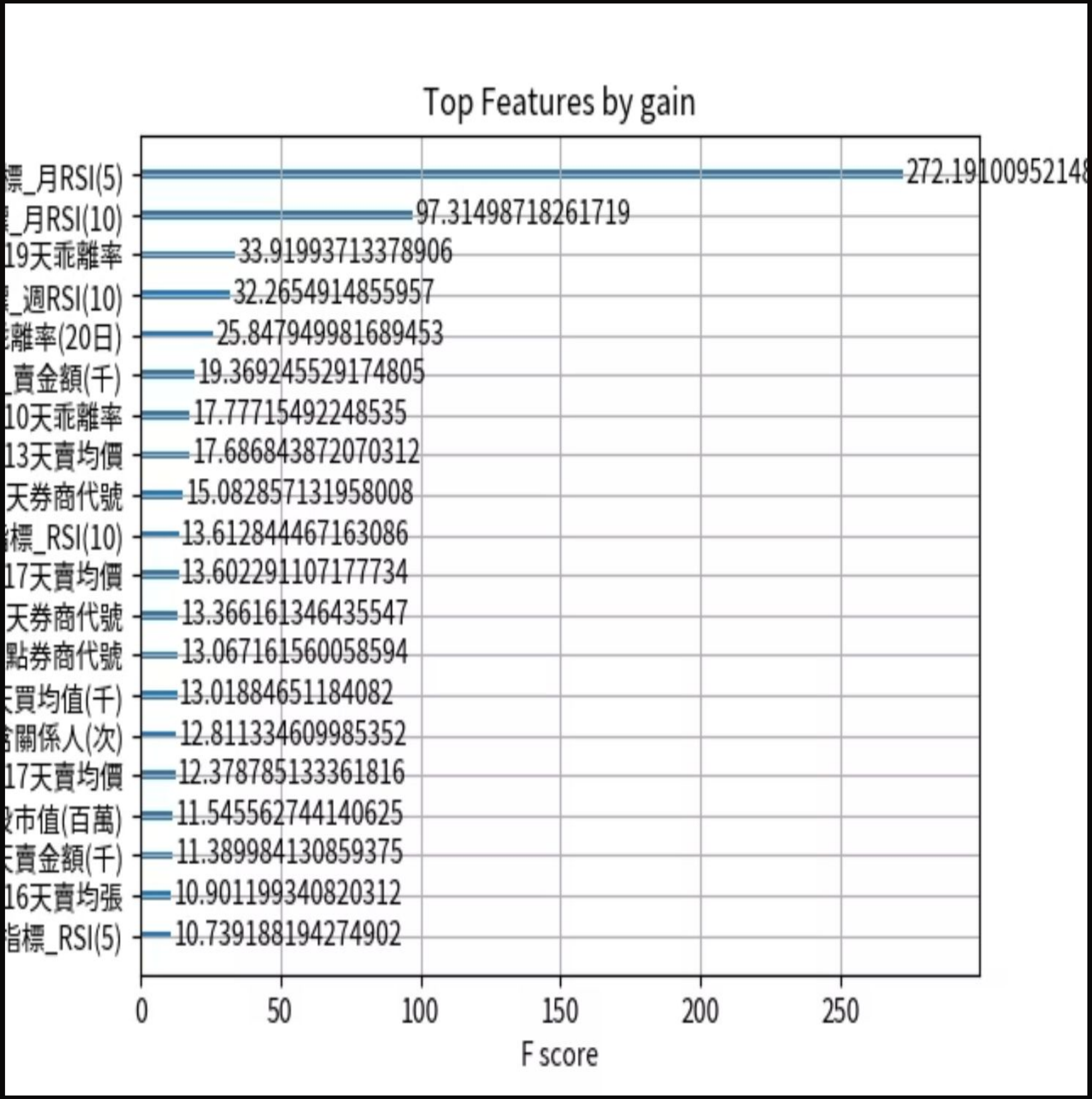
Advancing Our Predictive Capabilities

Dataset & FeatureSelection

Our initial dataset was a massive 12GB. To manage this, we focused on extracting the most salient features using **XGBoost**.

- The preliminary goal was to identify critical features even from data with significant missing values.

This process successfully reduced dimensionality to **1343** key columns/features, setting the stage for more focused analysis.



Refining Data Quality: KNN Imputation

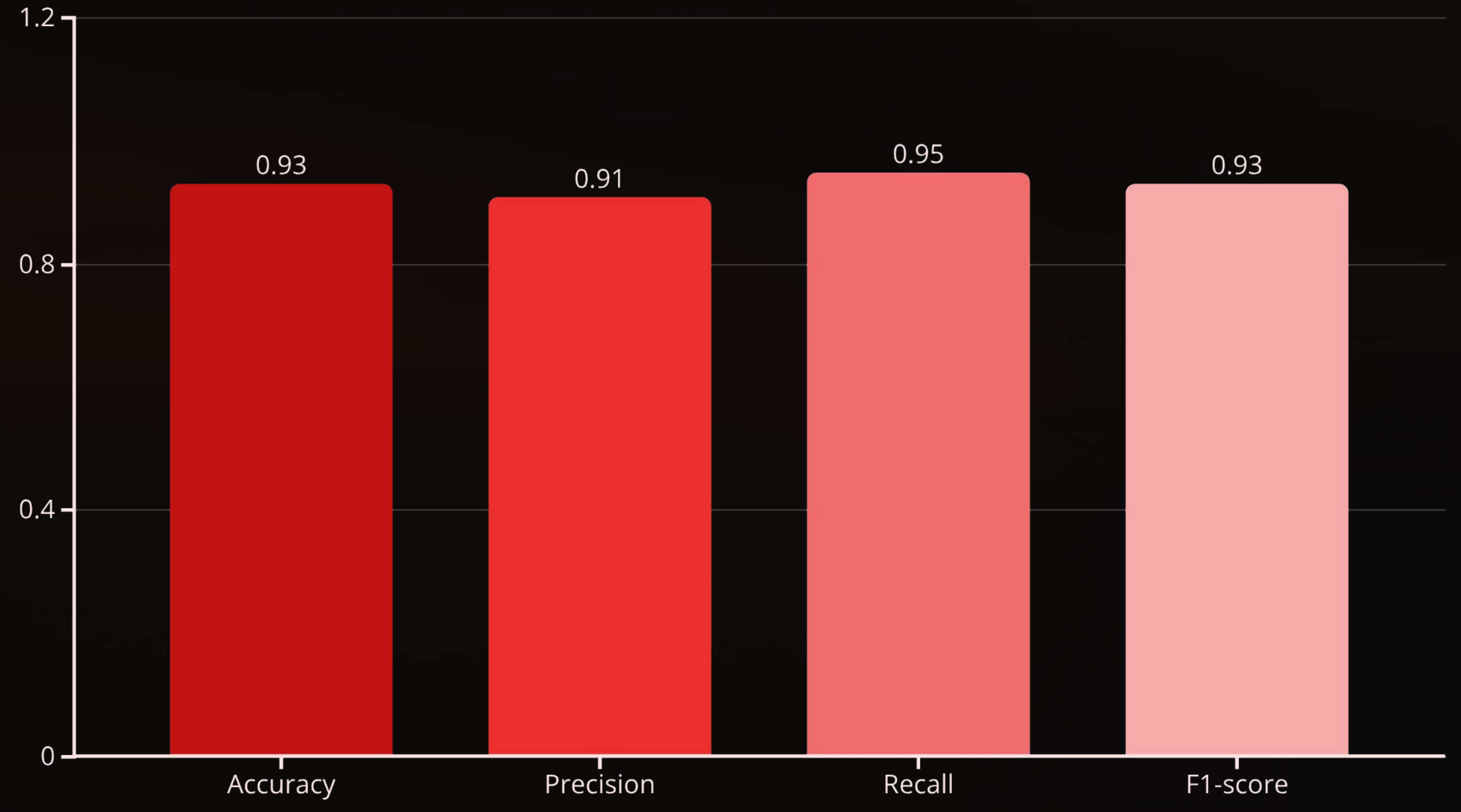
Data Imputation - KNNImputer

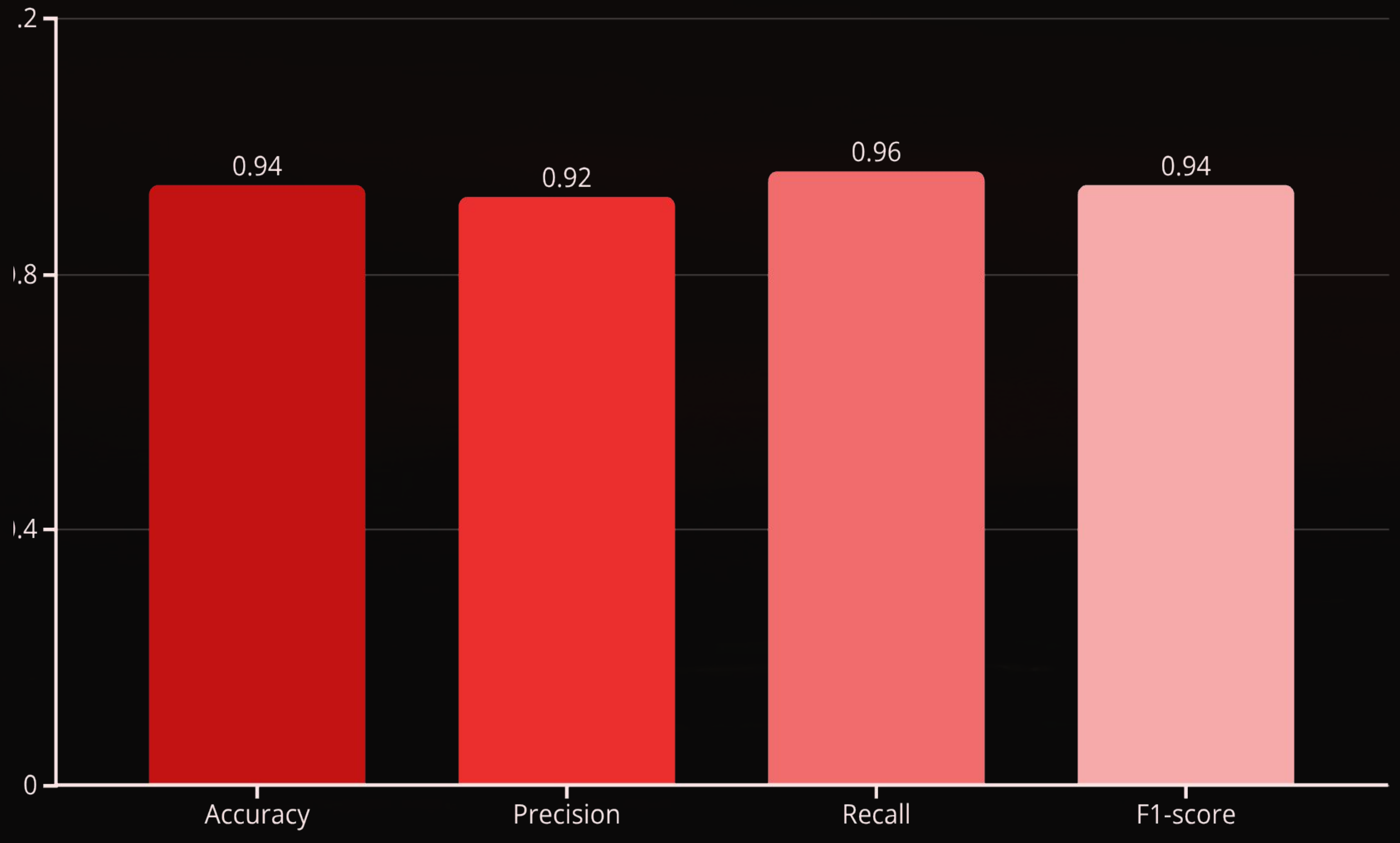
Even after feature selection, the dataset contained numerous columns with missing values. To address this, we applied **KNNImputer**.

- KNNImputer estimates missing values by considering the similarity of nearest neighbors.

Key Advantage: This method preserves non-linear relationships among variables, making it more robust than simpler imputation techniques (e.g., mean/median) and crucial for maintaining data integrity and predictive power.

Model Performance: XGBoost vs. MLP





XGBoost Results (Optimized)

Utilizing hyperparameter optimization via HalvingRandomSearchCV, XGBoost demonstrates robust performance, particularly in Recall, critical for identifying high-potential stocks.

Initial MLP Results (Optimized)

Trained on the KNN-imputed dataset with GA to optimize hyperparameters, the initial MLP shows promising results, indicating significant potential for future optimization.

Strategic Model Comparison

XGBoost: Current Leader

Achieved best overall performance, with a standout Recall of 0.95, indicating its effectiveness in capturing true positive high-return stocks. It appears near its optimization ceiling with current tuning.

MLP: High Potential with CV + GA

The initial MLP already delivers competitive results. Next, we will unlock more potential by combining robust cross-validati

Optimizing for Growth

While XGBoost delivers strong, stable predictions, the MLP offers a path to potentially superior performance. Future focus will be on unleashing the MLP's full capabilities through advanced optimization techniques.

This strategic comparison guides our future development, prioritizing models with the greatest potential for advanced stock prediction.

Next Steps: Unlocking MLP's Full Potential

1. Cross-Validation + GA

Run K-fold cross-validation combined with GA for hyperparameter search to improve generalization and reduce variance.

2. Integrate Cross-Validation

Incorporate robust cross-validation techniques to ensure model stability and generalizability across diverse datasets.

3. Performance Benchmark

Rigorously test the optimized MLP against XGBoost to determine if it can surpass current performance metrics.

Future Strategy

Current Focus

Our immediate priority is to maximize the performance of the MLP model through rigorous optimization, including Genetic Algorithms (GA) + cross validation.

Future Exploration

We plan to

Try advanced deep learning models for tabular data (e.g., **TabNet, FT-Transformer, ResMLP**).

Explore **ensemble approaches** combining tree-based models and neural networks.

Design and test a **custom Mixture of Experts (MoE)** tailored for financial stock prediction tasks.