# Indian Institute of Technology Roorkee


## Project Report

## Econometric, Machine Learning and Deep Learning

## Modelling of CMIE Financial Data

Submitted to - Dr. Gaurav Dixit

(Joint Faculty MFS, Core Faculty DOMS, IIT Roorkee)


Submitted By - Rohit Rawat

(B.S. Economic Science, IISER Bhopal)

# Declaration

I hereby declare that the project titled ‘Econometric, Machine Learning and Deep Learning Modelling of CMIE Data’ submitted to Mehta Family School of Data Science and Artificial Intelligence, Indian Institute of Technology Roorkee is an authentic and original record of my own efforts, under the guidance of Dr. Gaurav Dixit for the fulfilment of SPARK Summer Research Internship. I also declare that the project has not been submitted to any other institution in any form.

Date: Rohit Rawat

Place:

# Table of Content

# Abstract

Investment refers to an asset acquired with the objective of generating income or achieving an appreciation in value over time. When a firm, whether public or private, makes an investment, it is theoretically subject to the market's supply and demand dynamics. This means that investments in certain assets can be influenced by uncontrollable events. Consequently, investments in Information Technology (IT) and Information Systems (IS) are also susceptible to these external influences.

This study aims to explore the impact of government policies and other significant events on IT/IS investments in Indian companies. The analysis is based on data from 4009 companies spanning the period from 2010 to 2023. To understand these influences, a trend analysis of different metrics of the companies was performed and then compared to the potential external influences. Various analytical models were employed, including econometric models, machine learning decision tree models, and a deep learning model based on recurrent neural networks (RNN) known as Long Short-Term Memory (LSTM).

Keywords: Finance, IT/IS, Investment, Financial Modelling, CMIE.

# Introduction

Information Technology (IT) and Information Systems (IS) have profoundly influenced the operations of both public and private organizations. IT/IS investments generally encompass computer hardware (e.g., personal computers, servers, mainframes, and telecommunications equipment), software (system, general-purpose, custom, and utility software), and personnel (such as system administrators, maintenance staff, and technical experts). In 2021, IT investments in the United States reached USD 1.065 trillion, followed by China at USD 333 billion, Japan at USD 161 billion, and India at USD 72 billion. Global spending on IT hardware and communication services amounted to USD 4.7 trillion.

India, as a developing country, presents significant opportunities for growth in IT/IS investments. The Indian government has launched several initiatives, including Digital India, UPI, and Demonetisation, which have directly or indirectly influenced IT/IS spending. With its lower labour costs and expanding digital infrastructure, India is well-positioned to leverage technology to enhance productivity and economic growth.

In our study, we employed advanced econometric models, as well as machine learning (ML) and deep learning (DL) techniques, to analyse the impact of government policies, schemes, and other external factors on IT/IS investments in India. We also performed trend analysis to identify patterns and forecast future developments in IT/IS spending. Our findings aim to shed light on how these factors shape the investment landscape and provide insights into the key drivers of IT/IS adoption in India’s rapidly evolving market.

# Research Framework

This report presents a comprehensive analysis of panel data sourced from the Centre for Monitoring Indian Economy (CMIE). The dataset, which spans a broad timeframe, captures the nuances of economic cycles and fluctuations.

Statistical summaries and visualizations were utilized to identify patterns and anomalies over time. This exploration established a solid foundation for further analysis.

Econometric techniques such as the Fixed Effect Model, Random Effect Model, Fixed Difference OLS Model, and Generalized Method of Moments (GMM) were employed to untangle complex relationships between different factors, offering insights into their interdependencies and predictive power.

In addition, machine learning algorithms, including Gradient Boosting Machines, XGBoost, ADABoost, and Random Forest, were applied to forecast based on historical patterns. These methods provided predictive models with enhanced accuracy for anticipating future economic trends.

Furthermore, deep learning techniques, specifically Long Short-Term Memory (LSTM) networks and Gated Recurrent Unit (GRU) networks, were used to capture intricate temporal patterns in the data. LSTM models, known for their ability to learn from sequential data, proved ideal for forecasting economic indicators over extended timeframes.

# Data Preparation

Data from CMIE was acquired through ProwessIQ, an application for querying CMIE's database. A period of 2010-2023, since it encompasses a lot of newer government schemes and policies in practice and certain 'events' like COVID 19.

69 feature variables were handpicked, all related to company's metrics and investments in tangible and intangible assets. Other than that, 16 control variables were picked to categories the companies under different banners.

Feature variable observation was a float value in Million Rs. Unit.

Finally, data of 12175 Indian Companies were queried in ProwessIQ, taking ~15-20 variables in a batch, to avoid performance issues. The batches of data were stored in 4 different excel sheets.

The 4 batches were merged using inner join on the Company Name, giving a total of 12175*14*(69+16) = 14488250 observations.

The resulting data consisted of 12177 rows and 1191 columns.

Finally, the data was converted into Panel Data form using Python Pandas Library, resulting in 170451 rows and 87 columns.

The resulting panel data was then pre-processed using Python's Pandas library, to fill out missing value and remove irrelevant variables and companies.

The first step involved removal of irrelevant variables or variables with high levels of missing value. Threshold was set at 50%. If a given variable had more than 50% of observations as null, remove it. This reduced the feature variables from 69 to 19.

Continuing from the partially filtered panel data, the second step involved removal of companies with high levels of missing value. Threshold was set at 20%. If a given company had more than 20% of observations as null, remove it. This reduced the companies from 12175 to 6159 companies.

Finally, companies with low numbers of addition to computers and IT systems per year were removed. The threshold was set at Rs. 1 million averages for the year period. Companies with lower value were removed thus giving the final number of companies from 6159 to 4009.

The final data was then subject to pandas interpolation using the linear method, to fill the missing values.

# Methodology

## 1) Random Effect Model

A random effects model is a statistical method used in econometrics and social sciences to analyse panel data, where observations are made on multiple entities over time. Unlike fixed effect models that assume each entity has a specific, unchanging effect, random effects models treat these entity-specific effects as random variables that follow certain distributions, typically assumed to be normally distributed with mean zero.

The entity-specific effects are not observed but are estimated from the data. The model allows for both within-group and between-group variations by incorporating these random effects alongside the explanatory variables of interest. This approach accounts for heterogeneity across entities and captures the average effect of the explanatory variables on the outcome variable across all entities.

## 2) Fixed Effect Model

A fixed effect model is a statistical method used primarily in econometrics and social sciences to analyse panel data, where observations are made on multiple entities over time.

Unlike random effects models, fixed effect models assume that each entity (such as individuals, firms, or countries) has a specific, unchanging effect that influences the observed outcomes. These fixed effects are typically represented as dummy variables for each entity in the regression analysis. By controlling for these fixed effects, the model can account for individual heterogeneity and isolate the effects of other

explanatory variables of interest. Fixed effect models are valuable in identifying causal relationships within panel data by focusing on within-group variations over time while controlling for stable individual characteristics.

## 3) Fixed Difference OLS

Fixed difference OLS is a method used to control for both time-invariant individual characteristics (fixed effects) and time-varying shocks or changes (first differencing). First Differencing (FD): This technique involves differencing the data over time for each entity. It subtracts each entity's value in one period from its value in the previous period. First differencing helps in removing time-invariant individual effects and focuses on changes within entities over time.

Combining these two methods, fixed difference OLS estimates the coefficients by first differencing the data to eliminate fixed effects and then running Ordinary Least Squares (OLS) regression on the differenced data. This approach allows researchers to control for both time-invariant individual characteristics and time-varying shocks or changes, thereby isolating the effects of variables of interest more accurately.

## 4) Generalized Method of Moments

This estimation technique, known as the Generalized Method of Moments (GMM), is extensively employed in econometrics and statistics to mitigate issues such as endogeneity in regression analysis.

The fundamental idea behind the GMM estimator is to minimize a criterion function by selecting parameters that align the sample moments of the data as closely as feasible

with the population moments. This approach allows researchers to account for various sources of bias and inconsistency in their regression models, thereby improving the reliability and accuracy of their statistical analyses.

## 5) Gradient Boosting

Gradient Boosting is a powerful machine learning technique that builds predictive models in a sequential manner, leveraging the strengths of decision trees. It works by iteratively adding new models to correct errors made by existing models. Each new model is trained to predict the residual errors of the ensemble of models built so far.

The final prediction is obtained by summing the predictions of all the models. Gradient Boosting is particularly effective in handling complex datasets and often outperforms other ensemble methods due to its ability to capture intricate relationships between features and target variables.

## 6) XGBoost

XGBoost, short for Extreme Gradient Boosting, is an optimized and scalable implementation of the gradient boosting framework. It was developed to provide high performance and efficiency in machine learning tasks, especially in supervised learning problems such as regression and classification.

XGBoost enhances traditional gradient boosting by integrating several enhancements, including regularization techniques to prevent overfitting, parallel processing to

improve computational speed, and advanced tree pruning methods for better model generalization.

## 7) AdaBoost

AdaBoost Regression is a powerful ensemble learning technique that combines multiple weak learners, typically decision trees, to create a robust predictive model. Unlike traditional regression methods that fit a single model to the data, AdaBoost sequentially trains a series of models, each focusing on instances where previous models performed poorly.

It assigns higher weights to misclassified data points in each iteration, thereby ensuring subsequent models prioritize learning from these errors.

## 8) Random Forest

Random forest regression is a powerful machine learning technique that combines the concepts of ensemble learning and decision trees to predict continuous values. Unlike traditional regression methods that rely on a single model, random forest regression constructs multiple decision trees during training.

Each tree is trained on a subset of the data and makes predictions independently. The final prediction is then determined by averaging the predictions of all the individual trees, resulting in a robust and accurate model that can handle complex datasets with nonlinear relationships between variables.

## 9) Long Short-Term Memory

LST, or Long Short-Term Memory, is a type of recurrent neural network architecture designed to overcome the limitations of traditional RNNs in capturing long-term dependencies in sequential data.

LSTMs utilize a system of gates to regulate the flow of information through the network, consisting of input, forget, and output gates. These gates enable LSTMs to selectively remember or forget information from previous time steps.

## 10) Gated Recurrent Unit Network

GRUs are designed to handle sequential data and address the vanishing gradient problem by selectively updating and resetting their internal state.

GRUs have fewer parameters compared to LSTMs, as they merge the forget and input gates into a single update gate, simplifying the architecture without sacrificing performance.

# Target and Feature Variables

Panel Data used consists data of 4009 companies over a span of 2010-2023. This dataset includes the following columns:

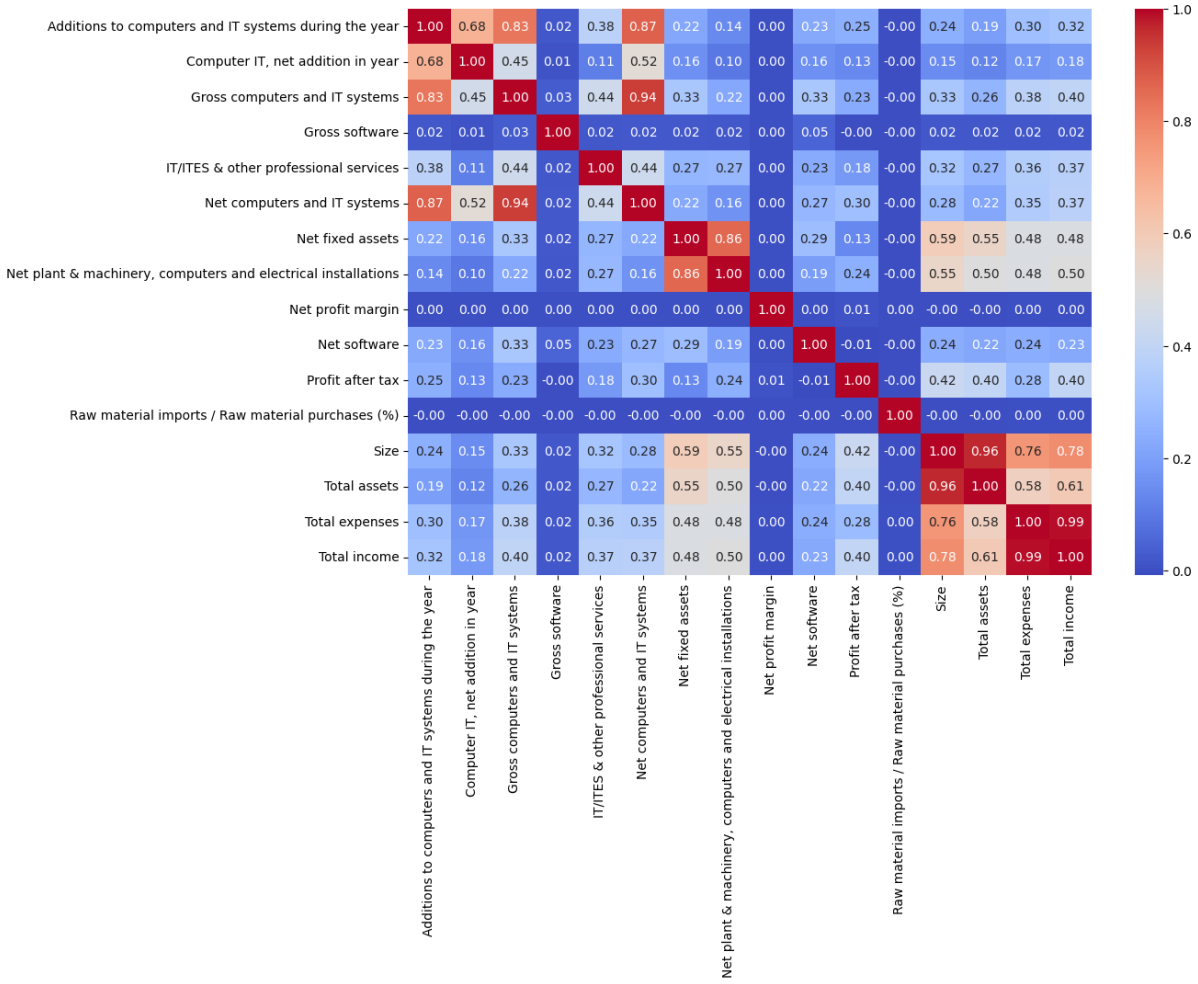
1) Additions to computers and IT systems during the year- This data field stores the value of all additions to the assets of computers made by the company during an accounting.
2) Computer IT, Net Addition in year- Like Additions to computers and IT systems during the year.
3) Gross computers and IT systems- This data field stores the gross value of computers and its peripherals owned by the company or leased by it during
   an accounting period.
4) Gross software- This data field stores the gross value of software of a company on the last day of the accounting period.
5) IT/ITES & other professional services- This data field is a child indicator under the parent 'Outsourced professional jobs', which captures expenses incurred by a company on services rendered by IT/ITES and other professional service providers.
6) Net computers and IT systems- This data field stores the net value of computers and its peripherals owned by the company or leased by it during an accounting period.
7) Net fixed assets- Net fixed assets is the net value of the fixed assets of a company after adjusting for additions/(deductions) to gross fixed assets and the cumulative depreciation on gross fixed assets.
8) Net plant & machinery, computers and electrical installations- The data field stores the net value of plant and machinery, computers and its peripherals and electrical installations, equipment and fittings as at the end of the accounting period

9) Net profit margin- A company's net profit margin ideally tells us how much after-tax profit the business makes for every rupee it generates as income.
10) Net software- This data field stores the net value of the software assets of the company at the end of the accounting period.
11) Profit after tax- This is the net profit of the company after tax. It is the residual after all revenue expenses are deducted from the sum of the total income and the change in stocks.
12) Size- The indicator 'Size' available in the 'Annual Financial Statements' query trigger provides the size of a company. Size is defined as the three-year average of the total income and total assets of a company.
13) Total assets- Total assets refer to sum of all current and non-current assets held by a company as on the last day of an accounting period.
14) Total expenses- This data field stores the sum of all revenue expenses incurred by a company during an accounting period.
15) Total income- Total income is the sum of all kinds of income generated by an enterprise during an accounting period. It includes income from continuing operations as well as income from discontinuing operations.

Total income was set as the target variable, as total income of a given company is a good indicator of growth and if the investments were fruitful. Remaining variables were feature variables.

Each observation had a value in Unit Million Rs.

# Correlation Matrix of all the Variables




Variables like (Additions to computers and IT systems during the years, Computer IT, Net Addition in year) have high correlations while other similar variables such as (Gross computers and IT systems, Net computers and IT systems), (Additions to computers and IT systems during the years, Gross computers and IT systems), (Size, Total Assets), (Size, Total Expenses) and more.

The results from the correlation matrix suggested presence of multicollinearity between the feature variables.

# VIF Test

| Feature | VIF |
|---|---|
| Additions to computers and IT systems during the years | 6.34 |
| Computer IT, Net Addition in year | 2.15 |
| Gross computers and IT systems | 10.57 |
| Gross software | 1.03 |
| IT/ITES & other professional services | 1.44 |
| Net computers and IT systems | 12.255 |
| Net fixed assets | 5.89 |
| Net plant & machinery, computers and electrical installations | 4.52 |
| Net profit margin | 1.00 |
| Net software | 1.25 |
| Profit after tax | 1.53 |
| Size | 115.54 |
| Total assets | 70.08 |
| Total expenses | 12.10 |

Results from the VIF Test show that multicollinearity exists among several variables.

After some feature engineering, VIF Test among the new selection of feature variables showed low multicollinearity.

| Feature | VIF |
|---|---|
| Computer IT, Net Addition in year | 1.37 |
| Net computers and IT systems | 1.67 |
| Net plant & machinery, computers and electrical installations | 1.35 |
| Net software | 1.15 |
| Profit after tax | 1.19 |
| Total expenses | 1.52 |

# Trend Analysis

Trend Analysis of the data was performed by dividing the companies into different sectors. The 4009 were categorised into the following sectors:

1) FMCG
2) BFSI
3) IT
4) Telecommunication
5) Media and Allied
6) Minerals and Allied
7) Other household goods
8) Electricity
9) Logistics and Transportation
10) Education
11) E-Commerce
12) Healthcare

A sudden increase or decrease certain variables in could be explained by following government schemes or policies and events:

1) Auction of 3G Spectrum Licenses- 2010
2) Bharat Net- 2011
3) 4g Commercially Available- 2012
4) Pradhan Mantri Jan Dhan Yojana -2014
5) Digital India- 2015
6) JIO Launch- 2016
7) UPI- 2016
8) Demonetisation- 2016

9) Startup India- 2016
10) GST- 2017
11) NDCP- 2018
12) COVID19 Digital Transformation- 2020
13) Everything Bubble-2021
14) Recession- 2022

Majority of the companies with high IT investment fell in the brackets of BFSI, IT industry and Telecommunication.

In the BFSI sector, significant spikes in feature variables were observed in 2021, primarily attributed to the rapid digitization of infrastructures necessitated by the social distancing measures during the COVID-19 pandemic. This sudden shift to digital platforms for financial services resulted in a notable increase in the adoption and utilization of digital tools and technologies.

Similarly, the IT industry experienced substantial spikes in many of its features in 2021, a trend largely driven by the COVID-19 pandemic. However, these spikes can also be traced back to earlier technological advancements such as the advent of 3G and 4G technologies, the implementation of the National Digital Communications Policy (NDCP), and initiatives like Digital India. These factors collectively contributed to the widespread digital transformation within the IT sector.

The telecommunications sector was notably influenced by the introduction of 3G and 4G technologies, which revolutionized mobile and internet services. This period also saw the emergence of JIO, which played a pivotal role in making affordable data services accessible to

a broader audience. Additionally, policies such as NDCP and initiatives like Digital India further accelerated the growth and development of the telecommunications infrastructure.

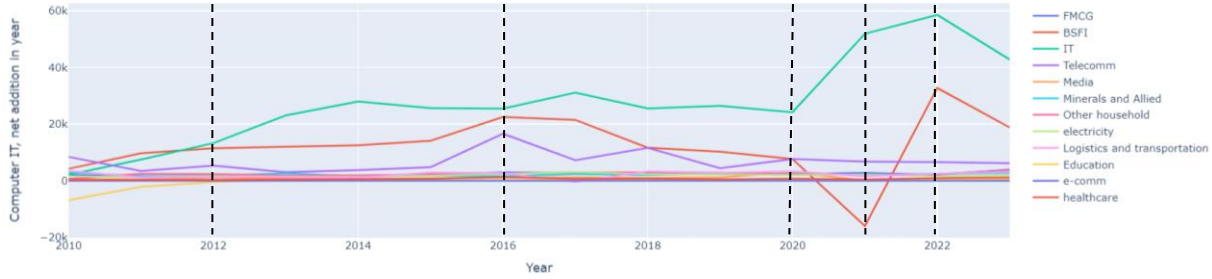
In contrast, the media and allied industries showed less dramatic spikes. The changes observed in this sector can mostly be explained by the digital transformation driven by COVID-19, alongside the broader impacts of Digital India, NDCP, and the rollout of 3G and 4G technology. These factors collectively facilitated the shift towards digital media consumption and production.

The education sector also showed notable spikes, closely linked to the emergence of 3G and 4G technologies, which enabled better access to online education platforms. Initiatives like Digital India and the necessity for digital transformation during COVID-19 further propelled this sector towards increased digital adoption and utilization.

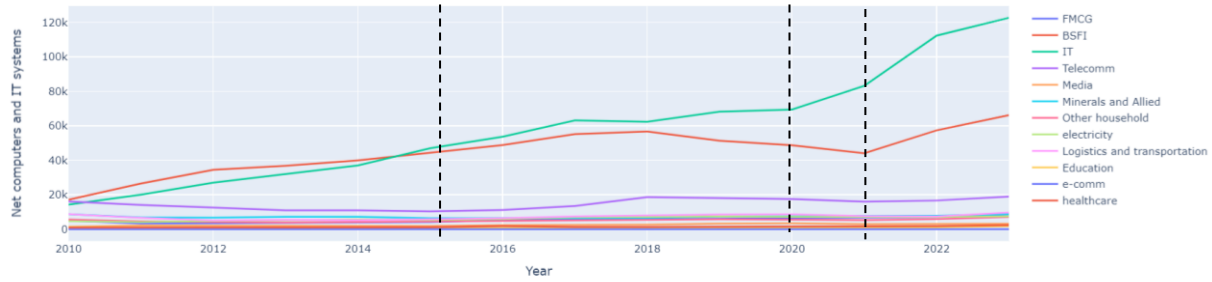
The FMCG and household goods sector exhibited fewer spikes compared to other sectors. The observed changes can be attributed to the implementation of the Unified Payments Interface (UPI), which revolutionized payment methods, the effects of demonetization which pushed for a cashless economy, and the digital transformation driven by COVID-19. These factors collectively influenced consumer behaviour and the operational dynamics within this sector.

Overall, while each sector experienced unique impacts, the common thread across these industries was the significant role of digital transformation, driven by both technological advancements and the imperative adaptations necessitated by the COVID-19 pandemic.

## Computer IT, net addition in year

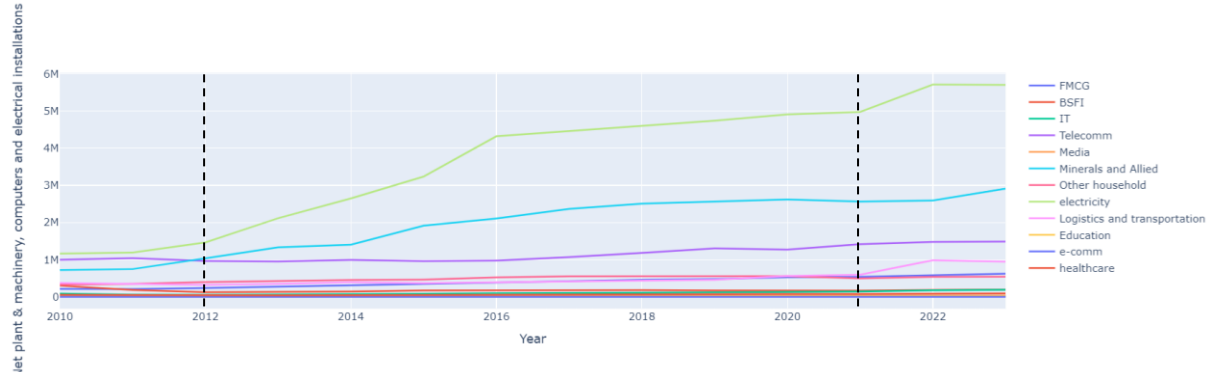



## Net Computers and IT systems

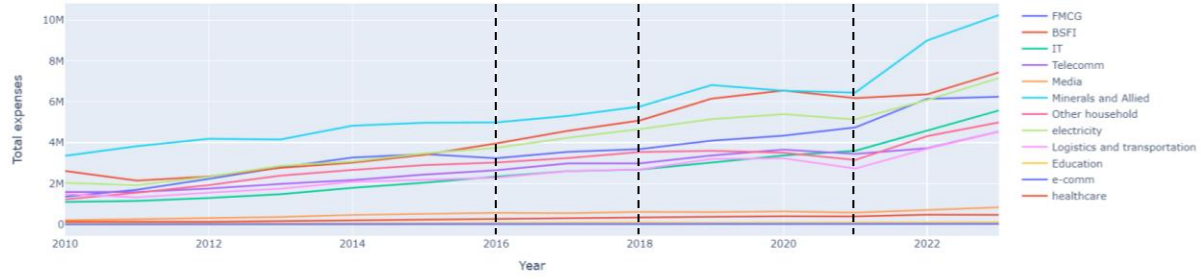



## Net Software

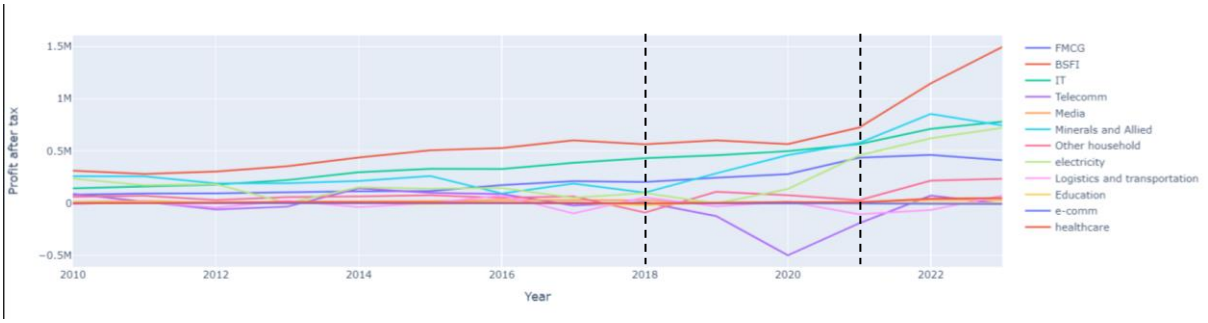

## Net plant & machinery, computers and electrical installations



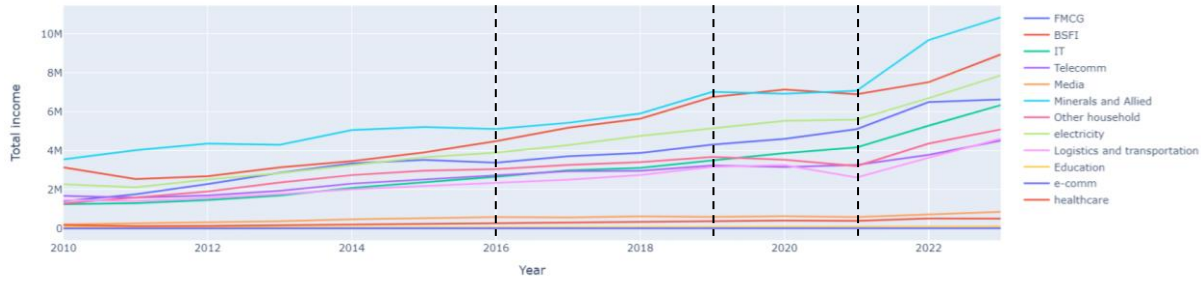
## Total Expenses




## Profit after Tax




## Total Income



# Results

Target Variable - Total income

Feature Variables - Computer IT, net addition in year, Net computers and IT systems, Net plant & machinery, computers and electrical installations, Net software, Profit after tax, Total expenses.

## 1) Random Effect Model:

F-statistic (robust) = 2.366e+05

P-value = 0

| Feature | Parameter | Std. Err. | T-stat | P-value | Lower CI | Upper CI |
|---|---|---|---|---|---|---|
| **Constant** | 118.13 | 37.081 | 3.1856 | 0.0014 | 45.447 | 190.81 |
| **Computer IT, net addition in year** | -0.0980 | 0.0481 | -2.0374 | 0.0416 | -0.1923 | -0.0037 |
| **Net computers and IT systems** | 0.4955 | 0.0389 | 12.741 | 0.0000 | 0.4193 | 0.5718 |
| **Net plant & machinery, computers and electrical installations** | 0.0111 | 0.0006 | 19.461 | 0.0000 | 0.0100 | 0.0122 |
| **Net software** | 0.3513 | 0.0605 | 5.8066 | 0.0000 | 0.2327 | 0.4698 |
| **Profit after tax** | 1.0014 | 0.0017 | 593.22 | 0.0000 | 0.9981 | 1.0047 |
| **Total expenses** | 0.9854 | 0.0003 | 3571.0 | 0.0000 | 0.9849 | 0.9860 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Year.2011** | -48.739 | 51.627 | -0.9441 | 0.3451 | -149.93 | 52.449 |
| **Year.2012** | -92.942 | 51.629 | -1.8002 | 0.0718 | -194.13 | 8.2504 |
| **Year.2013** | -127.92 | 51.630 | -2.4777 | 0.0132 | -229.12 | -26.727 |
| **Year.2014** | -50.547 | 51.631 | -0.9790 | 0.3276 | -151.74 | 50.651 |
| **Year.2015** | -100.67 | 51.635 | -1.9497 | 0.0512 | -201.88 | 0.5308 |
| **Year.2016** | -57.647 | 51.640 | -1.1163 | 0.2643 | -158.86 | 43.569 |
| **Year.2017** | -94.903 | 51.644 | -1.8376 | 0.0661 | -196.13 | 6.3197 |
| **Year.2018** | -77.422 | 51.648 | -1.4990 | 0.1339 | -178.65 | 23.809 |
| **Year.2019** | -92.727 | 51.659 | -1.7950 | 0.0727 | -193.98 | 8.5254 |
| **Year.2020** | -111.48 | 51.662 | -2.1578 | 0.0309 | -212.73 | -10.219 |
| **Year.2021** | 32.526 | 51.655 | 0.6297 | 0.5289 | -68.718 | 133.77 |
| **Year.2022** | -83.294 | 51.685 | -1.6116 | 0.1071 | -184.60 | 18.010 |
| **Year.2023** | -100.46 | 51.717 | -1.9425 | 0.0521 | -201.83 | 0.9054 |

# 2) Fixed Effect Model:

**Time Invariant:**

F-statistic (robust) = 1.095e+06

P-value = 0

| **Feature** | **Parameter** | **Std. Err.** | **T-stat** | **P-value** | **Lower CI** | **Upper CI** |
|---|---|---|---|---|---|---|
| **Constant** | 94.703 | 36.531 | 2.5924 | 0.0095 | 23.102 | 166.30 |
| **Computer IT, net addition in year** | -0.0743 | 0.0514 | -1.4471 | 0.1479 | -0.1750 | 0.0263 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Net computers and IT systems** | 0.2438 | 0.0628 | 3.8816 | 0.0001 | 0.1207 | 0.3668 |
| **Net plant & machinery, computers and electrical installations** | 0.0026 | 0.0010 | 2.6141 | 0.0089 | 0.0006 | 0.0045 |
| **Net software** | 0.1066 | 0.0816 | 1.3072 | 0.1912 | -0.0532 | 0.2665 |
| **Profit after tax** | 0.9896 | 0.0022 | 446.62 | 0.0000 | 0.9852 | 0.9939 |
| **Total expenses** | 0.9946 | 0.0006 | 1768.2 | 0.0000 | 0.9935 | 0.9957 |
| **Year.2011** | -51.665 | 51.432 | -1.0045 | 0.3151 | -152.47 | 49.142 |
| **Year.2012** | -103.58 | 51.437 | -2.0137 | 0.0440 | -204.40 | -2.7618 |
| **Year.2013** | -143.42 | 51.443 | -2.7880 | 0.0053 | -244.25 | -42.593 |
| **Year.2014** | -69.502 | 51.452 | -1.3508 | 0.1768 | -170.35 | 31.343 |
| **Year.2015** | -123.76 | 51.465 | -2.4047 | 0.0162 | -224.63 | -22.888 |
| **Year.2016** | -81.481 | 51.481 | -1.5827 | 0.1135 | -182.38 | 19.422 |
| **Year.2017** | -123.85 | 51.503 | -2.4046 | 0.0162 | -224.79 | -22.899 |
| **Year.2018** | -111.41 | 51.526 | -2.1621 | 0.0306 | -212.40 | -10.414 |
| **Year.2019** | -140.18 | 51.579 | -2.7177 | 0.0066 | -241.27 | -39.083 |
| **Year.2020** | -160.05 | 51.585 | -3.1027 | 0.0019 | -261.16 | -58.946 |
| **Year.2021** | -9.1878 | 51.566 | -0.1782 | 0.8586 | -110.26 | 91.882 |
| **Year.2022** | -140.45 | 51.704 | -2.7165 | 0.0066 | -241.79 | -39.113 |
| **Year.2023** | -175.58 | 51.860 | -3.3858 | 0.0007 | -277.23 | -73.939 |

**Time Effect:**

F-statistic (robust) = 7.223e+05

P-value = 0

F-test for Poolability: 1.5273

P-value: 0

| Feature | Parameter | Std. Err. | T-stat | P-value | Lower CI | Upper CI |
|---|---|---|---|---|---|---|
| **Constant** | -7.7341 | 11.325 | -0.6829 | 0.4946 | -29.931 | 14.462 |
| **Computer IT, net addition in year** | -0.0743 | 0.0514 | -1.4471 | 0.1479 | -0.1750 | 0.0263 |
| **Net computers and IT systems** | 0.2438 | 0.0628 | 3.8816 | 0.0001 | 0.1207 | 0.3668 |
| **Net plant & machinery, computers and electrical installations** | 0.0026 | 0.0010 | 2.6141 | 0.0089 | 0.0006 | 0.0045 |
| **Net software** | 0.1066 | 0.0816 | 1.3072 | 0.1912 | -0.0532 | 0.2665 |
| **Profit after tax** | 0.9896 | 0.0022 | 446.62 | 0.0000 | 0.9852 | 0.9939 |
| **Total expenses** | 0.9946 | 0.0006 | 1768.2 | 0.0000 | 0.9935 | 0.9957 |

# 3) First Difference OLS

F-statistic (robust) = 1.048e+05

P-value = 0

| Feature | Parameter | Std. Err. | T-stat | P-value | Lower CI | Upper CI |
|---|---|---|---|---|---|---|
| **Computer IT, net addition in year** | -0.0888 | 0.0524 | -1.6951 | 0.0901 | -0.1915 | 0.0139 |
| **Net computers and IT systems** | 0.3811 | 0.1444 | 2.6398 | 0.0083 | 0.0981 | 0.6640 |
| **Net plant & machinery, computers and electrical installations** | 0.0068 | 0.0025 | 2.6943 | 0.0071 | 0.0018 | 0.0117 |
| **Net software** | 0.1266 | 0.1365 | 0.9274 | 0.3537 | -0.1409 | 0.3941 |
| **Profit after tax** | 0.9849 | 0.0030 | 323.75 | 0.0000 | 0.9790 | 0.9909 |
| **Total expenses** | 0.9663 | 0.0013 | 746.87 | 0.0000 | 0.9637 | 0.9688 |

# 4) Generalised Method of Moments:

Instruments: Lag2 and Lag3 of Total income, Lag1 and following of Total expenses, Computer IT, net addition in year as IV.

**Hansen test of overid. restrictions**:

Chi-squared statistic: 158.782

Degrees of freedom: 136

p-value: 0.088

The Hansen test of overidentifying restrictions is a statistical method used to assess the validity of instruments in a dynamic panel data model. It evaluates whether the chosen instruments are correlated with the error term, which would indicate potential model misspecification. The test provides a chi-squared statistic and a p-value; a high p-value (typically above 0.05) suggests that the overidentifying restrictions are valid, indicating that the instruments used in the model are appropriate and do not introduce bias.

**Arellano-Bond test:**

AR(1): z = -1.46, P > z = 0.145: no significant first-order autocorrelation in the first differences.

AR(2): z = -0.48, P > z = 0.630: no significant second-order autocorrelation.

The Arellano-Bond test is a statistical procedure used to assess the presence of autocorrelation in dynamic panel data models, particularly those estimated using Generalized Method of Moments (GMM). It specifically tests for first-order (AR(1)) and second-order (AR(2)) autocorrelation in the first differences of the data. A significant AR(1) result indicates the presence of first-order autocorrelation, while a

non-significant AR(2) result suggests that the model does not suffer from second-order autocorrelation, which is crucial for the model's validity.

**Durbin-Watson:**

Result = 1.869

The test produces a statistic ranging from 0 to 4, where a value around 2 indicates no autocorrelation, values below 2 suggest positive autocorrelation, and values above 2 indicate negative autocorrelation.

| **Feature** | **Coeff.** | **Corrected Std. Err.** | **z** | **P>\|z\|** |
|---|---|---|---|---|
| **Computer, IT net addition in year** | 0. 068159 | 0.1035340 | 0.6583282 | 0.5103273 |
| **Net computers and IT systems** | 0.3610928 | 0.1402061 | 2.5754424 | 0.0100112 |
| **Net plant & machinery, computers and electrical installations** | 0.0089745 | 0.0044682 | 2.0085133 | 0.0445888 |
| **Net software** | 0.3305597 | 0.2313008 | 1.4291329 | 0.1529660 |
| **Profit after tax** | 0.9633002 | 0.0220621 | 43.6631184 | 0 |
| **Total expenses** | 0.9454708 | 0.0244629 | 38.6492266 | 0 |

## 5) Gradient Boost:

| Metric | Training Data | Testing Data |
|---|---|---|
| **Mean Absolute Error** | 400.65 | 992.62 |
| **Root Mean Squared Error** | 1421.56 | 13283.26 |
| **Mean Absolute Percentage Error** | 29.89514 | 9.19714 |
| **R-squared** | 0.998849 | 0.96302 |

## 6) XG Boost:

| Metric | Training Data | Testing Data |
|---|---|---|
| **Mean Absolute Error** | 433.84 | 2185.18 |
| **Root Mean Squared Error** | 1703.96 | 21047.52 |
| **Mean Absolute Percentage Error** | 23.89062 | 6.26622 |
| **R-squared** | 0.99834 | 0.90717 |

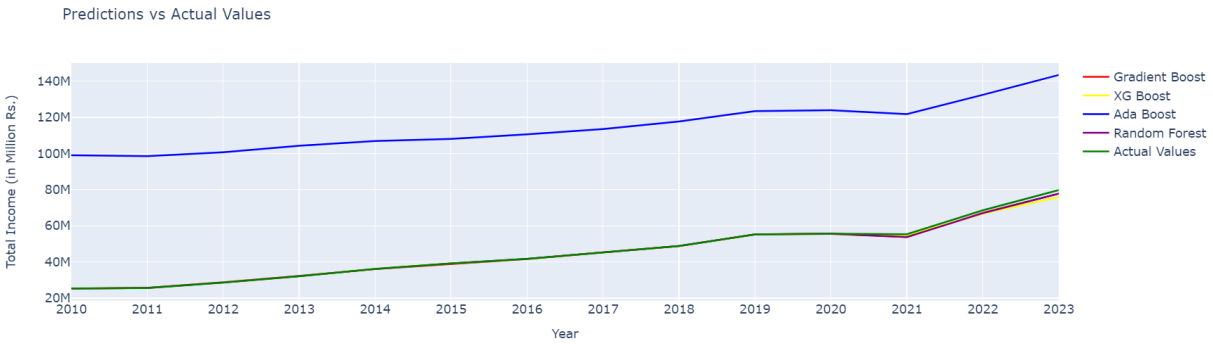
## 7) Ada Boost:

| Metric | Training Data | Testing Data |
|---|---|---|
| **Mean Absolute Error** | 18567.09 | 18682.38 |
| **Root Mean Squared Error** | 21280.64 | 25918.56 |
| **Mean Absolute Percentage Error** | 617.83158 | 273.95086 |
| **R-squared** | 0.74206 | 0.85924 |

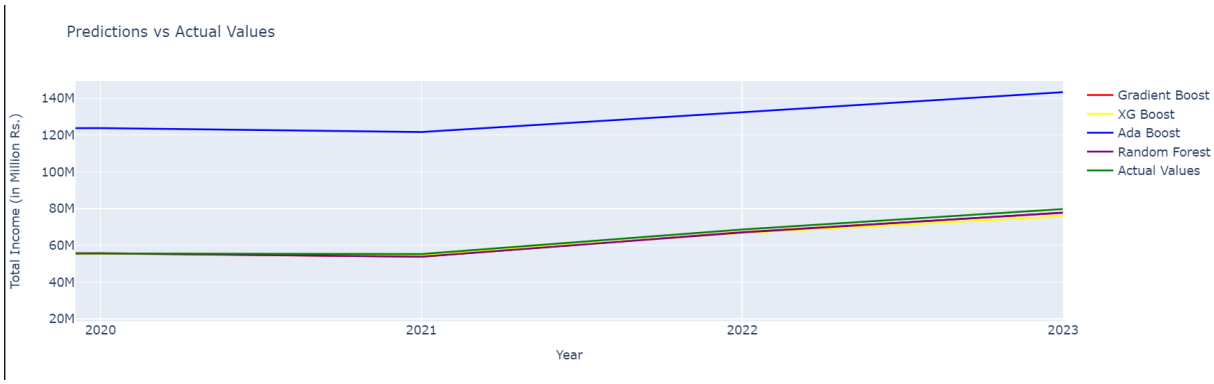
## 8) Random Forest:

| Metric | Training Data | Testing Data |
|---|---|---|
| **Mean Absolute Error** | 136.24 | 947.09 |
| **Root Mean Squared Error** | 1229.30 | 14965.60 |
| **Mean Absolute Percentage Error** | 10.91857 | 3.13830 |
| **R-squared** | 0.99913 | 0.95307 |

## Predicition and Actual Aggregate Total Income for ML models:



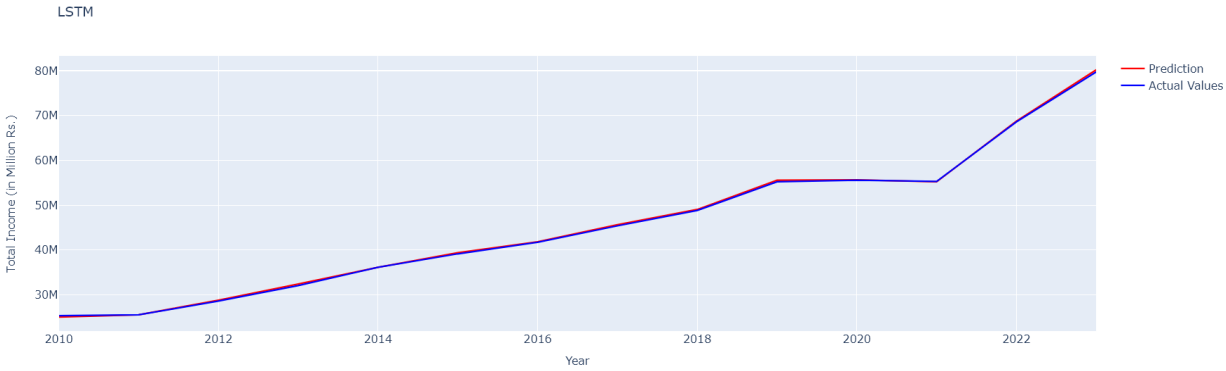

## Predicition and Actual Aggregate Total Incomes for ML models (between 2020 and 2023)



## 9) LSTM

| Metric | Training Data | Testing Data |
|---|---|---|
| **Mean Absolute Error** | 15.05 | 18.59 |
| **Root Mean Squared Error** | 2091.06 | 3046.13 |
| **Mean Absolute Percentage Error** | 4.83369 | 0.87278 |
| **R-squared** | 0.99875 | 0.99902 |

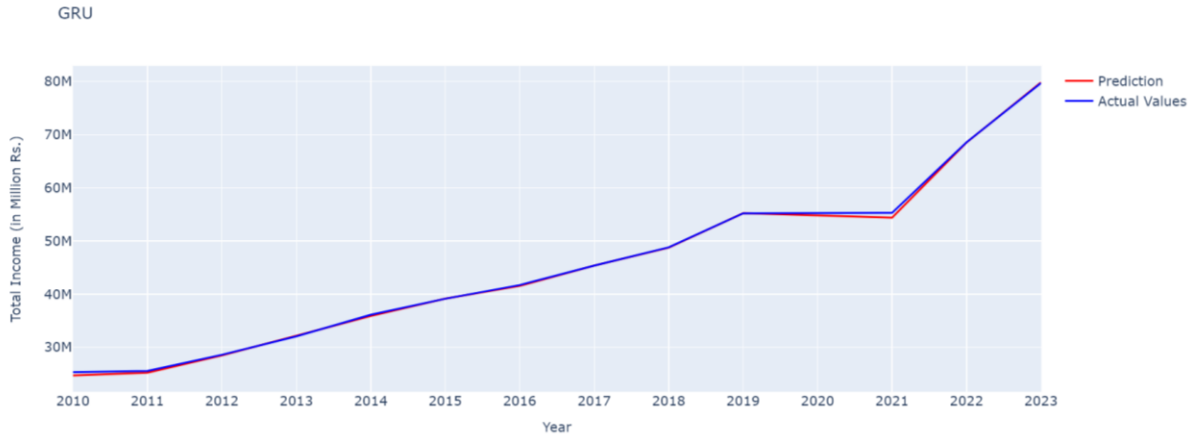
**Predicition and Actual Aggregate Total Income for LSTM**




## 10) GRU

| Metric | Training Data | Testing Data |
|---|---|---|
| **Mean Absolute Error** | 15.32 | 19.22 |
| **Root Mean Squared Error** | 2099.39 | 2721.17 |
| **Mean Absolute Percentage Error** | 4.56454 | 1.24426 |
| **R-squared** | 0.99874 | 0.99930 |

**Predicition and Actual Aggregate Total Income for Gru**



# Conclusion

## General Result of Trend Analysis:

In 2021, the BFSI and IT sectors saw significant digitization due to COVID-19, leading to increased digital tool adoption. The IT sector's growth was also driven by 3G/4G technologies, NDCP, and Digital India. Telecommunications experienced a revolution with 3G/4G and JIO's affordable data services. Media changes were mainly due to digital transformation during COVID-19, supported by Digital India and NDCP. The education sector's spikes were enabled by 3G/4G and accelerated by COVID-19. The FMCG sector saw fewer changes, influenced by UPI, demonetization, and digital transformation. Overall, digital transformation and COVID-19 were key drivers across sectors.

## Random Effect Model:

**Negative Impact of Computer IT Net Addition**: The coefficient for "Computer IT, net addition in year" is negative (-0.0980) with a p-value of 0.0416, indicating that an increase in net addition of IT systems is associated with a decrease in total income.

**Highly Significant Predictors**: "Net computers and IT systems," "Net plant & machinery," "Net software," "Profit after tax," and "Total expenses" are highly significant predictors ($p < 0.0001$).

**Yearly Effects**: The year variables show varying significance, with some years like 2013, 2015, 2019, and 2020 showing significant negative coefficients, indicating lower total income in these years.

## Fixed Effect Model:

**Time Invariance**: The time-invariant model shows high robustness (F-statistic: 1.095e+06) and significance.

**Profit After Tax Dominance:** "Profit after tax" has a very high coefficient (0.9896) and significance ($p < 0.0001$), reinforcing its critical role in predicting total income.

**Yearly Effects:** Similar to the random effect model, certain years (e.g., 2013, 2015, 2017, 2019, 2020, 2022, and 2023) have significant negative impacts on total income.

## First Difference OLS:

**Consistent Predictors:** "Net computers and IT systems," "Net plant & machinery," "Profit after tax," and "Total expenses" remain significant predictors with positive coefficients.

**Lower Significance for IT Net Addition:** "Computer IT, net addition in year" has a less significant impact (p-value: 0.0901) compared to other predictors.

## Generalised Method of Moments:

**Overidentifying Restrictions Valid:** The Hansen test's high p-value (0.088) indicates valid overidentifying restrictions, meaning the instruments used are appropriate.

**No Significant Autocorrelation:** The Arellano-Bond test shows no significant first-order or second-order autocorrelation, suggesting the model's validity.

## Gradient Boost:

**High Training Accuracy:** The model performs exceptionally well on training data (R-squared: 0.998849), indicating it fits the training data closely.

**Lower Testing Accuracy:** Testing data performance is lower (R-squared: 0.96302), suggesting potential overfitting.

## XG Boost:

**Good Training Fit:** High R-squared (0.99834) on training data.

**Performance Drop on Testing:** Noticeable drop in R-squared (0.90717) for testing data, indicating a slight overfitting.

## Ada Boost:

**Overfitting Issue:** The large difference between training and testing errors (MAE and RMSE) indicates significant overfitting, with much poorer performance on testing data.

## Random Forest:

**High Accuracy:** Random forest shows high R-squared for both training (0.99913) and testing data (0.95307), indicating good generalizability and fit.

**Low MAE and RMSE:** Relatively low errors for both training and testing data.

## LSTM:

**Excellent Fit:** Very high R-squared for both training (0.99875) and testing data (0.99902), indicating an excellent fit and predictive power.

**Low MAE and RMSE:** Extremely low mean absolute error and root mean squared error, demonstrating high accuracy.

## GRU:

**Top Performer:** Similar to LSTM, GRU shows outstanding performance with very high R-squared for both training (0.99874) and testing data (0.99930).

**Low MAE and RMSE:** Minimal errors, reflecting precise predictions.

## General Result of Models:

**Profit After Tax and Total Expenses:** These two variables consistently emerge as the strongest predictors of total income across all models, with high coefficients and significance.

**IT System and Software:** "Net computers and IT systems" generally have a positive impact, "Computer IT, net addition in year" shows a negative or less significant impact, Net plant & machinery, computers and electrical installations have low impact, while “Net Software” shows a positive and a significant impact at a higher threshold.

# References

## Academic References


1) **Thakurta & Deb, 2023.** Limited Effectiveness of IT/IS Investments in an Emerging Economy: Evidence from India and Implications.
    - https://dl.acm.org/doi/abs/10.1145/3583581.3583587

2) **Borenstein et al., 2010.** A basic introduction to fixed-effect and random-effects models for meta-analysis.
    - https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.12

3) **Liker et al., 1985.** Panel data and models of change: A comparison of first difference and conventional two-wave models
    - https://www.sciencedirect.com/science/article/abs/pii/0049089X85900134

4) **Ogaki, 1993.** Generalized method of moments: Econometric applications.
    - https://www.sciencedirect.com/science/article/abs/pii/S0169716105800525

5) **Mason et al., 1999. Boosting Algorithms as Gradient Descent.**
    - https://www.researchgate.net/publication/221618845_Boosting_Algorithms_as_Gradient_Descent

6) **Biau, 2012. Analysis of a Random Forests Model.**
    - https://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

7) **Hochreiter & Schmidhuber, 1997.** Long Short-term Memory.

- https://www.researchgate.net/publication/13853244_Long_Short-term_Memory

8) **Kyunghyun Cho et al., 2014.** Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

- https://arxiv.org/abs/1406.1078

9) **Pedregosa et al. 2011.** Scikit-learn: Machine Learning in Python.

- https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

## Online References

1) **ProwessIQ Data Dictionary:** https://prowessiq.cmie.com/

2) **Scikit-learn docs:** https://scikit-learn.org/0.21/documentation.html

3) **Pydynpd docs:** https://github.com/dazhwu/pydynpd

4) **Tensorflow docs:** https://www.tensorflow.org/api_docs

5) **Medium.com:** https://melaniesoek0120.medium.com/covid-19-global-data-time-series-prediction-with-lstm-recurrent-neural-networks-f7825c4a1f6f