Team 05 Jan 2025

Financial Forensics End - Term Project

Analyzing Key Trends and Creating Portfolio for Dalal Street

Problem Statement

The objective of this project is to analyze the financial data of all companies listed on the Bombay Stock Exchange (BSE) and create an optimal investment portfolio within a budget of INR 10,00,000.

Input and Constraints

The dataset includes metrics from the Balance Sheet, Profit and Loss (P&L) Statement, Cash Flow Statement, and other financial ratios, along with stock market prices at two specific time points, T1 and T2.

The task focuses on selecting a portfolio as of time-period T1 that maximizes returns and aligns with sound financial decision-making principles.

Provided Metrics

We are provided with the following financial documents for all companies listed on the Bombay Stock Exchange (BSE) for the time periods T1 and T2:

1. **Annual_P_L_1.csv :** Annual profit and loss data – Part 1.
2. **Annual_P_L_2.csv :** Annual profit and loss data -Part 2.
3. **Balance_Sheet.csv :** Balance sheet data providing a snapshot of financial position at T1/T2.
4. **Quarter_P_L_1.csv :** Quarterly profit and loss data – Part 1.
5. **Quarter_P_L_2.csv :** Quarterly profit and loss data – Part 2.
6. **Cash_flow_statements.csv :** Cash flow statement detailing cash inflows and outflows at T1/T2.
7. **Other_metrics.csv :** Additional metrics relevant for financial analysis at T1/T2.
8. **Price.csv :** Stock price data for companies at T1/T2.
9. **Ratios_1.csv :** Financial ratios - Part 1.
10. **Ratios_2.csv :** Financial ratios – Part 2

Approach

Our project follows a well-defined pipeline for stock price prediction and portfolio construction, enabling informed, data-driven decision-making. This structured approach ensures an optimized stock selection strategy that effectively balances risk and returns, making it a valuable tool for financial analysis and investment planning.

Data Loading & Exploration	Data Preprocessing	Model Training	Portfolio construction	Back Testing & Validation
Imported financial datasets (T1 & T2) containing key metrics like P&L, Balance Sheet, and Market Data. Performed exploratory analysis to understand feature distributions, missing values, and correlations.	Handled missing data, encoded categorical variables, and selected the most relevant features. Applied transformations to improve data quality for model training.	Trained XGBoost and Gradient Boosting models to predict stock prices, achieving high accuracy. Tuned hyperparameters and evaluated performance on test and future datasets.	Predicted stock returns and were ranked based on model-predicted scores. Selected top stocks with the highest returns .	Validated predictions on future data (T2) by comparing actual vs. predicted prices. Assessed model performance through visualizations and error metrics.

Data Loading & Exploration

- Loaded datasets T1 and T2, containing Profit & Loss statements, Balance Sheets, Ratios, and Market Data.
- Conducted Exploratory Data Analysis (EDA) using Pandas, Matplotlib, and Seaborn to understand data distributions, correlations, and missing values.
- Key financial indicators such as P/E Ratio, Debt-to-Equity, and Return on Equity (ROE) were examined to assess their impact on stock prices.

Data Preprocessing

- Missing values were handled:
  - Dropped columns where the current price was missing.
  - Numerical features were imputed using the median.

- Categorical features (like sector/industry type) were OneHotEncoded for machine learning compatibility.

- Feature Selection:
  - Used mutual_info_regression to identify the most impactful features.
  - Removed irrelevant features to reduce noise in predictions.

Model Training

- Machine Learning Models:
  - Implemented XGBoost and Gradient Boosting models for stock price prediction.
  - XGBoost achieved $R^2 = 0.77$ on test data and $R^2 = 0.71$ on T2 data.
  - Gradient Boosting performed better with $R^2 = 0.99$, indicating higher accuracy.

- Hyperparameter Tuning:
  - Optimized parameters like learning rate, number of estimators, and max depth for better prediction accuracy.

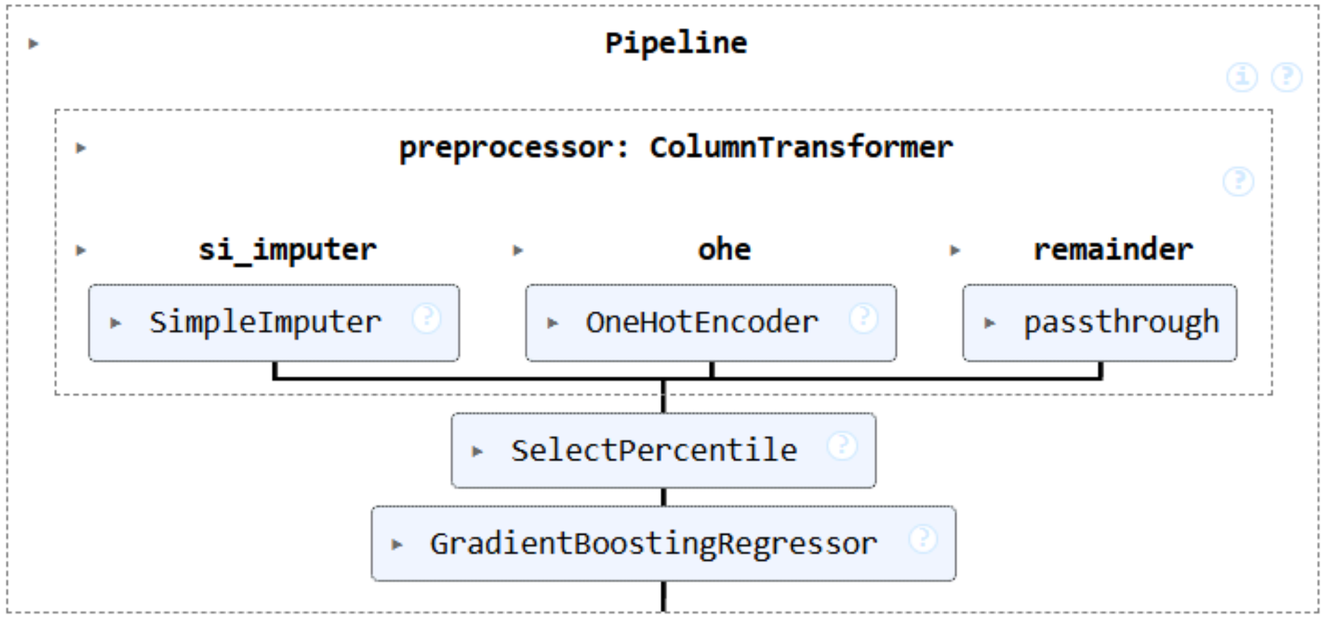
- Training Strategy:
  - Split data into training (80%) and testing (20%) sets.
  - Evaluated models using RMSE and R^2 scores.

Pipeline Used

- **Preprocessor: ColumnTransformer**
  - si_imputer (SimpleImputer)
    - strategy = 'median'
  - ohe (OneHotEncoder)
    - handle_unknown = 'ignore'
  - remainder
    - passthrough (keeps the remaining columns unchanged)

- **Feature Selection: SelectPercentile**
  - score_func = mutual_info_regression
  - percentile = 75

- **Model: GradientBoostingRegressor**
  - n_estimators = 300
  - learning_rate = 0.1
  - max_depth = 3
  - random_state = 42
  - loss = ‘squared_error’



Portfolio construction

- Stock Price Predictions:
  - Used trained models to predict future stock prices based on historical trends and financial indicators.
  - Computed predicted returns using the formula:

$$\frac{\text{Predicted Price} - \text{Current Price}}{\text{Current Price}}$$

- Final Stock Selection:
  - Sorted stocks by final score, which combined predicted returns and other relevant metrics.
  - Selected the top-performing stocks based on this ranking for portfolio inclusion.

Back Testing & Validation

- Validated model predictions on the T2 dataset (future stock data).
- Compared predicted vs. actual stock prices to measure performance.
- Used scatter plots, heatmaps, and bar charts to visualize model accuracy.
- Refined the model based on validation results to improve stock selection accuracy.

Top 5 Performing Stocks

BSE Code	NSE Code	Company Name	Units
544105	--	Harshdeep	322
543244	--	Shine Fashions	112
538882	--	Emerald Finance	641
543828	--	Sudarshan Pharma	246
543766	--	Ashika Credit	238

For the final portfolio, click here ↓

Final Portfolio

Team 05 Jan 2025

Thank You For Your Attention