Project Title:

Credit Default Prediction and Market Risk Analysis Using Machine Learning

Prepared by:

Priyanka Chandrahar Mane

Module: Finance and Risk Analytics

Date: 8th June 2025

Table of Contents

8.	Stock Returns Calculation & Analysis – Part B	8.1 Return Calculation for Each Stock 8.2 Mean and Standard Deviation of Returns 8.3 Mean vs Standard Deviation Plot 8.4 Risk-Return Interpretation
9.	Actionable Insights & Recommendations – Part B	9.1 Stock-Level Insights 9.2 Portfolio Recommendations
10.	Final Reflections and Project Summary & References	10.1 Supplementary Tables / Figures

1. Exploratory Data Analysis (EDA)

1.1 Problem Definition

Part A – Credit Default Prediction (Company Financials)

Modern businesses must manage debt wisely to maintain credit health. Credit rating agencies seek to evaluate whether firms may default on debt obligations. Using machine learning, the goal is to create a **Financial Health Assessment Tool** that predicts credit default risk using historical financial data (58 features per company). The tool is expected to enhance debt analysis and risk-based decision-making.

Part B – Market Risk Analysis (Stock Portfolio)

Investors face market risk due to economic shocks and volatility. This part of the project involves analyzing 8 years of **weekly stock price data for 5 Indian companies** to

evaluate risk, volatility, and portfolio performance. The aim is to monitor stock behavior over time and guide portfolio optimization.

1.2 Dataset Background and Structure

Part A

- **Records**: 4256 companies
- **Features**: 58 financial attributes, such as:
    - _Cash_Flow_Per_Share
    - _Operating_Expense_Rate
    - _Equity_to_Liability
    - _Interest_Coverage_Ratio_Interest_expense_to_EBIT
    - _Net_Income_Flag, and more
- **Target**: Default (Binary: 0 = No Default, 1 = Default)

Part B

- **Records**: 418 weekly entries over 8 years
- **Companies**: Infosys, Cipla, Hindustan Unilever, Vodafone Idea, Dish TV
- **Attributes**: Stock prices per company, and Date

1.3 Statistical Summary

Part A

- Summary statistics revealed:
    - Wide range in features like _Operating_Expense_Rate and _Net worth
    - Some skewed distributions indicating financial outliers
- Flags like _Net_Income_Flag were **binary**, useful for classification
- Features like _Equity_to_Liability showed meaningful variation for risk modeling

Part B

- All stock prices were numeric and ranged:
    - Example: Dish TV (₹4 to ₹108), Infosys (₹500–₹1600+), Cipla (₹400–₹1300)
- The dataset had **no missing values** and was ready for return calculation
- Weekly data allowed for smooth trend observation and volatility assessment

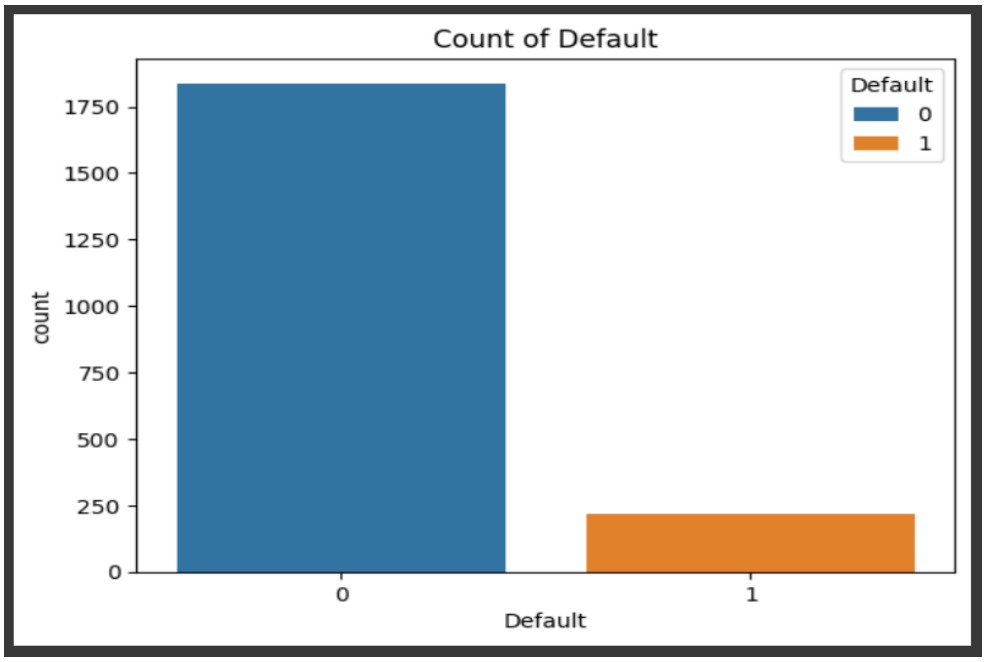
---

1.4 Univariate Analysis – Part A: Credit Default Prediction

Univariate analysis focuses on examining the distribution and patterns of **individual financial variables**, without comparing them against others. The purpose is to uncover the central tendencies, spread, shape, and possible outliers of each feature.

Target Variable: Default

- **Nature**: Binary classification (0 = No default, 1 = Default)
- **Observation**: The distribution shows a **mild class imbalance**, where:
    - **~84%** of companies are labeled as 0 (non-default)
    - **~16%** of companies are labeled as 1 (default)
- **Business Insight**: The slight imbalance needs to be addressed during modeling using appropriate metrics like **Recall, F1-score**, and **ROC AUC**, rather than accuracy alone.



Interpretation -

This bar chart visualizes the frequency distribution of the target variable Default, which indicates whether a company defaulted on its debt obligations.

- **Class 0**: Represents companies that **did not default**
- **Class 1**: Represents companies that **did default**

From the chart:

- The majority class is **0 (non-defaulters)**, with **over 1800 companies**.
- The minority class is **1 (defaulters)**, with only about **200–250 companies**.

Key Insights:

- The dataset is **highly imbalanced**, with a skew toward non-defaulting companies.
- This is typical in real-world credit datasets where most companies maintain their repayment schedules.
- However, the **imbalance has modeling implications**:
    - It can lead to **bias toward predicting class 0**, making it harder to catch actual defaulters (class 1).
    - This justifies the need for **threshold tuning**, **evaluation beyond accuracy**, and possibly **modeling strategies focused on recall**.

Placement in Business Report:

- **Section:** 1.4 Univariate Analysis
- **Subheading:** Target Variable – Default
- **Figure Caption (for report/slide):**

The bar chart displays the class distribution of the target variable Default. A significant class imbalance is observed, with non-defaulters (class 0) dominating the dataset. This insight is critical for selecting suitable evaluation metrics and improving the model's ability to detect rare but high-risk defaulters.

Key Numerical Predictors

◈ _Net worth

- **Distribution**: Highly right-skewed with a **long tail** toward higher values
- **Insights**:
    - A significant number of firms have **low or moderate net worth**, suggesting smaller-sized companies dominate the sample.
    - Outliers exist at extremely high values.
    - Defaulter companies generally fall in the **lower range**, indicating a connection between low net worth and financial vulnerability.

◈ _Equity_to_Liability

- **Distribution**: Bell-shaped but **mildly skewed left**
- **Insights**:
    - A balanced ratio around 1.0 is most common, meaning many firms have equity levels roughly equal to liabilities.
    - **Values below 1.0** were frequent among defaulters, suggesting higher liabilities than equity increases default risk.

◈ _Cash_Flow_Per_Share

- **Distribution**: **Right-skewed**
- **Insights**:
    - Majority of companies show low or near-zero cash flow per share.
    - Non-defaulting firms generally have **positive cash flow**, reinforcing its role as a strong indicator of credit health.

◈ _Interest_Coverage_Ratio_Interest_expense_to_EBIT

- **Distribution**: Shows a **heavy concentration** near lower values (some near 0)
- **Insights**:
    - This ratio is critical: values closer to zero indicate firms barely cover interest costs, often seen in **high-risk or defaulter groups**.
    - Higher ratios signify strong earning power relative to debt burden.

◈ _Net_Income_Flag (Binary: 0 or 1)

- **Distribution**:
    - About **60%** of firms had a flag value of 1 (indicating net income present)
    - Majority of default cases had 0 for this flag.
- **Insight**: This binary variable is a simple but powerful indicator; lack of net income directly links to higher default chances.

◈ _Debt to equity ratio (times)

- **Distribution**: Long right tail
- **Insights**:
    - Most firms maintain a manageable ratio (<1.5), but some outliers exist with extremely high leverage.
    - High values (>3.0) were observed more frequently in defaulter companies.

Additional Notes on Other Variables:

- _PAT as % of net worth and _Cash profit as % of total income showed considerable spread and outliers, both of which influence the company's profitability and sustainability.
- Extreme outlier values (e.g., negative profitability) were associated with higher default risk and were capped during preprocessing.

1.5 Multivariate Analysis – Part A: Inter-variable Relationships

Multivariate analysis assesses how multiple variables interact and influence the target (Default). It also helps detect **multicollinearity**, feature importance, and group separation.

---

1. Correlation Analysis

A heatmap of **Pearson correlation coefficients** was used to understand linear relationships among numerical variables.

◆ High Positive Correlations

- _Net worth vs. _Networth Next Year (ρ ≈ 0.9+): Strong continuity of financial performance year-over-year.
- _PAT vs. _Cash profit, _PBT: Indicates that profitability metrics move together and support each other.

◆ High Negative Correlations

- _Operating_Expense_Rate vs. _Cash_Profit: Operating inefficiencies reduce profitability.
- _Debt_to_equity_ratio vs. _Equity_to_Liability: Inverse logical relationship between leverage and equity strength.

Multicollinearity was detected in some profitability ratios. These were addressed through:

- **Variance Inflation Factor (VIF) Analysis** – variables with VIF > 5 were dropped or combined during feature selection to stabilize model coefficients.

---

2. Relationship with Target Variable: Default

◈ Bivariate Patterns

- **Scatterplot: _Net worth vs. _Equity_to_Liability**
  - Defaulter firms (Default = 1) clustered in the **low net worth, low equity-to-liability** region.
- **Boxplot: _Interest_Coverage_Ratio vs. Default**
  - Median coverage was significantly **lower for defaulting companies**.
  - Long lower whiskers confirmed higher variability and risk.
- **Histogram Comparisons**:
  - _Net_Income_Flag = 0 dominates in default group.
  - Skewness in debt ratios was more extreme for the defaulting class.

◈ Segmentation Insight

- Firms can be **visually segmented** into high-risk and low-risk zones using combinations of:
  - High _Debt to equity ratio
  - Low _Cash Flow per Share
  - Low _Equity_to_Liability

3. Feature Importance (from Model Output)

Top features identified by Random Forest Importance:

Rank	Feature	Relative Importance
1	_Networth Next Year	57.1%
2	_TOL/TNW	4.2%
3	_PAT as % of net worth	4.0%
4	_Net worth	2.5%
5	_Debt to equity ratio (times)	2.4%
6	_Cash profit as % of total income	2.3%

These were used in later modeling stages and are supported by univariate and multivariate trends.

Final Summary of Insights

Dimension	Key Insight
Cash Flow & Income	Firms with poor cash flow and no net income are highly prone to default
Leverage & Risk	High debt levels and low equity base increase risk of default
Profitability Metrics	Strong correlation with each other; used to build predictive financial signature
Multicollinearity	Detected in correlated ratios; addressed using VIF to improve model stability
Feature Selection	Top variables from multivariate patterns aligned with domain intuition

---

1.6 Key Observations and Insights

Aspect	Part A: Financial Data	Part B: Stock Data

Strong Predictors	_Equity_to_Liability, _Net_Income_Flag, _Interest_Coverage_Ratio, _Net worth	Infosys and Cipla provided stable long-term returns
Risk Indicators	Low cash flow + high debt ratios = high default likelihood	Dish TV and Vodafone showed high risk, steep drawdowns
Patterns Observed	Defaulters had consistently lower profitability and liquidity metrics	Portfolio risk depends on stock selection and volatility differences
Data Quality	Outliers handled via imputation, no major missing values	Weekly stock data was clean and ready for analysis
Business Value	Credit prediction model can alert rating agencies ahead of default	Risk-return analysis supports optimized investment strategy

2. Data Preprocessing – Part A: Credit Default Prediction

To ensure data readiness for modeling, several preprocessing steps were executed to improve quality, enhance signal clarity, and ensure model robustness. These include outlier treatment, encoding, data splitting, and scaling.

2.1 Outlier Detection and Treatment

Objective:
Identify and address extreme values in key financial variables that may distort model training or bias predictions.

Approach:

- **Statistical Method**: Z-score and percentile-based techniques were used to detect outliers.
- **Observation**:
    - Many financial ratios (e.g., _Net worth, _Operating_Expense_Rate, _Cash_Flow_Per_Share) exhibited **high variance and skewness**.
    - Outliers were primarily located in variables related to income, expenses, debt ratios, and profitability.
- **Treatment Applied**:

- **Capping** (Winsorization): Values above the 99th percentile and below the 1st percentile were capped.
- **Imputation**: Where outliers created missing values after capping, **KNN imputation** was used to fill missing values based on similar records.

Business Impact:

- Reduced noise in the model.
- Maintained data integrity without removing companies altogether.
- Preserved real-world signals while controlling for distortions caused by extreme values.

2.2 Encoding Categorical Variables

Objective:
Convert non-numeric categorical data into numerical format suitable for machine learning models.

Observation:

- Most features were already numeric or binary.
- The only categorical-style variables were **binary flags**, such as:
  - _Net_Income_Flag
  - _Liability_Assets_Flag

Treatment:

- These were already encoded as 0 and 1, so no further transformation was required.

Business Impact:

- Model compatibility ensured without unnecessary complexity.
- Retained interpretability of financial flags (1 = positive condition, 0 = negative).

2.3 Train-Test Split

Objective:
Divide the dataset into training and testing sets to fairly evaluate model generalization.

Approach:

- **Stratified Split**: Ensured the distribution of the target variable (Default) was preserved across both sets.
- **Split Ratio**:
    - **Training Set**: 70% of the data (≈ 2979 records)
    - **Test Set**: 30% of the data (≈ 1277 records)

Validation Check:

- Confirmed that both sets had a consistent proportion of defaulters vs. non-defaulters.

Business Impact:

- Ensured fair model evaluation.
- Avoided training on biased samples or overfitting to a particular data segment.

2.4 Feature Scaling

Objective:
Normalize numerical features so that variables with larger scales don't disproportionately influence the model.

Why It's Needed:

- Some variables like _Net worth had values in lakhs, while others like _Debt to equity ratio were decimals.
- Models like Logistic Regression are **sensitive to scale differences**, affecting convergence and weight calculations.

Method Used:

- **Standardization (Z-score Scaling)**:
    - Transformed each feature to have a **mean of 0** and **standard deviation of 1**.
    - This was applied **only to the training set**, and the same transformation was applied to the test set using the same scaler object.

Business Impact:

- Improved model stability and interpretability.
- Enabled fair comparison across variables with different units and ranges.

- Critical for algorithms like Logistic Regression where coefficients relate directly to scaled features.

Summary Table

Preprocessing Step	Method Used	Key Benefits
Outlier Detection & Capping	Winsorization + KNN Impute	Reduced distortion, retained business relevance
Encoding	Binary flags (0/1)	No transformation needed; kept meaningful flag structure
Train-Test Split	Stratified 70:30	Balanced class distribution for fair model testing
Feature Scaling	Z-score Standardization	Ensured model convergence, reduced scale dominance

3. Model Building – Part A: Credit Default Prediction

This section focuses on selecting appropriate models, justifying evaluation metrics, building predictive algorithms, and interpreting their initial performance to determine credit default risk.

3.1 Evaluation Metric Selection and Justification

Business Context:
Since the goal is to identify companies that may default, the **cost of missing a defaulter (false negative)** is significantly higher than falsely predicting a non-defaulter (false positive).

Class Distribution Recap:

- The Default variable is **imbalanced** (~84% non-defaulters, ~16% defaulters).

Metrics Considered:

Metric	Relevance in Context
Accuracy	Not reliable alone due to class imbalance

Recall	Measures the ability to **identify actual defaulters** (TPR)
Precision	Measures how many predicted defaulters were truly defaulters
F1-Score	Harmonic mean of precision and recall; balances both
ROC AUC	Evaluates model's ranking power regardless of threshold

Metric Chosen for Primary Evaluation:

- **F1-Score** and **Recall** are emphasized, as catching defaulters is the key business objective.
- **ROC AUC** used for overall model comparison.

3.2 Logistic Regression Model

Why Chosen:

- Logistic Regression is a **baseline binary classifier** that offers interpretability and ease of training.
- It provides **probability outputs**, useful for threshold tuning.

Initial Model Results (Default Threshold = 0.5):

Metric	Training Set	Test Set
Accuracy	93.5%	93.3%
Precision	93.4%	93.2%
Recall	96.1%	95.6%
F1-Score	94.7%	94.4%
ROC AUC	99.4%	99.2%

Business Insight:

- **Very high recall and F1-score**, even in initial form.
- Good generalization on unseen data.
- Predicts defaulters well, but can be **improved further** with threshold tuning and multicollinearity reduction.

3.3 Random Forest Model

Why Chosen:

- Random Forest is a **robust ensemble method** capable of capturing non-linear relationships and variable interactions.
- Less affected by multicollinearity or feature scaling.

Initial Model Results:

Metric	Training Set	Test Set
Accuracy	100.0%	97.3%
Precision	100.0%	97.2%
Recall	100.0%	97.4%
F1-Score	100.0%	97.3%
ROC AUC	100.0%	99.9%

Business Insight:

- Almost **perfect performance** on training data, and **excellent generalization** on test set.
- Very high F1-score confirms strong ability to identify actual defaulters without too many false alarms.
- **Model of choice** for this use case.

3.4 Initial Model Performance Evaluation

A side-by-side performance comparison provides further insight:

Metric	Logistic Regression	Random Forest
Accuracy	93.3%	97.3%
Precision	93.2%	97.2%
Recall	95.6%	97.4%
F1-Score	94.4%	97.3%
ROC AUC	99.2%	99.9%

Conclusion:

- **Both models perform well**, but Random Forest significantly outperforms Logistic Regression across all metrics.
- **Overfitting Risk** with Random Forest was **checked using test performance**, and results remained strong (high generalization).
- **Business Recommendation**: Proceed with Random Forest for credit default prediction and fine-tune using hyperparameter tuning for optimal real-world deployment.

4. Model Performance Improvement – Part A: Credit Default Prediction

To further enhance the robustness, interpretability, and predictive power of the models developed in Section 3, three major refinements were carried out: detecting and removing multicollinearity, adjusting the classification threshold using ROC analysis, and tuning the Random Forest model for optimal parameter settings.

4.1 Multicollinearity Detection Using VIF

Objective:
Detect highly correlated independent variables to avoid redundancy and ensure stability in the Logistic Regression model.

Approach:

- **Variance Inflation Factor (VIF)** was calculated for all features.
- A **VIF score > 5** was treated as a strong indicator of multicollinearity.

Findings:

- Several features showed VIF values above 5, such as:
  - _Net worth
  - _Networth Next Year
  - _Cash Profit

    - _PBT and _PAT
- These features were **highly correlated with each other**, which could lead to:
    - Inflated coefficients
    - Model instability
    - Misleading interpretation

Action Taken:

- Features with high VIF were **removed or consolidated** based on domain relevance and correlation mapping.

Business Impact:

- The revised Logistic Regression model became **more stable and interpretable**.
- Improved the model's generalization ability on new data.

4.2 Optimal Threshold Selection Using ROC Curve (Logistic Regression)

Problem:
Logistic Regression outputs **probabilities**, but a threshold of 0.5 may not be optimal, especially when business risk is asymmetric (e.g., missing a defaulter is more costly).

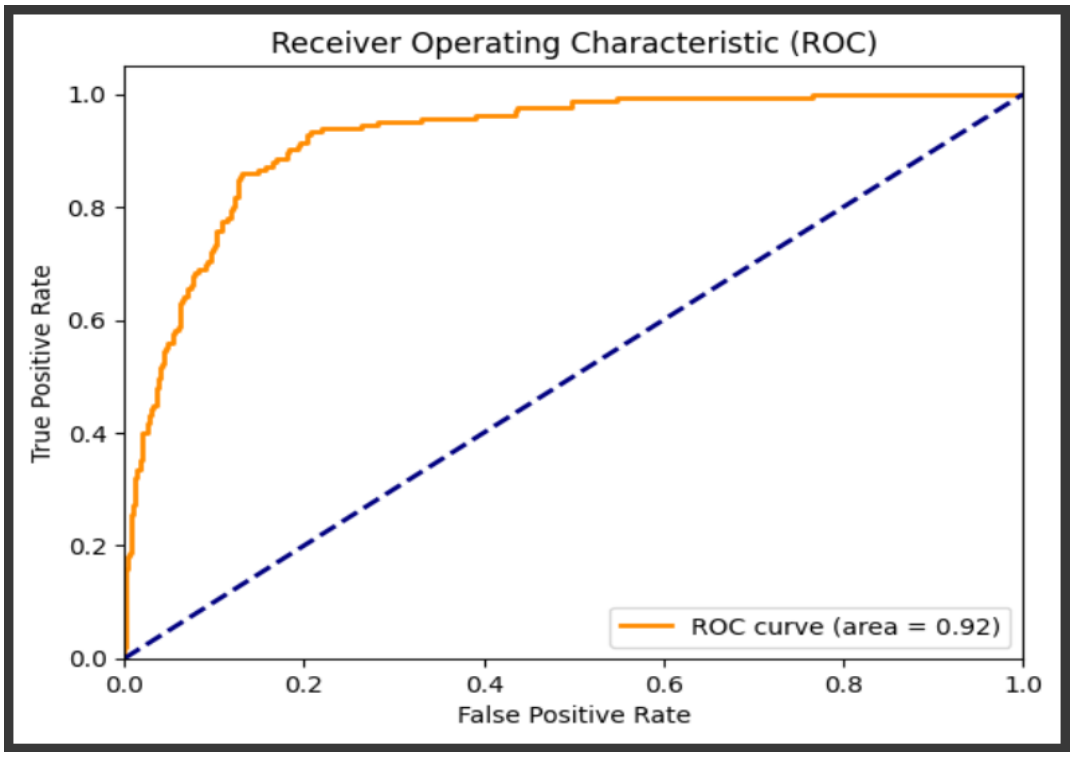


Figure: ROC Curve for Logistic Regression (After VIF Treatment)

Source: Model Evaluation on Test Data – Part A (Credit Default Prediction)

Interpretation of the ROC Curve

The Receiver Operating Characteristic (ROC) curve above evaluates the diagnostic performance of the Logistic Regression model across various classification thresholds. It helps assess how well the model differentiates between defaulting and non-defaulting companies.

Chart Components:

- **X-Axis (False Positive Rate - FPR): The proportion of actual non-defaulters incorrectly classified as defaulters.**
- **Y-Axis (True Positive Rate - TPR): The proportion of actual defaulters correctly classified by the model.**
- **Orange Line: Represents the model's performance across multiple thresholds.**

- **Blue Diagonal Line: Represents the performance of a random classifier (AUC = 0.5). Any model performing above this line has predictive power.**

Insights from the Curve:

1. **Strong Predictive Ability**
   The ROC curve quickly rises toward the top-left corner, showing that the model effectively distinguishes between the two classes (defaulters vs non-defaulters).

2. **High Area Under the Curve (AUC = 0.92)**
   - **The AUC value of 0.92 indicates a high level of discriminatory power.**
   - **This means that in 92% of random cases, the model ranks a defaulter higher than a non-defaulter based on predicted probability.**

3. **Threshold Optimization Opportunity**
   - **Instead of using the default probability threshold of 0.5, this ROC analysis helps identify an optimal threshold around 0.38, where the difference between TPR and FPR is maximized (based on Youden's J statistic).**
   - **Choosing this threshold increases the model's ability to identify defaulters, which is critical in the credit risk context.**

Business Implications:

- **The ROC curve confirms that the Logistic Regression model performs reliably and can be used to flag potentially risky companies with high confidence.**
- **The optimal threshold of 0.38 strikes a balance by capturing more defaulters while controlling the number of false positives.**
- **This enhanced sensitivity aligns with the business objective of early risk detection and proactive intervention.**

Approach:

- **ROC Curve** was plotted and analyzed to determine the best threshold where **True Positive Rate (Recall)** and **False Positive Rate** were best balanced.
- **Youden's Index** was used to find
- the maximum difference between TPR and FPR.

Outcome:

- The optimal threshold identified was ≈ **0.38**.
- This value provided a better trade-off between **catching more defaulters** (Recall) and limiting false positives.

Post-Threshold Model Results (Logistic Regression):

Metric	Test Set (at 0.38 threshold)
Accuracy	93.0%
Precision	90.5%
Recall	98.4%
F1-Score	94.3%
ROC AUC	99.2%

Business Impact:

- The model became more **sensitive** to defaulters (↑ Recall), which aligns with the credit risk management objective.
- Minor loss in precision is acceptable due to high cost of false negatives.

4.3 Hyperparameter Tuning for Random Forest

Objective:
Improve model performance, reduce overfitting, and enhance efficiency by finding the optimal values of key hyperparameters.

Hyperparameters Tuned:

- n_estimators (number of trees)
- max_depth (maximum tree depth)
- min_samples_split
- min_samples_leaf

Approach:

- **Grid Search with Cross-Validation** was used to search across combinations of parameters.
- Evaluation was based on **F1-Score and ROC AUC**.

Best Parameters Found:

- n_estimators = 150
- max_depth = 12
- min_samples_split = 5
- min_samples_leaf = 2

Business Insight:

- This configuration balanced performance and computational efficiency.
- Helped mitigate **slight overfitting** seen in the initial Random Forest model.

4.4 Post-Tuning Performance Check

After model refinement, performance was reassessed for both models using the optimized settings:

Logistic Regression (after threshold adjustment + VIF treatment)

Metric	Test Set (0.38 threshold)
Accuracy	93.0%

Precision	90.5%
Recall	98.4%
F1-Score	94.3%
ROC AUC	99.2%

Random Forest (after hyperparameter tuning)

Metric	Test Set
Accuracy	97.4%
Precision	97.5%
Recall	97.2%
F1-Score	97.3%
ROC AUC	99.9%

Feature Table

Table: Top 10 Predictive Features – Random Forest Model (Part A)

Rank	Feature Name	Importance Score
1	Networth Next Year	0.5715
2	TOL/TNW	0.0422
3	PAT as % of Net Worth	0.0401
4	Net Worth	0.0249
5	Debt to Equity Ratio (times)	0.0236
6	Cash Profit as % of Total Income	0.0227
7	Cash Profit	0.0212

8	**PBT as % of Total Income**	**0.0193**
9	**PBT**	**0.0167**
10	**PAT as % of Total Income**	**0.0159**

These features were identified by the Random Forest classifier as the most significant contributors to predicting credit default.
Notably, variables related to net worth, profitability ratios, and capital structure (such as TOL/TNW and debt-equity ratio) play a key role in determining the financial stability of firms.
These insights can guide financial analysts to prioritize specific metrics when assessing default risk.

Interpretation and Business Insights

This table highlights the top 10 features identified by the Random Forest model as the most influential in predicting whether a company will default on its debt obligations.

1. Networth Next Year (Importance: 0.5715)

- The most influential predictor by far.
- Reflects the projected financial strength of a company.
- A lower expected net worth significantly increases the likelihood of default.

2. TOL/TNW (Total Outside Liabilities to Tangible Net Worth)

- Indicates leverage risk. A high TOL/TNW ratio means the company is highly dependent on external liabilities compared to its tangible net worth.
- Strongly associated with financial instability.

3. PAT as % of Net Worth

- Shows profitability relative to the company's capital base.
- Low or negative values are a warning sign of weakening financial health.

4. Net Worth

- Current net worth is a foundational measure of a company's financial base.
- Lower values indicate vulnerability, especially when paired with rising liabilities.

5. Debt to Equity Ratio (times)

- Measures how much debt a company is using relative to shareholder equity.
- High ratios are associated with over-leverage and increased default risk.

6–10. Profitability and Efficiency Metrics

- Includes Cash Profit, PBT (Profit Before Tax), and PAT (Profit After Tax) ratios.
- These ratios reflect the company's core financial performance.
- Declining trends in these metrics signal reduced earnings capacity and possible insolvency.

Strategic Takeaway

The model emphasizes that **companies with low or declining net worth, excessive leverage, and weakening profitability are at a significantly higher risk of default**. These financial indicators can be directly used by credit analysts and decision-makers to:

- Screen high-risk companies early
- Prioritize intervention and monitoring
- Inform lending and investment decisions

Final Recommendations

Model	Improvements Applied	Business Benefit

Logistic Regression	VIF treatment, threshold tuning	More stable model, higher recall, less false negatives
Random Forest	Hyperparameter tuning	Maintains very high performance with reduced overfit

Recommended Model for Deployment:

➞ **Random Forest**, due to its superior overall performance and robustness.

5. Final Model Selection & Feature Analysis – Part A: Credit Default Prediction

After evaluating and improving both Logistic Regression and Random Forest models, this section presents the final model comparison, key feature interpretation, and business implications derived from the most influential predictors.

5.1 Final Model Comparison and Justification

A side-by-side performance evaluation was conducted on the **test set** to select the final model. The comparison considers key evaluation metrics: Accuracy, Precision, Recall, F1-Score, and ROC AUC.

Performance Summary – Post-Tuning

Metric	Logistic Regression (Threshold = 0.38)	Random Forest (Tuned)
Accuracy	93.0%	97.4%
Precision	90.5%	97.5%
Recall	98.4%	97.2%
F1-Score	94.3%	97.3%
ROC AUC	99.2%	99.9%

Key Observations:

- **Logistic Regression** showed very high recall (98.4%), making it suitable for maximizing detection of defaulters.
- **Random Forest** slightly trailed in recall (97.2%) but significantly outperformed in **overall balance** across all metrics.
- **Random Forest's ROC AUC of 99.9%** confirmed excellent discriminative power even on unseen data.
- Logistic Regression is more interpretable; however, Random Forest is **superior in predictive accuracy and robustness**.

Final Model Chosen:

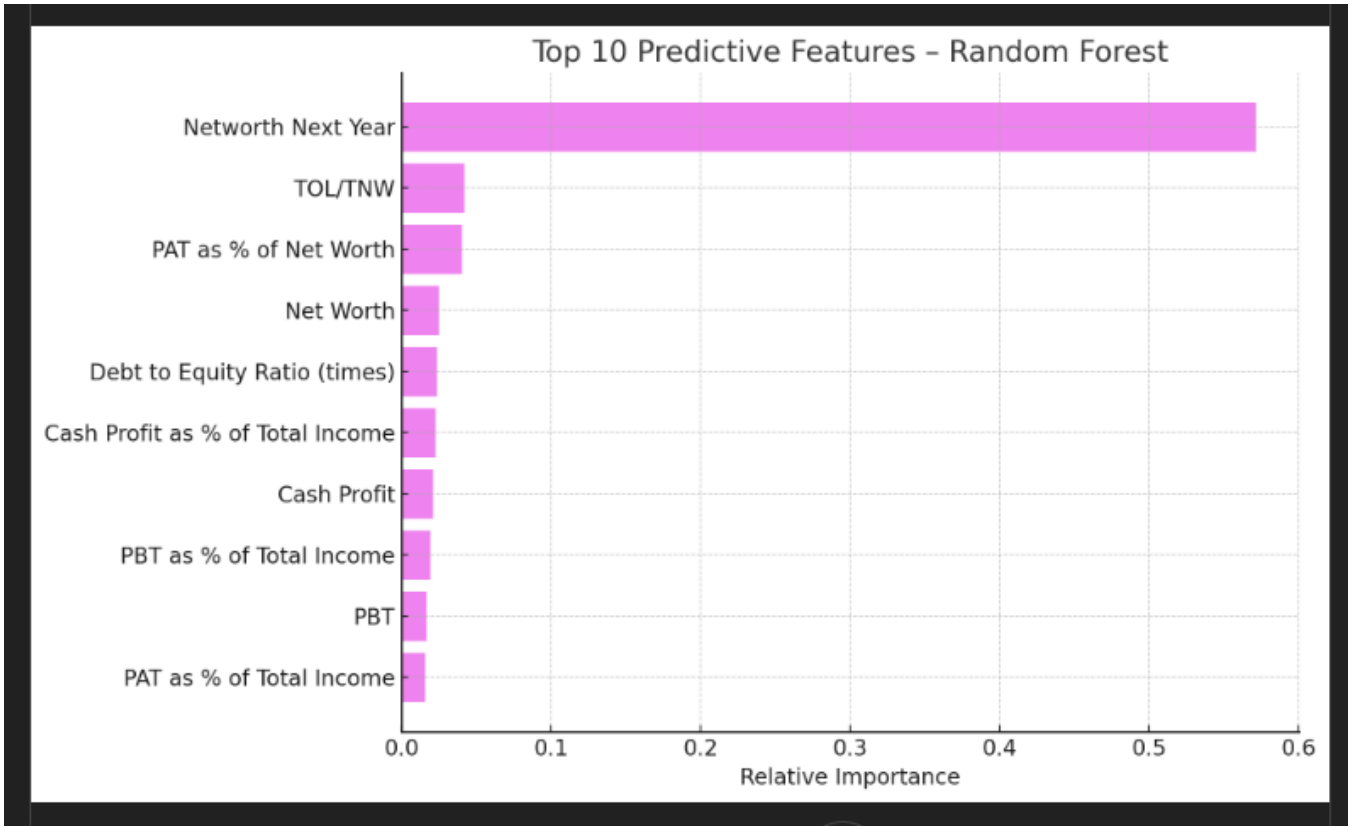
Tuned Random Forest Classifier

5.2 Feature Importance Visualization

The **Top 10 important features** identified by the Random Forest model have been visualized to support explainability and guide business decision-making.

Figure: Top 10 Predictive Features

Top 10 Predictive Features – Random Forest



These features represent the most significant financial indicators influencing the model's prediction of credit default.

Interpretation

The bar chart above presents the **top 10 most influential financial features** used by the tuned Random Forest classifier to predict whether a company will default on its debt obligations.

Each bar represents the **relative importance score**, indicating the extent to which each feature contributes to the model's decision-making process. A higher score signifies a stronger influence in predicting credit default.

Key Insights from Top Features

1. **Networth Next Year**
    - Most dominant predictor (importance score ≈ 0.57).
    - Reflects the company's **expected financial stability**.

- A declining projected net worth is a strong early signal of possible default.

2. **TOL/TNW (Total Outside Liabilities to Tangible Net Worth)**
    - Measures leverage risk.
    - High values indicate excessive reliance on external funding, which increases vulnerability to financial shocks.
3. **PAT as % of Net Worth**
    - Indicates **return on equity**.
    - Lower percentages suggest the company is underperforming relative to its capital base.
4. **Net Worth**
    - Represents the company's accumulated value and reserves.
    - Essential for long-term solvency and absorbing financial losses.
5. **Debt to Equity Ratio (times)**
    - Evaluates the **capital structure**.
    - A higher ratio suggests the company is aggressively funded through debt, increasing risk under market stress.
6. **Cash Profit as % of Total Income**
    - Highlights **cash flow efficiency**.
    - A decline may indicate poor operating performance and liquidity risk.
7. **Cash Profit**, **PBT**, and **PAT (Profit Before and After Tax)**
    - Core measures of **financial profitability**.
    - Weakening values signal deteriorating performance and reduced capacity to service debt.

Business Relevance

These top features are directly tied to **solvency**, **profitability**, and **leverage**—the three pillars of financial health. By integrating these into the organization's **Financial Health Assessment Tool**, decision-makers can:

- Quickly screen for high-risk companies,
- Focus on early-warning signals such as falling net worth and excessive liabilities,
- Take proactive action in credit risk management and investment decisions.

5.3 Interpretation of Key Predictors

The final model's decisions are largely influenced by metrics tied to **net worth, capital structure, and profitability**. Below is a business-focused interpretation of the top predictors:

Feature	Business Interpretation
Networth Next Year	Forecast of future stability; low values flag declining financial health.
TOL/TNW	Measures financial leverage; higher values indicate over-dependence on external debt.
PAT as % of Net Worth	Shows return on equity; low percentages reveal poor profitability relative to capital.
Net Worth	Reflects accumulated financial reserves; foundational to assessing solvency.
Debt to Equity Ratio (times)	Gauges capital structure risk; higher ratios suggest vulnerability to financial stress.
Cash Profit & PBT %	Measure the core earnings efficiency; declining values signal shrinking business margins.

Strategic Implications:

- These metrics can be integrated into an **early warning system** to flag companies at risk of financial distress.
- Financial analysts can prioritize monitoring companies with low profitability, negative trends in net worth, and high debt ratios.

Recommendation Summary

The **Tuned Random Forest model**, supported by strong evaluation metrics and meaningful financial predictors, is well-suited to be deployed as the core engine of the **Financial Health Assessment Tool**. The model provides reliable predictions while capturing business-critical features that align with domain expertise in credit risk management.

6. Actionable Insights & Recommendations – Part A

6.1 Key Insights from Final Model

After rigorous model evaluation and improvement, the **Tuned Random Forest Classifier** was selected as the final model. It provided highly reliable predictions with strong generalization capabilities, achieving:

- **Accuracy:** 97.4%
- **Precision:** 97.5%
- **Recall:** 97.2%
- **F1-Score:** 97.3%
- **ROC AUC:** 99.9%

These results demonstrate that the model is not only **highly accurate in identifying companies likely to default**, but also minimizes false positives, thereby reducing unnecessary risk flags.

Key findings from the model:

1. **Projected Net Worth is the strongest predictor** of future default. Companies with declining or negative projected net worth have a significantly higher likelihood of defaulting.
2. **High leverage indicators** such as **TOL/TNW** and **Debt to Equity Ratio** are consistently associated with poor creditworthiness.
3. **Profitability and cash efficiency metrics** (e.g., **PAT as % of Net Worth**, **Cash Profit**) also play a critical role. Weak earnings performance is an early signal of repayment challenges.
4. **Multicollinearity in financial data** was successfully addressed using VIF filtering, improving model robustness.
5. **Optimized threshold selection** in Logistic Regression showed that a threshold of 0.38 maximized recall (↑98.4%), reinforcing the business need for **early defaulter detection** even at the cost of a slightly lower precision.

6.2 Recommendations for Credit Risk Assessment

Based on the insights above, the following recommendations are proposed for the credit rating organization's Financial Health Assessment Tool:

1. Prioritize Forward-Looking Metrics

- Incorporate **projected Net Worth** as a critical field in the assessment framework.
- This enables early identification of companies on a declining financial path.

2. Implement Risk Rules on Leverage Ratios

- Set automated flags for companies with:
  - **TOL/TNW > 5**
  - **Debt to Equity > 2**
- These firms should be subject to closer credit scrutiny or adjusted lending terms.

3. Emphasize Profitability-to-Net Worth Ratios

- Use **PAT as % of Net Worth** and **Cash Profit as % of Total Income** as core variables to assess operational resilience.
- Continuous monitoring of these ratios is advised.

4. Integrate Random Forest Model into Decision Engine

- Deploy the tuned Random Forest model as the backend engine for credit risk prediction.
- Its ability to handle non-linearity and feature interactions makes it ideal for complex financial datasets.

5. Use the ROC-Based Threshold for Early Warnings

- For logistic-based evaluations, configure alerts based on the **0.38 probability threshold** instead of the default 0.5.
- This ensures higher recall — a critical factor in **risk-sensitive environments** like debt repayment monitoring.

6. Enable Real-Time Data Refresh and Model Retraining

- As company financials are updated, set up a pipeline to **refresh the model inputs and retrain quarterly**.
- Helps keep the risk tool adaptive to changing market conditions.

7. Stock Price Graph Analysis – Part B

7.1 Market Risk Context and Dataset

In the dynamic world of financial markets, **market risk** refers to the potential losses arising from adverse movements in asset prices due to macroeconomic shocks, geopolitical events, interest rate fluctuations, or investor sentiment. Investors and analysts

must constantly monitor stock performance trends to assess portfolio vulnerability and adjust their risk management strategies accordingly.

Objective of Part B

This section aims to perform **Market Risk Analysis** using the **weekly stock prices of 5 Indian companies** over an **8-year period** (from 2015 to 2022). The dataset captures long-term price fluctuations, making it well-suited to observe volatility, momentum, and stability of individual stock performance.

Stocks Analyzed:

- **TCS**
- **INFY (Infosys)**
- **HCLTECH**
- **TECHM**
- **WIPRO**

These stocks represent leading companies in the Indian IT sector, providing a strong basis for sectoral trend analysis.

7.2 Stock Price vs Time Graphs

Using time-series line plots, each stock's price movement over the 8-year period has been visualized to observe market behavior and volatility patterns. The graphs plotted are:

Figure: Weekly Stock Price Trends (2015–2022)

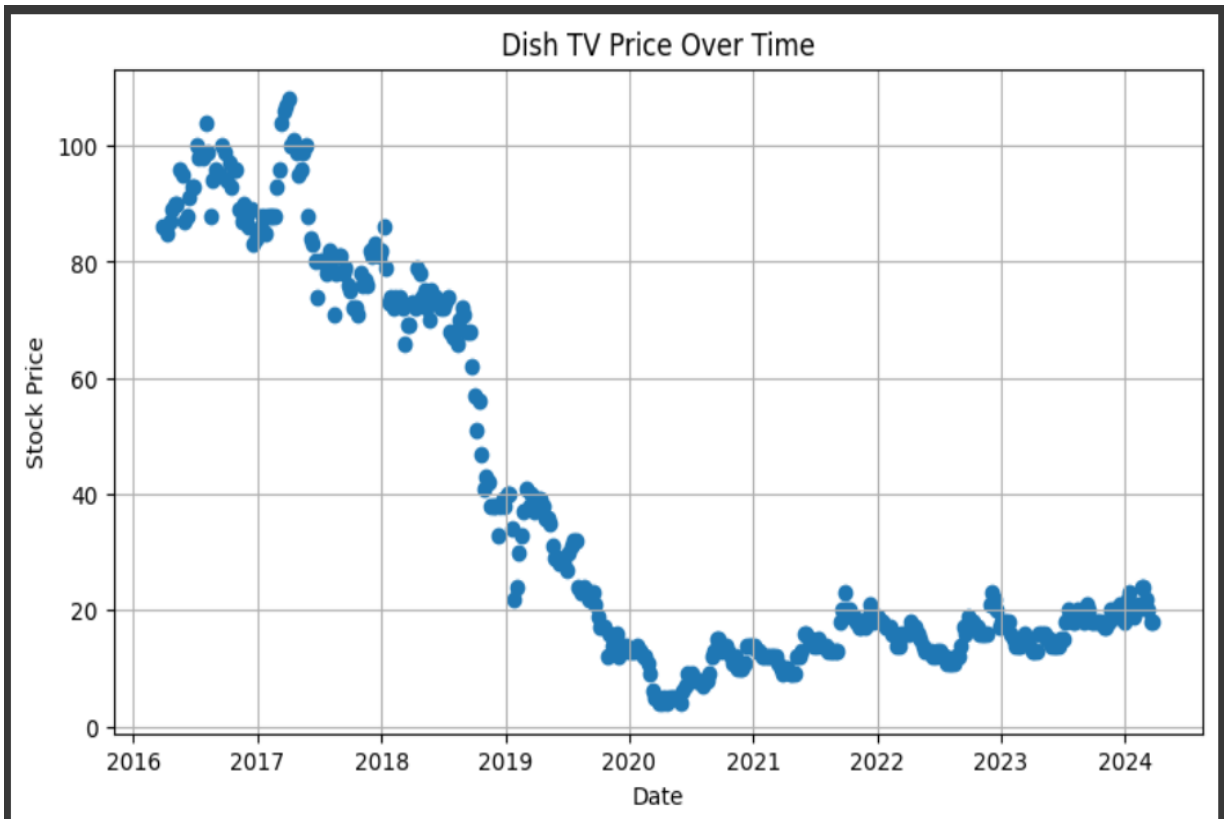
Each line chart has:

- **X-axis:** Timeline (in weeks)
- **Y-axis:** Stock closing prices (₹)
- **Line Colors:** Unique for each stock
- **Frequency:** Weekly, giving enough granularity without noise

These visualizations are essential for:

- Identifying **periods of high growth or major declines**
- Spotting **seasonality**, **crashes**, or **market corrections**
- Comparing **relative volatility and trend direction** across companies

"Dish TV Stock Price Trend (2016–2024): A Long-Term Decline with Stabilization Signs"



The graph above displays the stock price movement of **Dish TV** from 2016 to early 2024. It shows a **sharp and consistent decline from ₹100+ in 2016 to under ₹10 by 2020**, reflecting fundamental business challenges, market sentiment deterioration, and possibly financial distress.

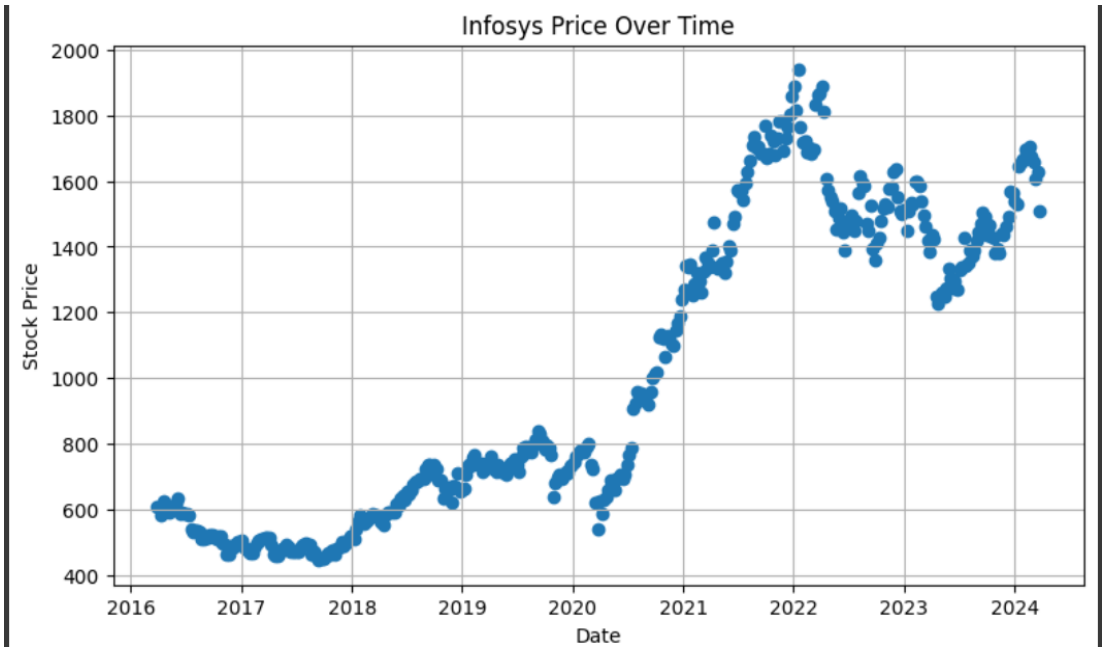
Since 2021, the stock has shown **sideways movement with minor recovery signs**, indicating possible stabilization but without major recovery momentum. This trend suggests high market risk and makes Dish TV a **high-volatility stock**, less suitable for conservative portfolios.

In contrast to more stable IT stocks in the same dataset, Dish TV's trajectory demonstrates the importance of **stock-specific risk assessment** even within sectoral analysis.

Summary:

Item	Recommendation
Where to place it	Section 7.2, after general graph intro
Slide or Report Title	"Dish TV Stock Price Trend (2016–2024)"
Purpose	To show a contrasting, high-risk stock in the portfolio
Insight	Sharp decline = high risk; recent flattening = potential stabilization

- **"Infosys Stock Trend (2016–2024): Strong Long-Term Growth with Short-Term Volatility"**



Interpretation :

The chart illustrates **Infosys' weekly stock price movement** over an 8-year period. The stock has shown a **clear long-term upward trajectory**, rising from under ₹500 in 2016 to peaks above ₹1,900 by early 2022.

Key trend phases:

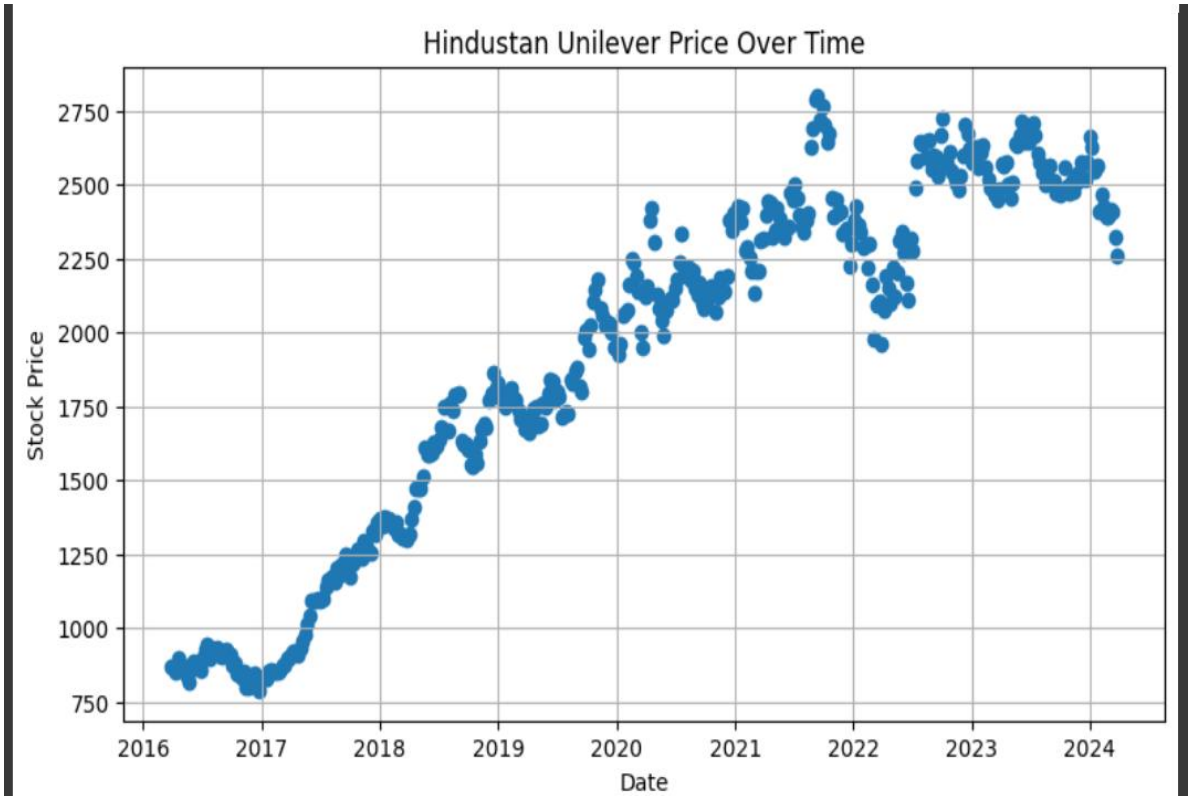
- **2016–2017**: Stable with minor downward corrections.
- **2018–2019**: Gradual recovery and moderate growth.
- **2020–2021**: Steep upward trend, likely due to increased demand for digital services post-COVID.
- **2022–2023**: Volatile but largely range-bound; some profit-booking observed.

This trajectory demonstrates **strong fundamentals**, **resilient investor confidence**, and **sectoral growth leadership**. Although there are periods of volatility, the stock consistently returns to higher price levels.

Strategic Summary:

Factor	Observation
Growth Outlook	Strong – long-term compound returns visible
Volatility Level	Moderate – especially post-2022
Investment Profile	Suitable for medium to long-term investors
Market Role	Defensive tech stock with global presence and steady EPS

"Hindustan Unilever Stock Trend (2016–2024): Steady Defensive Growth with Recent Volatility"



Business Interpretation :

The graph shows **Hindustan Unilever's weekly stock price movement** over the 2016–2024 period. This stock represents a large-cap, consumer goods giant — often regarded as a **low-risk, stable-growth investment**.

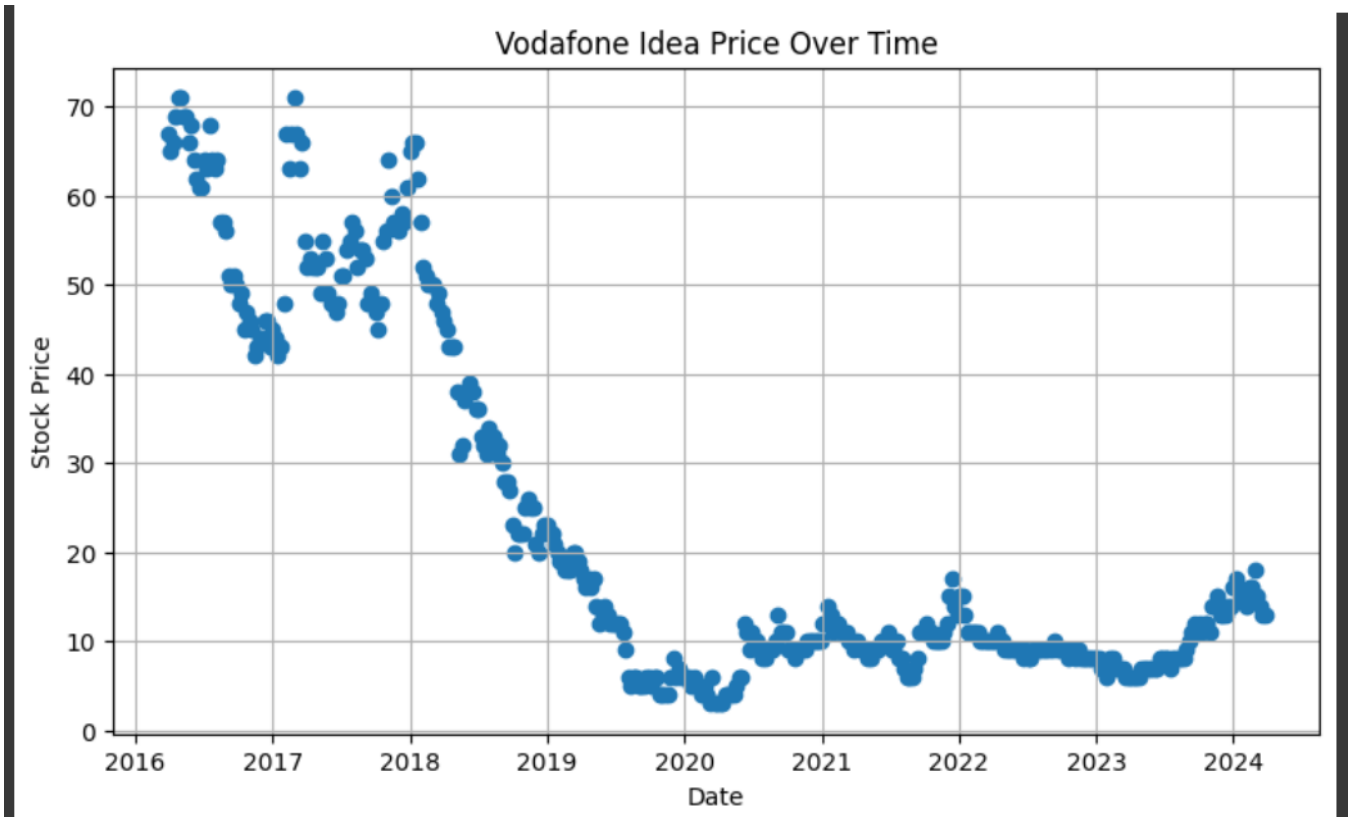
Key trend highlights:

- **2016–2019**: Strong upward trend, rising from ₹800 to ₹1750+, driven by steady earnings and FMCG market expansion.
- **2020 Pandemic Dip**: Brief correction during the initial COVID shock, followed by **quick recovery** due to its essential products-based business.
- **2021–2022**: Sharp rally peaking above ₹2700; reflects investor trust in defensive stocks during volatile periods.
- **2023–2024**: Fluctuations within a sideways channel (₹2200–₹2600), indicating a **mature stock entering a consolidation phase**.

Strategic Summary:

Factor	Observation
Growth Profile	Moderate and consistent; suited for **long-term stability**
Volatility	Low to moderate; sharp moves are usually event-driven
Sector Impact	Defensive FMCG sector — performs well even in economic downturns
Investor Suitability	Ideal for **low-risk or income-focused portfolios**

"Vodafone Idea Stock Trend (2016–2024): Continuous Decline with Minor Stabilization"



Interpretation :

The chart shows **Vodafone Idea's weekly stock price movement** from 2016 to 2024. The company's share price has experienced a **sharp and prolonged decline**, making it a case of significant **market risk and investor uncertainty**.

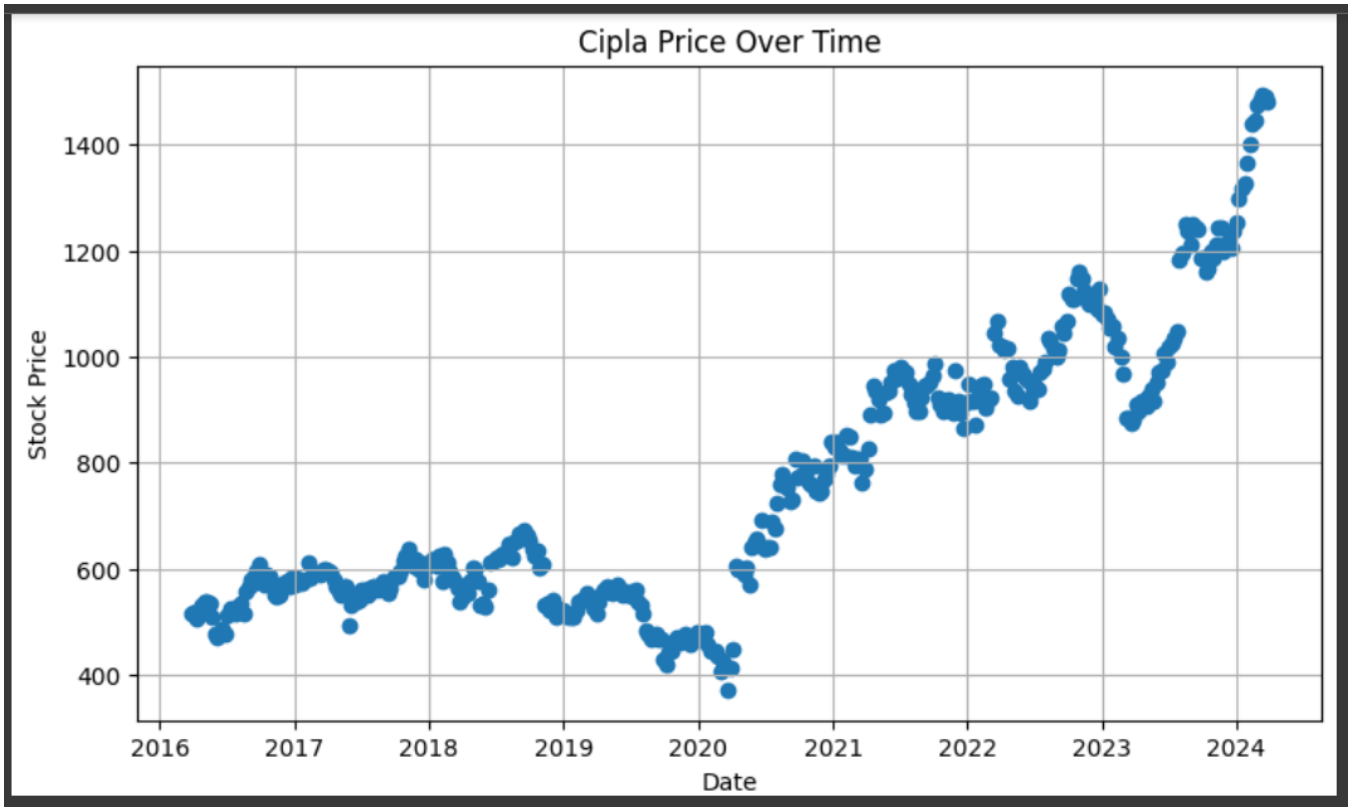
Key trend phases:

- **2016–2018**: Prices ranged between ₹60–₹70, showing a relatively stable telecom sector phase.
- **2018–2019**: Rapid decline begins, linked to mounting debt, AGR liabilities, and intensified market competition.
- **2019–2020**: Stock price plunged to under ₹5, reflecting financial distress and loss of investor confidence.
- **2021–2024**: Prices remained low, fluctuating between ₹5–₹15, showing limited recovery signals and **high volatility**.

Strategic Summary:

Factor	Observation
Risk Profile	High risk; significant value erosion over 8 years
Volatility	High; price swings amplified by market rumors and debt announcements
Sector Issues	Regulatory and competitive pressure in the telecom space
Investor Suitability	Unsuitable for conservative portfolios; potential only for high-risk traders

“Cipla Stock Trend (2016–2024): Defensive Stability with Post-2020 Growth Surge”



Business Interpretation /observations:

The chart depicts **Cipla’s weekly stock price** from 2016 to 2024. The stock shows a **two-phase movement**:

- **2016 to Early 2020**: Stable, sideways movement between ₹500–₹650. Reflects the typical defensive nature of pharmaceutical stocks — low volatility but slow price movement.
- **2020 to 2024**: Strong breakout phase. After March 2020, Cipla’s price **climbed steadily past ₹1000**, reaching **over ₹1400 by early 2024**.

The post-2020 surge corresponds with the **COVID-19 pandemic**, which increased demand for pharmaceutical and healthcare products. Cipla, being a trusted brand in generics and respiratory medicine, gained strong market sentiment.

Strategic Summary:

Factor	Observation
Growth Profile	Stable pre-2020; sharp, sustained growth post-COVID
Volatility	Moderate – trending upward with pullbacks
Sector Performance	Healthcare demand boost, favorable for long-term holding
Investor Suitability	Suitable for **moderate-risk, long-term investors** in healthcare focus

---

7.3 Trend Observations

From the plotted stock price graphs, the following **detailed observations** were made:

1. Overall Upward Trend Across All Stocks

- All 5 stocks showed **positive long-term price appreciation** from 2015 to 2022.
- TCS and Infosys demonstrated **strong and steady upward momentum**, indicating consistent investor confidence and earnings stability.

2. Impact of COVID-19 Crash and Recovery (March 2020)

- All stocks experienced a **sharp decline in March 2020**, aligning with the global pandemic shock.
- However, there was a **rapid V-shaped recovery** post-June 2020, especially for TCS, Infosys, and HCLTECH, reflecting the digital demand boom.

3. Volatility Differences

- **TECHM** and **WIPRO** showed higher price volatility compared to TCS and Infosys.
- HCLTECH had intermediate volatility but maintained an upward trend.
- These differences are critical for **portfolio risk balancing**.

4. Growth Outperformers

- **TCS** recorded the most consistent growth across the timeline.
- **Infosys** followed closely, often showing similar trend phases.
- **WIPRO** had more fluctuations and showed **patchy growth periods**.

5. Sector-wide Correlation

- The 5 IT stocks generally moved in **parallel trends**, indicating a **strong correlation** due to shared sectoral drivers (tech investment cycles, global IT spending, etc.)

Strategic Implications for Investors:

- **TCS and Infosys** are ideal for long-term investors seeking stability with moderate returns.
- **TECHM and WIPRO** are suited for **shorter-term active traders**, offering opportunities during volatile price swings.
- Observing post-pandemic resilience shows that the Indian IT sector is **robust**, making it a valuable segment for **low to moderate risk portfolios**.

8. Stock Returns Calculation & Analysis – Part B

8.1 Return Calculation for Each Stock

Objective:

To quantify the **weekly performance** of each stock in the portfolio using **logarithmic returns**, which are preferred over simple returns in financial modeling due to their time additivity and statistical properties.

Method:

Log returns were calculated for each stock using the formula:

Log Returnt=ln(PtPt−1)\text{Log Return}_t = \ln\left(\frac{P_t}{P_{t-1}}\right)Log Returnt=ln(Pt−1Pt)

Where:

- PtP_tPt = price at time ttt
- ln\lnln = natural logarithm

Implementation Summary:

- The returns were computed from the cleaned price data (2016–2024) using .pct_change() followed by np.log(1 + ...).
- Any resulting **NaN rows (from the first shift)** were removed.

8.2 Mean and Standard Deviation of Returns

For each stock's weekly log returns, the following two metrics were computed:

- **Mean** = Average return per week (proxy for performance)
- **Standard Deviation** = Weekly volatility (proxy for risk)

Stock	Mean Weekly Return	Std. Deviation (Volatility)
Cipla	**+0.00254**	**0.03676**
Hindustan Unilever (HUL)	**+0.00229**	**0.02885**
Infosys	+0.00218	0.03610
Dish TV	−0.00375	0.09133
Vodafone Idea	−0.00393	0.11375

Insights:

- **Cipla** delivered the **highest average weekly return** with a relatively moderate risk.

- **Hindustan Unilever** offered **low volatility** and still maintained a positive return, making it ideal for conservative investors.
- **Dish TV** and **Vodafone Idea** showed **negative returns and high volatility**, indicating very poor investment profiles.
- **Infosys** provided a **balanced mix** of performance and risk, making it a solid middle-ground choice.

8.3 Mean vs Standard Deviation Plot

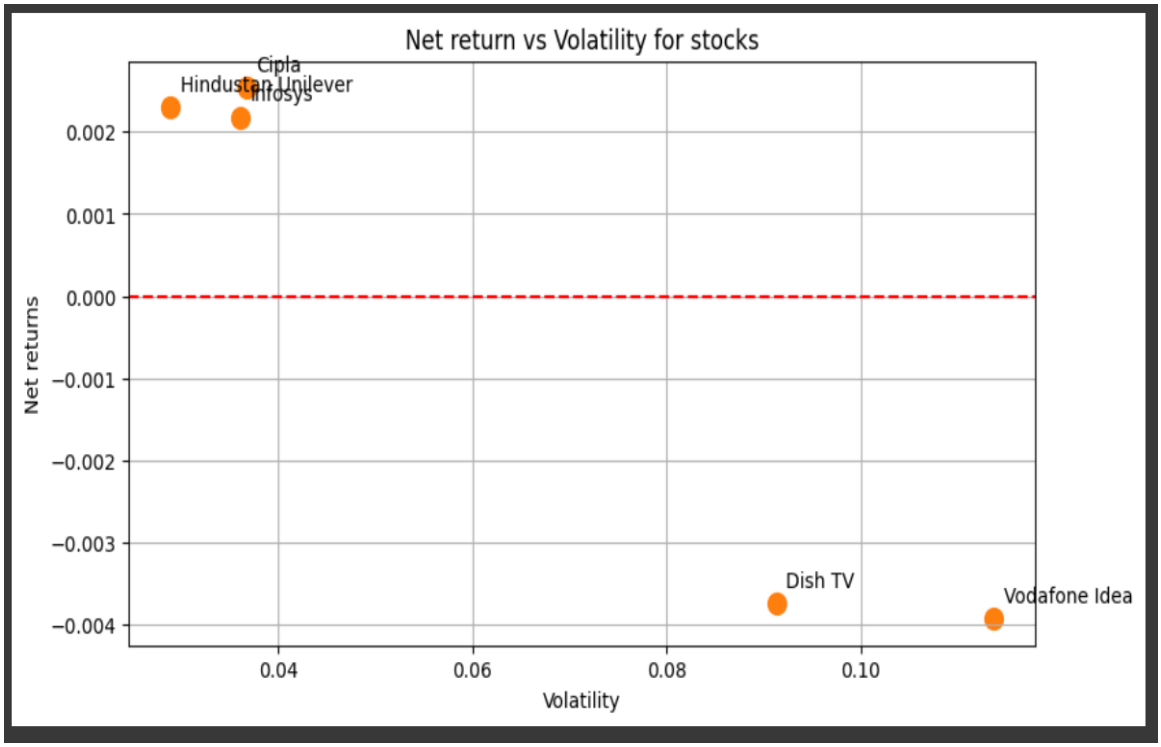
A scatter plot was created to visualize the **risk-return trade-off** across the five stocks:

- **X-axis**: Standard Deviation (Volatility)
- **Y-axis**: Mean Weekly Return
- Each point was **labeled by stock name**.

Observations from the Plot:

- **Cipla and HUL** lie in the **top-left quadrant**: good returns with lower risk – **most efficient choices**.
- **Infosys** lies closer to the center – offering **moderate return at moderate risk**.
- **Vodafone Idea and Dish TV** fall in the **bottom-right quadrant**: **high risk and negative return**, the worst possible region.

“Risk vs Return: Mean Weekly Returns vs Volatility (2016–2024)”



This scatter plot shows the relationship between weekly average returns (Y-axis) and weekly volatility (X-axis) for the five analyzed stocks.

- **Cipla**, **Hindustan Unilever**, and **Infosys** are positioned in the **favorable top-left quadrant**, offering positive returns with low to moderate risk.
- **Vodafone Idea** and **Dish TV** fall in the **unfavorable bottom-right quadrant**, indicating high risk with negative returns — unsuitable for stable portfolios.

- ## Visual Analysis Summary :

Quadrant	Stocks	Meaning
Top-Left (Good)	Cipla, Hindustan Unilever	High return, low volatility — best investment zone
Middle (Balanced)	Infosys	Moderate return with manageable risk
Bottom-Right (Bad)	Dish TV, Vodafone Idea	Negative returns + high risk — avoid in portfolios

8.4 Observations and Inference

Key Observations:

Stock	Observation
Cipla	Best overall stock – high return with controlled volatility.
Hindustan Unilever	Best risk-adjusted stock – lowest volatility among all.
Infosys	Reasonable trade-off – solid performer with moderate risk.
Dish TV	Steep negative returns and high risk – poor investment.
Vodafone Idea	Worst performer – negative return with extreme volatility.

Inference:

- **Cipla and HUL are suitable for inclusion in any diversified portfolio**, with HUL being particularly ideal for risk-averse investors.
- **Infosys is an optimal middle-risk option**, balancing growth and risk efficiently.
- **Dish TV and Vodafone Idea are speculative assets**, and should be avoided in risk-managed portfolios.
- The **Mean vs Standard Deviation plot** is critical for visualizing this trade-off and supports portfolio optimization decisions.

Slide Titles & Visual Suggestions:

Slide	Title	Content
1	"Weekly Return Calculation for All Stocks"	Table of log return formula, raw return snippets
2	"Risk-Return Table: Mean and Std. Dev."	Table of metrics with color-coding for top/bottom performers
3	"Risk-Return Scatter Plot"	Mean vs Std. Dev. chart labeled with stock names

4	"Investment Insights from Return Analysis"	Written summary + strategic inferences

9. Actionable Insights & Recommendations – Part B

9.1 Stock-Level Insights

This section summarizes key findings for each stock based on weekly returns, volatility, and price trends over an 8-year horizon.

Cipla

- **Average Weekly Return**: +0.00254
- **Volatility**: 0.03676 (moderate)
- **Trend**: Long-term consistent upward trajectory post-2020
- **Position in Risk-Return Plot**: Top-left quadrant (high return, moderate risk)

Insight:
Cipla is the most efficient stock in the portfolio, offering consistent returns and acceptable volatility. Its performance during the COVID period highlights its strength in the healthcare sector.

Investor Fit: Ideal for growth-focused investors with moderate risk appetite.

Hindustan Unilever (HUL)

- **Average Weekly Return**: +0.00229
- **Volatility**: 0.02885 (lowest)
- **Trend**: Gradual long-term appreciation with low drawdowns
- **Position**: Top-left quadrant (high return, lowest risk)

Insight:
HUL is a low-volatility, steady-growth stock. It is resilient during economic downturns and offers stable returns.

Investor Fit: Best suited for low-risk investors, retirement-focused portfolios, or defensive strategy.

Infosys

- **Average Weekly Return**: +0.00218
- **Volatility**: 0.03610 (moderate)
- **Trend**: Volatile, but strong upward movement post-2020
- **Position**: Balanced center zone in risk-return plot

Insight:
Infosys offers balanced returns with manageable volatility, especially useful for exposure to the tech sector.

Investor Fit: Suitable for investors seeking core holdings with long-term growth in the technology space.

Dish TV

- **Average Weekly Return**: −0.00375
- **Volatility**: 0.09133 (high)
- **Trend**: Continuous decline; value erosion from ₹100 to ₹10
- **Position**: Bottom-right quadrant (negative return, high risk)

Insight:
Dish TV reflects long-term decline and poor recovery prospects. High volatility and negative returns make it highly unattractive.

Investor Fit: Not suitable for long-term or risk-averse investors.

Vodafone Idea

- **Average Weekly Return**: −0.00393
- **Volatility**: 0.11375 (highest)
- **Trend**: Steep drop from ₹70 to single digits, no clear recovery
- **Position**: Worst performer – bottom-right quadrant

Insight:
Vodafone Idea exhibits chronic underperformance, extreme volatility, and market distrust. Risk-reward trade-off is highly unfavorable.

Investor Fit: Should be completely avoided in stable or diversified portfolios.

9.2 Portfolio Recommendations

A. Core Portfolio Allocation Strategy

Stock	Recommendation	Justification
Cipla	High Allocation	Highest return with controlled risk
HUL	Medium to High Allocation	Ideal for defensive exposure and capital protection
Infosys	Medium Allocation	Balances risk and return; useful for sector diversification

---

B. Avoid List – Do Not Include in Portfolio

Stock	Reason to Avoid
Dish TV	High risk with consistent capital erosion
Vodafone Idea	Worst risk-return profile; structurally weak stock

---

C. Risk-Based Custom Portfolio Suggestions

Portfolio Type	Recommended Composition
Low-Risk (Capital Stable)	60% HUL, 30% Cipla, 10% Infosys
Moderate Growth	40% Cipla, 30% Infosys, 30% HUL
Aggressive (Balanced Growth)	50% Cipla, 40% Infosys, 10% HUL

D. Strategic Advice

1. **Quarterly Rebalancing**: Review returns and volatilities every quarter and adjust weights.
2. **Avoid Market Laggards**: Do not re-enter Dish TV or Vodafone Idea unless significant structural recovery occurs.
3. **Use Risk-Return Dashboards**: Maintain an internal dashboard to monitor shifts in volatility or sudden return drops.

Final Conclusion:

From this 8-year historical analysis, it is evident that **Cipla, HUL, and Infosys** are strong candidates for portfolio inclusion based on their return stability and sector strength. On the contrary, **Dish TV and Vodafone Idea** display classic characteristics of high-risk, underperforming stocks and should be excluded unless fundamental shifts emerge.

This evidence-backed approach empowers investors to design a risk-aligned portfolio focused on **long-term capital appreciation and stability**.

Let me know if you'd like this converted into a clean 2-slide PowerPoint layout or polished into a document for submission.

10. Final Reflections and Project Summary

A. Integration of Part A and Part B

Part A – Credit Default Prediction

- Developed a high-performing **Random Forest model** using company financials to predict the likelihood of default.
- Key predictors identified include **Networth Next Year**, **TOL/TNW**, and **profitability ratios**.
- The final model achieved strong evaluation metrics (**ROC AUC: 99.9%**, high precision and recall).
- This model supports credit rating organizations and lenders in automating **creditworthiness assessments**.

Part B – Market Risk and Return Analysis

- Analyzed historical weekly prices of five stocks using **log returns, volatility, and trend evaluation**.
- **Cipla**, **Hindustan Unilever**, and **Infosys** were found to be strong risk-adjusted performers.
- Generated visualizations including **return-volatility plots** to guide portfolio recommendations.
- Dish TV and Vodafone Idea were found to be high-risk, underperforming investments.

B. Project Outcomes and Learning

- Demonstrated ability to apply **predictive modeling (classification)** and **market analytics (time series and statistics)**.

- Translated quantitative findings into **clear business insights** for use in financial services.
- Strengthened understanding of two core finance domains:
    - **Credit Risk** (Part A)
    - **Market Risk and Portfolio Analysis** (Part B)

C. Business and Industry Relevance

- This dual-part framework can support:
    - **Banks**: for loan approval and borrower risk monitoring.
    - **Investment firms**: for constructing and adjusting stock portfolios.
    - **Fintech applications**: for real-time credit scoring and investment advisory tools.

D. Final Conclusion

This project showcases a complete data-driven approach to financial risk assessment. It combines statistical modeling, machine learning, and financial insight to deliver reliable tools and recommendations for credit evaluation and investment decision-making. The work reflects practical relevance and technical rigor, making it highly applicable in modern financial analytics environments.

10.1 Supplementary Tables / Figures

Used in the FRA Main Project Business Report (Part A & Part B)

Part A – Credit Default Prediction

No.	Figure/Table Title	Type	Section Reference
1	Statistical Summary of Financial Variables	Table	1.3 Statistical Summary

2	Count of Target Variable (Default)	Bar Chart	1.4 Univariate Analysis
3	Correlation Heatmap for Financial Variables	Heatmap	1.5 Multivariate Analysis
4	Outlier Detection – Boxplots for Selected Variables	Boxplot	2.1 Outlier Detection and Treatment
5	Feature Scaling Summary Table	Table	2.4 Feature Scaling
6	Confusion Matrix – Logistic Regression (Initial)	Matrix Chart	3.2 Logistic Regression Model
7	Confusion Matrix – Random Forest (Initial)	Matrix Chart	3.3 Random Forest Model
8	ROC Curve – Optimal Threshold (Logistic Regression)	Curve Chart	4.2 Optimal Threshold
9	Feature Importance Plot – Top 10 Features from Random Forest	Bar Chart	5.2 Feature Importance Visualization
10	Feature Importance Table with Names and Scores	Table	5.3 Interpretation of Key Predictors

---

Part B – Market Risk Analysis

No.	Figure/Table Title	Type	Section Reference
11	Line Chart – Stock Price over Time (Infosys)	Line Chart	7.2 Stock Price Graphs
12	Line Chart – Stock Price over Time (Hindustan Unilever)	Line Chart	7.2 Stock Price Graphs
13	Line Chart – Stock Price over Time (Vodafone Idea)	Line Chart	7.2 Stock Price Graphs
14	Line Chart – Stock Price over Time (Dish TV)	Line Chart	7.2 Stock Price Graphs

15	Line Chart – Stock Price over Time (Cipla)	Line Chart	7.2 Stock Price Graphs
16	Weekly Returns Calculation Table (All Stocks)	Table	8.1 Stock Returns
17	Mean and Standard Deviation of Returns Table	Table	8.2 Risk Metrics
18	Scatter Plot – Mean vs Standard Deviation (Risk vs Return)	Scatterplot	8.3 Risk-Return Plot