Part 1: Technical Assessment

Alternate data used :

a) Data Exploration and Cleaning

2 out of 3 of the datasets shared (i.e. RBIB Table No. 18 _ Consumer Price Index (Base 2010=100).xlsx, simulated_auto_sales.csv) had the Time defining column defined on a monthly basis. Moreover, most of the macro economic indicators use a monthly scale to record the data values and hence I've followed according to the monthly scale.

- To ensure that the date scale remains common across all the datasets, the 'Date' or equivalent columns in the datasets were converted to datetime format for easier manipulation and then the date was reformatted to a consistent 'MM-YYYY' format to facilitate merging later.

Data Transformation

1. Dataset 1 : Consumer Price Indices (CPIs, HICPs), COICOP 1999 Consumer Price Index Total for India.csv (alternative data)
    - Consumer Price Index : Measures the average change in prices paid by consumers over time for a basket of goods and services
    - Steps followed: Date formatting, and retention of only the 'Month' and 'CPI' columns.
2. Dataset 2 : Producer Price Index by Commodity Metals and Metal Products Iron and Steel.csv (alternative data)
    - Producer Price Index : Measures the average change over time in the selling prices received by domestic producers of goods and services
    - Steps followed: Date formatting, and Retention of only the 'Month' and 'PPI' columns.
3. Dataset 3 : HBS Table No. 43 _ Major Monetary Policy Rates and… (Repo, Reverse Repo and MSF) Rates, CRR & SLR.pdf (given)
    - Columns retained: Bank Rate, Fix Range LAF Repo Rate, Fix Range LAF Reverse Repo Rate, Cash Reserve Ratio
    - Columns dropped: Standing Deposit Facility (SDF) Rate, Marginal Standing Facility (MSF), Statutory Liquiditory Ratio (SLR) due to very scarce data in the dataset and non availability of alternate datasets for these variables in the internet
    - Steps followed:

    - The 'Effective Date' column was converted to datetime.
    - Rows with invalid dates were dropped, and a new 'Month' column was created for grouping.
    - A custom function was defined to merge rows by month using the median for numeric columns.

4. Dataset 4 : Indian Rupees to U.S. Dollar Spot Exchange Rate.csv (alternative data)
    - Steps followed: Date formatting
5. Dataset 5 : RBIB Table No. 18 _ Consumer Price Index (Base 2010=100).csv (given)
    - Columns retained: Rural Index, Rural Inflation (%), Urban Index, Urban Inflation (%), Combined Index, Combined Inflation (%)
    - Columns dropped: Commodity Description, Provisional/Final as they were not meaningful, non-categorical columns
6. Dataset 6 : auto_sales.csv (alternative data)
   I initially used the dataset provided, simulated_auto_sales.csv , to train the model. But when generated the Pearson Correlation numbers, I found that there was a negative correlation between this data and the stock price which shouldn't be the case and hence cross verified multiple sites on the data and obtained a verified dataset from a reliable online source.
7. Dataset 7 : gdp.csv
    - Steps followed: Date formatting
8. Dataset 8 : consumer confidence interval.csv
    - Steps followed: Date formatting
9. Dataset 9 : Gasoline Price (USD/Litre).csv
    - Steps followed: Date formatting
10. Dataset 10 : Stock Price.csv
    - Steps followed: Date formatting, Cleaning stock values (text to numeric by removing comma)

Merging datasets

- Objective: To create a comprehensive dataset that includes all relevant economic indicators.

- Implementation:
    - Each processed DataFrame was merged on the 'Month' column using an outer join to ensure that all data points were included, even if some datasets had missing values for certain months.
    - The merged DataFrame was sorted by month to maintain chronological order.

Handling missing values

- Objective: To ensure the integrity of the dataset by addressing missing values.
- Implementation:

a) For the columns 'CPI', 'PPI', 'Exchange Rate', 'Rural Index', 'Rural Inflation (%)', 'Urban Index', 'Urban Inflation (%)', 'Combined Index', 'Combined Inflation (%)', and 'Sales', forward fill and if necessary backward fill was used. Values were missing only for the last few months and not significant variances were observed in these columns towards the end of the dataset, data rather remained quite stale. Hence forward/backfill used.
b) For the columns 'GDP growth rate' and 'Consumer Confidence Index', linear interpolation was used. Cross verified from the plots of the varying of the data values across time from multiple websites that were continuous.

Handling outliers

- Objective: To clean the dataset by removing outliers that could skew analysis.
- Implementation:
    - The Interquartile Range (IQR) method was used to identify outliers in numeric columns.
    - A mask was created to filter out rows that contained outlier values, resulting in a cleaner dataset for analysis.

Standardizing across multiple datasets

StandardScalar was used to scale the values in the columns in the merged dataset so that variance in the scale of the values wouldn't affect the MSE

b) Data Analysis

- Pearson correlation was done first to find how correlated the variables were. And to drop if any poorly co related values were present.

```
Closing Price                  1.000000
Rural Index                    0.845117
actions                        0.837532
Urban Index                    0.827736
Sales                          0.821746
Exchange Rate                  0.746052
CRR                            0.719498
Bank Rate                      0.658722
Rural Inflation (%)            0.462236
PPI                            0.408811
Gasoline Price (USD/Litre)     0.320906
Combined Inflation (%)         0.308890
Repo Rate                      0.299242
GDP growth rate                0.286710
Consumer Confidence Index      0.147043
Urban Inflation (%)            0.009769
CPI                           -0.612165
Reverse Repo Rate             -0.700489
Name: Closing Price, dtype: float64
```

Since there was almost 0 correlation between Urban Inflation (%) and Closing Price, I dropped the column. Negatively correlated values (i.e. CPI, Reverse Repo Rate both of which indicate about inflation) were kept intact to learn how they affect the Closing Price inversely.

- Following, I did a Granger Causality Test and ranked the variables according to their minimum p value. These would be the predictive variables, ranked.
- Predictive variables (threshold 0.25): CPI, CRR, Gasoline Price, Bank Rate, Rural Inflation, Reverse Repo Rate, Consumer Confidence Index, Repo Rate, Exchange Rate
- The trends of the first 5 predictive variables against Closing Prices are plotted out in the code.

```
Predictive Variables by Granger p-value (lower is better):
CPI: min p = 0.009
CRR: min p = 0.063
Gasoline Price (USD/Litre): min p = 0.078
Bank Rate: min p = 0.0794
Rural Inflation (%): min p = 0.1301
Reverse Repo Rate: min p = 0.1343
Consumer Confidence Index: min p = 0.1378
Repo Rate: min p = 0.1906
Exchange Rate: min p = 0.2308
Combined Inflation (%): min p = 0.2799
Urban Index: min p = 0.3781
PPI: min p = 0.4855
Combined Index: min p = 0.5528
Rural Index: min p = 0.5863
GDP growth rate: min p = 0.7332
```

c) Signal Creation

- Signal 1 : Real Monetary Stress Index

Quantifies monetary pressure using real interest rates, liquidity (CRR), and currency stress.

- Higher Repo Rates → costlier borrowing → lower investment/spending.
- CRR (Cash Reserve Ratio) → tighter liquidity → slower growth.
- INR depreciation (Exchange Rate) → costlier imports (e.g., auto components, metals).
- RMSI captures macro tightening, which negatively affects asset prices.

Let

- RtR_tRt = Repo Rate
- ππt\pi_tππt = Combined Inflation (%)
- CRRt\text{CRR}_tCRRt = CRR
- FXt\text{FX}_tFXt = Exchange Rate (USD/INR)

Then:

RMSIt=(Rt−πt)+CRRtFXt\text{RMSI}_t = \frac{(R_t - \pi_t) + \text{CRR}_t}{\text{FX}_t}RMSIt=FXt(Rt−πt)+CRRt

- Signal 2 : Consumer Demand Stress Oscillator (CDSO)

Measures stress on consumer demand based on price inflation vs. confidence and rural/urban split.

  - Rising inflation and falling confidence → suppressed demand
  - Rural vs. Urban comparison shows stress in core Tata Motors markets (rural for commercial vehicles, urban for PVs)
  - Increased gasoline prices -> increase costs of petrol/diesel lowers demand/willingness to spend on/use/buy automobiles
  - Captures elasticity and macro headwinds to consumer purchasing power.

CDSOt=CPIt−CRRt+γ⋅GasolinePricetCCIt\text{CDSO}_t = \frac{\text{CPI}_t - \text{CRR}_t + \gamma \cdot \text{GasolinePrice}_t}{\text{CCI}_t}CDSOt=CCItCPIt−CRRt+γ⋅GasolinePricet

Where:

  - CPI = Consumer Price Index → proxy for inflation
  - CRR = Cash Reserve Ratio → affects liquidity & interest rates
  - GasolinePrice = Fuel price (USD/Litre) → direct consumer burden
  - CCI = Consumer Confidence Index → reflects sentiment

- γ (gamma) = scaling factor for gasoline (to match units), e.g., 10 or 100

Interpretation:

- High CDSO → High inflation, expensive fuel, and tight liquidity relative to poor consumer confidence → high financial stress on consumers.
- Low CDSO → Low inflation/fuel, looser monetary policy, strong confidence → low stress, possibly higher auto demand.

Why it's predictive for Tata Motors:

- High CDSO → lower expected revenue/sales → potential negative impact on returns.
- Low CDSO → tailwind for demand → potential positive impact.

- Signal 3 : CPI Acceleration Signal

CPI (Consumer Price Index) reflects consumer inflation. A rising rate of CPI acceleration (i.e., second derivative) might precede tighter monetary policy or reduced purchasing power, which often negatively affects auto sector stock returns.

Mathematical Formulation:

CPI_Accelt=CPIt−2·CPIt−1+CPIt−2\text{CPI_Accel}_t = \text{CPI}_t - 2 \cdot \text{CPI}_{t-1} + \text{CPI}_{t-2}CPI_Accelt=CPIt−2·CPIt−1+CPIt−2

Reasoning:

It captures the acceleration or curvature in the CPI trend over time. It is a second-order difference or discrete approximation of the second derivative of the CPI.

- First-order difference (e.g., CPIt−CPIt−1\text{CPI}_t - \text{CPI}_{t-1}CPIt−CPIt−1) captures momentum - is CPI increasing or decreasing?
- Second-order difference (like the formula above) tells us how that change itself is changing - i.e., is CPI speeding up or slowing down?

Interpretation:

- If CPI_Accel > 0: CPI is accelerating upwards -> inflation is increasing faster than before.
- If CPI_Accel < 0: CPI is decelerating -> inflation may still be rising, but more slowly, or even starting to fall.
- If CPI_Accel ≈ 0: Inflation trend is linear or stable, not changing its slope.

Why it matters for stocks (e.g., Tata Motors):

- Rapidly accelerating CPI might signal strong inflation pressure, possibly leading to higher interest rates → negatively impacts auto sales & margins.
- Decelerating CPI might indicate easing inflation, which is often positive for consumer spending and borrowing.

d) Model Development (Stock Predictor)

Finalized Model : XGBoost + L1, L2 Regularization + Time features

Lasso and Ridge Regularization were added to reduce overfitting. Time features were deployed to capture variance over time (gives importance to the time period at which a certain variance occured) and predict possible variations in the future accurately.

Result :

- Train Set Performance:

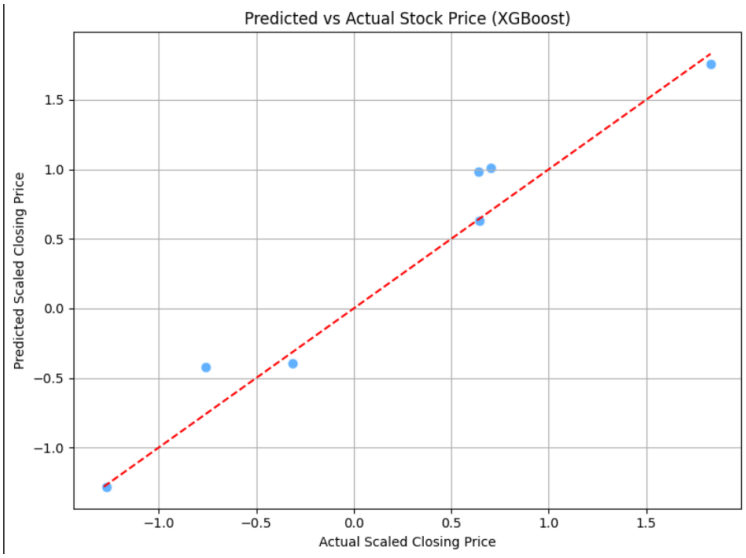
R^2: 0.9879 | MAE: 0.0734 | RMSE: 0.1114

- Validation Set Performance:

R^2: 0.9350 | MAE: 0.1762 | RMSE: 0.2142

- Test Set Performance:

R^2: 0.9489 | MAE: 0.1668 | RMSE: 0.2203



Other Models tried: XGBoost + Time Features + Bagging, LighBGM, LSTM (all weren't as efficient as the above)

XGBoost + Time Features + Bagging

- Train Set Performance:

R^2: 0.9601 | MAE: 0.1466 | RMSE: 0.2026

- Validation Set Performance:

R^2: 0.9232 | MAE: 0.1748 | RMSE: 0.2328

- Test Set Performance:

R^2: 0.8895 | MAE: 0.2791 | RMSE: 0.3240

LightGBM

- Train Set Performance:

R^2: 0.8194 | MAE: 0.3284 | RMSE: 0.4309

- Validation Set Performance:

R^2: 0.6909 | MAE: 0.3837 | RMSE: 0.4671

- Test Set Performance:

R^2: 0.7865 | MAE: 0.4252 | RMSE: 0.4504

e) Model Development (Quarterly Revenue Predictor)

Preprocessing

- Multi collinearity reduction was done using Upper Triangular Correlation Matrix and columns 'Rural Index', 'Urban Index', 'Combined Index', 'Combined Inflation (%)' were dropped

Finalized Model : XG Boost + GridSearchCV

Result

- Train Set:

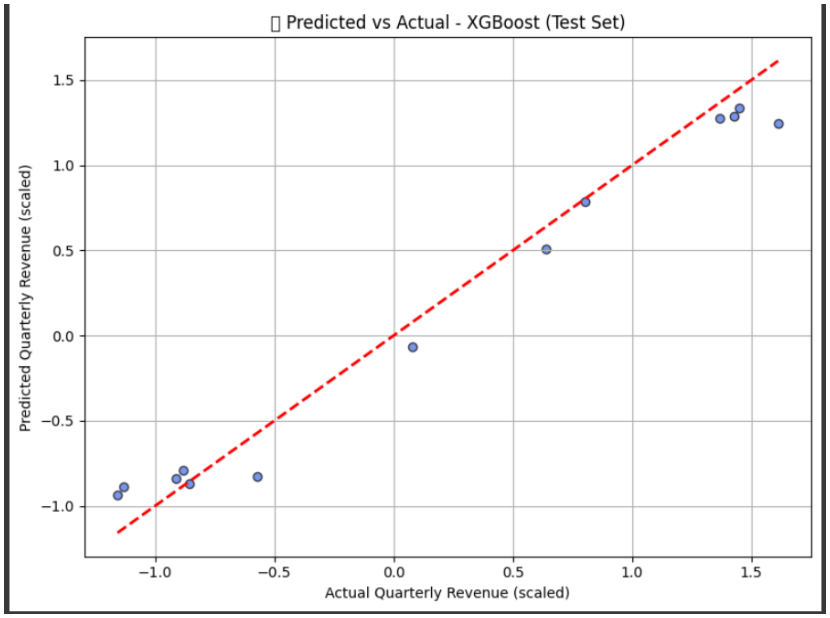
R^2: 0.9863 | MAE: 0.0696 | RMSE: 0.1113

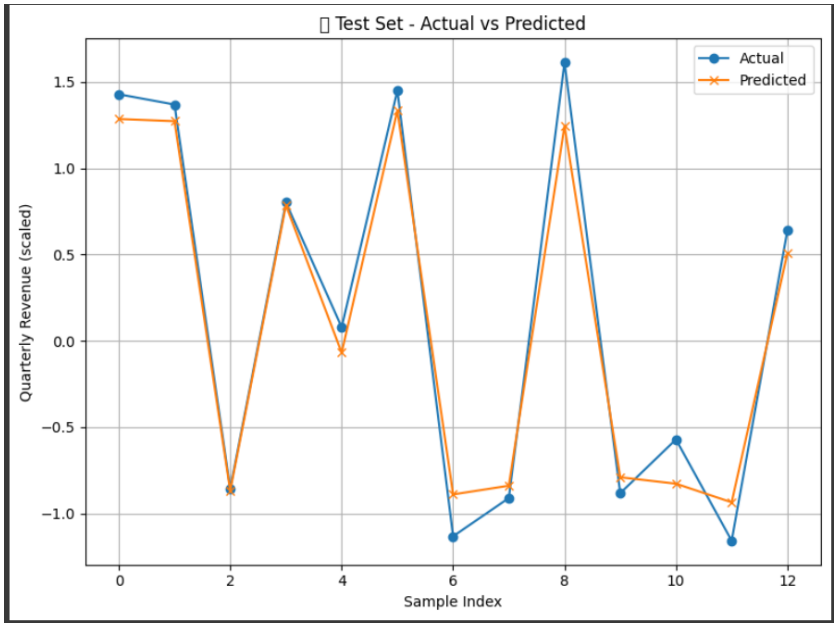
- Validation Set:

R^2: 0.9411 | MAE: 0.2550 | RMSE: 0.2751

- Test Set:

R^2: 0.9724 | MAE: 0.1473 | RMSE: 0.1766





f) Fundamental Validation

Signal 1: Real Monetary Stress Index (RMSI)

Why It Matters:

- High repo rate increases borrowing cost → higher interest expense for consumers and corporates

- Higher CRR limits liquidity → banks lend less → slower credit growth
- INR depreciation (↑ FX rate) raises import costs of auto parts → compresses Tata's margins

Effect on Margins:

- ↑ RMSI → Macroeconomic tightening → input cost pressure + demand slowdown → ↓ margins
- ↓ RMSI → Easier credit, cheaper imports → ↑ margins

Signal 2: Consumer Demand Stress Oscillator (CDSO)

Why It Matters:

- Higher inflation and gasoline prices reduce disposable income
- Low consumer confidence reflects weak intent to purchase vehicles
- Tight liquidity (↑ CRR) → credit constraints on auto loans

Effect on Margins:

- ↑ CDSO → Falling demand → lower pricing power → discounts, higher inventory costs → ↓ margins
- ↓ CDSO → Stronger demand → better pricing → ↑ margins

Signal 3: CPI Acceleration (CPI_Accel)

Why It Matters:

- Acceleration of inflation (↑ CPI_Accel) predicts policy tightening, cost escalation
- Deceleration may signal easing, boosting consumer spending

Effect on Margins:

- ↑ CPI_Accel → Higher input costs, reduced demand → ↓ margins
- ↓ CPI_Accel → Stability, potential rate cuts → ↑ margins

Part 2: Coding Challenge

Signals :

1. AVMD (Adjusted Volume Momentum Differential)

Captures unusual volume-driven momentum by adjusting recent return momentum with volume spikes.

Let:

- RtR_tRt: Return at time t
- VtV_tVt: Volume at time t
- μV\mu_VμV: Mean volume
- σV\sigma_VσV: Std. deviation of volume

Then:

$$AVMD_t = R_t \times \left(\frac{V_t - \mu_V}{\sigma_V}\right)AVMDt=Rt×(σVVt−μV)$$

Interpretation:

- High AVMD → Strong positive return with abnormally high volume → institutional accumulation → bullish signal
- Low or negative AVMD → Weak returns or strong volume on negative price days → distribution or bearish signal

Why it matters:

- Volume-supported price movement is often considered more trustworthy in technical analysis.
- In the context of Tata Motors, sudden volume surges might indicate news-driven moves, earnings reactions, or macro events.

2. EWIPI (Economic-Wide Inflation Pressure Index)

Measures inflationary heat in the economy by combining CPI and PPI, capturing both consumer-level and producer-level inflation pressures.

$$EWIPI_t = \alpha \cdot CPI_t + (1 - \alpha) \cdot PPI_tEWIPIt=α·CPIt+(1−α)·PPIt$$

Where:

- α\alphaα is a weighting factor (e.g., 0.6 for consumer-driven economy)

Interpretation:

- High EWIPI → Both consumers and producers are facing high inflation → likely margin pressure on firms
- Low EWIPI → Inflation is under control → favorable for stable cost structures

Why it matters:

- Increases in input (PPI) and output (CPI) costs can squeeze margins or shift consumer behavior.
- Tata Motors, being in the auto sector, is sensitive to both raw material prices (steel, semiconductors) and end consumer affordability.

3. DSES (Demand-Side Elasticity Stress Signal)

Measures stress on demand based on how real interest rates and fuel burden affect consumers' purchasing power.

Let:

- RIRt=RepoRatet−InflationtRIR_t = \text{RepoRate}_t - \text{Inflation}_tRIRt=RepoRatet −Inflationt
- GPt=GasolinePricetGP_t = \text{GasolinePrice}_tGPt=GasolinePricet DSESt=RIRt+γ·GPt\text{DSES}_t = RIR_t + \gamma \cdot GP_tDSESt=RIRt+γ·GPt Where γ\gammaγ is a scaling factor (e.g., 10–100 to match unit scales.

Interpretation:

- High DSES → Consumers are squeezed due to high interest rates and fuel prices → lower demand for discretionary goods like automobiles
- Low DSES → Easier borrowing & cheaper fuel → more willingness to spend

Why it matters:

- Tata Motors' passenger and commercial vehicles are fuel-sensitive discretionary purchases.
- Tight credit conditions and fuel inflation can sharply cut demand, especially in rural India.

Models

1. Random Forest Regressor (For Long-term trading)

```
ML Strategy Backtest:
Sharpe Ratio: 1.52
Max Drawdown: -1.52%
Annualized Return: 11.41%
Monthly Turnover: 8.90%
Estimated Daily Turnover: 0.40%
```

2. Random Forest Regressor + Optimizing threshold (For short-term trading)

```
✅ Optimal Threshold Found:
Threshold: 0.0100
Sharpe Ratio: 2.18
Avg Monthly Turnover: 27.59%
Estimated Daily Turnover: 1.31%
Max Drawdown: -11.68%
Annualized Return: 88.17%
```