Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 2 (2025)

Predictive Analysis of Organized Retail Penetration in India: A Study Using Multivariate Regression Techniques.

Kalipada Senapati[1], Dr. Ayan Chattopadhyay[2], Prof. (Dr.) Ranajit Chakrabarty[3]

[1]Assistant Professor, Dept. of Management Science & Humanities, MCKV Institute of Engineering, Howrah, WB and Research Scholar, Maulana Abul Kalam Azad University of Technology, India.
[2]Associate Professor, Army Institute of Management Kolkata, Affiliated to Maulana Abul Kalam Azad University of Technology, WB.
[3]Professor (Retd.), Department of Business Management, University of Calcutta, 157L, Prince Gulam Hossain Shah Road, Jadavpur, Kolkata - 700032.

Abstract:
This study investigates the influence of various economic and demographic factors on India's organized retail sector using 1997 to 2018 time series data. Due to data constraints from early organized retailing and the exclusion of the Covid – 19 period, the research aimed to develop a multivariate forecasting model for organized retail sales. Independent variables included population, median age, total personal disposable income, household consumption, employment, infrastructure investment, mall space and internet/smartphone users. Employing linear, polynomial, Ridge and Lasso regressions, the study addressed stationarity and multicollinearity. The Augmented Dickey Fuller (ADF) test identified employment, infrastructure investment and growing mall culture as key predictors. Final regression analysis, performed in Python addressed model performance and accuracy through comparative analysis.
JEL: C22, C51, C55

Keywords:
Organized retail sales, Augmented Dickey-Fuller (ADF), Ordinary least Square (OLS), Polynomial, Ridge and Lasso regressions.

1. Introduction

Forecasting and model development are crucial in management, yet challenging in India's retail sector due to data. Despite contributing 10% India's GDP, organized retail makes up only 18% of the total retail market- including brick-and-mortar and e-commerce ((CARE Ratings, 2019). As of 2024, organized retail and e-commerce were valued at US$ 95 billion and US$ 53.8 billion respectively.

Since India's liberalization in 1990, organized retail penetration has lagged behand developed nations. Existing studies often provide descriptive rather than analytical insights. Global research employs models like exponential smoothing, Box-Jenkins ARIMA, and machine learning methods such as LSTM) and ANN. However, a robust data driven model tailored to India remains absent.Despite projected growth rates of 15-20% CAGR, actual performance has often fallen short, underscoring the need for a reliable model.
The study addresses that gap by developing a quantitative model using 22 years of secondary data (1997-2018) to analyze the relationship between organized retail penetration (ORP) and its economic and demographic drivers.

The article proceeds as follows: Section 2 reviews literature, Section 3 outlines research gap, Section 4 presents research objectives and methodology. Section 5 covers the finding and analysis. Section 6 reports the conclusion with future research directions.

2. Literature Review

India's retail development has been linked to the macroeconomic stability, rising disposable income, urbanization and growing middle class (Chattopadhyay, 2018). Demographic changes have influenced consumer behavior, driving demand for branded goods, growth of organized retail and malls (Manoj, 2013), though organized retail still dominates (Chowdappa, 2020). Investments in infrastructure have further supported and digital technologies have further supported organized retail expansion. Digitalization, smart-phones use and internet access – especially post-pandemic have accelerated online retail adoption (Mittal, 2020; Roy et al., 2018).
The study uses a prediction model to analyze organized retail penetration (ORP) in India using variables such as population (POP), median age (MA), total disposable personal income (TDPI), household consumption expenditure (HHCE), employment (EMPL), infrastructure investment (INFRAINVEST), mall space (MALLSAREA), internet (INTUSER) and smartphone usage (SMPHUSER).

Regression models have been widely used in economic forecasting. Abdul-Muni & Quaidoo (2016) applied bound testing to analyze remittances' impact on Ghana's inflation, finding long-run but no short-run effects. Amral et al. ((2007) used Multiple Linear Regression (MLR) with polynomial terms for short-term load forecasting in Indonesia. Bianco et al. (2009) found GDP, not price, drives Itali's electricity consumption. Chaiyasoonthorn & Suksa-ngiam (2011) identified purchase intention and income as key factors influencing retail purchase in Bangkok. Upadhya et al. (2012) used logistic regression on financial ratios to predict stock performance on Indian firms with 56.8% accuracy. Gandhi and Shankar (2014) applied DEA (Data Envelopment Analysis), the Malmquist Productivity Index (MPI), and Bootstrapped Tobit Regression (BTR) to assess Indian retailer efficiency, finding outlet numbers and M&A activity significant. Wang (2016) used ARIMA to forecast economic growth in Shenzhen, China. Lalou et al. (2020) examined demand forecasting for retail networks, emphasizing analytics for 3PL providers. Young (2020) compared regression model and neural networks for Covid – 19 trends - using state level and Twitter data. Doi et al. (2025) proposed kernel Ridge regression in mixed data sampling for economic forecasting, outperforming traditional models.

3. Research gap and uniqueness of the study

The literature reveals extensive use of multivariate regression models in forecasting across various domains, including economic growth (Doi et al., 2025), electricity consumption (Amral et al., 2007; Bianca et al., 2009), stock performance (Upadhyay et al., 2012) and the efficiency of Indian retailers (Gandhi & Shankar, 2014). However, a comprehensive analysis of the Indian organized retail sector remains limited. While, Siddiqui and Tripathi (2016) explored the food and grocery retail, broader sector-wide insights are lacking.

Addressing data limitations and nonlinearity, Young (2020) effectively employed Ridge, Lasso and polynomial regressions. Yet, organized retail development in India using these advanced methods.

This study is the first to apply multivariate OLS, polynomial, Ridge and Lasso regression models to assess India's organized retail sector using economic and demographic variables under severe data constraints.

4. Research Objective and Methodology

The study aims to examine the relationship between macroeconomic and demographic variables and organized retail penetration (ORP) in India by developing a comprehensive model. It utilizes both linear and nonlinear regression techniques on annual time series data from 1997 to 2018, excluding the Covid-19 period (2019-2022) to avoid outlier effects. Nine independent variables - population, median age, disposable incomes, consumptions, employment, infrastructure investment, mall space, internet users and smartphone users – were selected based on literature and sourced from platforms like Tradingeconomics, E&Y, IBEF and Statista.

Table 1: Description of Variables

Variables with measuring units	Abbreviation	Variable type	Description
Organized retail penetration (INR billion)	ORP	Dependent	Organized retail sales value (INR billion)
Population in India (million)	POP	Independent	Total population in India
Median age (years)	MA	Independent	Median age of the Indian population
Total disposable personal income (INR billion)	TDPI	Independent	Aggregate disposable income
Household consumption expenditure (INR billion)	HHCE	Independent	Total household consumption
Employment (million)	EMPL	Independent	Number of employment in India
Infrastructure investment (INR billion)	INFRAINVEST	Independent	Investment in infrastructue in India
Mall Space (million Sq. ft.)	MALLSAREA	Independent	Mall Space in India
Number of internet users (million)	INTUSER	Independent	Number of interner users
Number of Smartphone Users (million)	SMPHUSER	Independent	Nuber of smartphone users

Source: Researchers' analysis.

The researchers employed these datasets to employ various regression models such as, multiple linear regression model, polynomial model for nonlinear relationship, Ridge regression and Lasso regression. This section focusses on sequentially describing these regression methods to explain their theoretical basis. Table 2 documents the values of the input variables.

Table 2: Input Variables

Year	ORP	POP	MA	TDPI	HHCE	EMPL	INFRAINVEST	MALLSAREA	INTUSER	SMPHUSER
1997	119.8	964.0	20.4	13000.0	9655.2	272.0	694.6	0.5	0.7	0.2
1998	158.5	983.0	20.8	15200.0	11130.0	284.0	869.2	1.0	1.3	0.4
1999	220.7	1000.0	21.0	16700.0	13655.6	296.0	937.9	1.5	2.7	0.7
2000	223.8	1020.0	21.2	18300.0	15931.8	330.0	1069.6	2.5	5.4	0.8
2001	254.8	1030.0	22.9	20200.0	17389.0	332.6	1184.5	2.9	6.8	1.0
2002	335.4	1060.0	23.1	21400.0	18609.1	324.0	1754.8	4.5	16.3	1.9
2003	352.1	1070.0	23.4	23600.0	20488.4	319.0	2412.8	6.5	18.0	2.9
2004	415.6	1090.0	23.6	32600.0	22097.3	320.6	2909.5	16.5	21.5	5.1
2005	466.6	1100.0	23.8	37100.0	24509.3	323.9	6705.8	19.0	26.3	8.0
2006	596.3	1120.0	24.1	43600.0	28080.8	327.3	10362.9	21.6	31.4	14.6
2007	397.0	1130.0	24.3	50500.0	33016.1	330.6	57598.5	30.0	44.6	20.0
2008	832.8	1150.0	24.6	56800.0	35062.1	460.8	169137.7	39.4	50.4	25.0
2009	618.7	1170.0	24.8	65800.0	43796.2	463.6	225470.1	45.6	59.9	30.0
2010	1299.2	1190.0	25.1	77900.0	50348.7	467.4	249966.1	53.4	89.3	34.0
2011	1721.2	1210.0	25.4	89600.0	57264.1	472.2	428045.6	65.4	121.8	58.7
2012	1937.7	1230.0	25.8	102000.0	65570.9	477.3	507843.3	68.6	136.5	90.6
2013	2217.5	1240.0	26.1	115000.0	71334.8	483.9	600154.0	72.9	152.5	129.1
2014	2662.7	1260.0	26.4	127000.0	87137.3	483.1	745451.2	74.7	170.1	190.0
2015	3400.4	1270.0	26.4	140000.0	91999.2	482.7	845240.6	77.7	189.2	250.7
2016	4260.1	1290.0	26.6	156000.0	106136.6	483.1	963468.6	81.0	212.9	304.5
2017	5928.2	1300.0	26.7	173000.0	124937.0	487.2	1078764.0	83.7	236.6	394.8
2018	7990.3	1320.0	26.8	192000.0	141681.1	491.1	1236956.7	89.3	265.1	400.0

Source: Tradingeconomics.com, IBEF.Org and Statista.

4.1 Multiple linear regression model

$$\mathbf{Y = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\ldots\ldots\ldots+\beta_kX_k+€} \quad \textbf{(1)}$$

Where β_0 is the intercept

β1,……βk regression coefficients of k explanatory or independent variables

€ = model error

Y is the dependent variable

The letter b is used for sample estimates of a β parameter and hence the equation takes the form:

$\mathbf{Y = b_0+ b_1X_1 + b_2X_2+ b_3X_3+\ldots\ldots\ldots+_{bkXk} +e,}$ e is the error term.

4.2 Polynomial Regression (Non-Linear Relationships)

For a polynomial regression model of degree 2, the equation is represented as:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{\beta_0} + \sum_{i=1}^{n} \boldsymbol{\beta_i}\, \boldsymbol{X_I} + \sum_{i=1}^{n}\sum_{i=1}^{n} \boldsymbol{\beta_{ij}X_iX_j} + \boldsymbol{\varepsilon} \quad \textbf{(2)}$$

Where, β0 is the intercept,

βi represents the coefficients for linear terms,

βij represents the coefficients for interaction and quadratic terms, ϵ is the error term.

4.3 Ridge Regression equation

$$\widehat{\boldsymbol{Y}} = \boldsymbol{\beta_0} + \sum_{i=1}^{n} \boldsymbol{\beta_i}\, \boldsymbol{X_I} + \boldsymbol{\lambda} \sum_{i=1}^{n} \boldsymbol{\beta_i^2} \quad \textbf{(3)}$$

where

$\hat{Y}$ is **ORP** (dependent variable)

iXi are the independent variables

βi are the Ridge regression coefficients

λ is the regularization parameter (controls shrinkage of coefficients).

Ridge regression retained all variables with reduced magnitudes, demonstrating its ability to prevent overfitting while controlling for multicollinearity.

4.4 Lasso Regression

Lasso regression eliminated some variables entirely (coefficient = 0), performing feature selection automatically. It selectively removes non-important predictors.
Helps with automatic feature selection.
The **Lasso Regression equation is given by-**

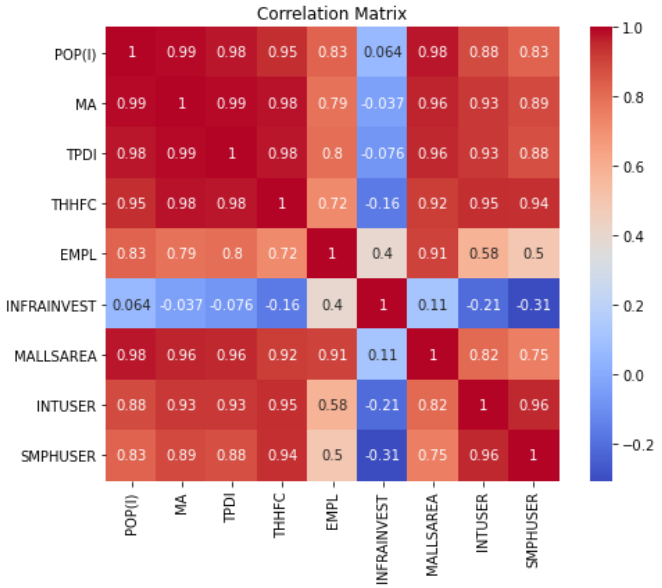
$$\hat{Y} = \beta_0 + \sum_{i=1}^{n} \beta_i X_I + \lambda \sum_{i=1}^{n} |\beta_i| \quad (4)$$

$\hat{Y}$ is **ORP** (dependent variable).
X_I are the independent variables
β_0 is the intercept.
βi are the Lasaso regression coefficients.
λ is the regularization parameter (controls shrinkage of coefficients).

5.Findings and analysis

A significant challenge in this data analysis was substantial multicollinearity among the independent variables. To address this, prior to conducting regression analysis using various methods, a correlation heatmap was generated in Python to visualize correlations among the independent variables.

Exhibit 1: Heatmap of Correlation Matrix (with absolute values)



Source: Researchers' analysis.

Exhibit 1 demonstrates strong correlations within the heatmap, confirming the presence of multicollinearity. These correlations, with values largely between 0.83 to 0.99 (excluding INFRAINVEST), suggested the necessity of data transformation, specifically through logarithmic or first differencing techniques. Following this, further generation of heatmap and estimation of VIFs for the variables are necessary.

5.1 Designing sequence of experiments

To address multicollinearity – a common issue in multiple regression, that leads to unstable coefficients estimates - researchers followed a structured approach. First, they measured Variance Inflation Factor (VIF) using Python to identify and reduce correlated predictors. Next, they tested for stationary before applying various regression models. Table 3 presents VIF results, confirming high multicollinearity and the need for corrective measures.

Table 3: Multicollinearity analysis

Variable	*VIF*
POP	79983.630
MA	86446.320
TDPI	363.580
HHCE	531.120
EMPL	955.320
INFRAINVEST	10.310
MALLSAREA	395.900
INTUSER	54.610
SMPHUSER	100.750

Source: Researchers' analysis

The researchers transformed the dataset using first differences, and assessed multicollinearity. Table 4 presents the new reduced VIFs of the independent variables.

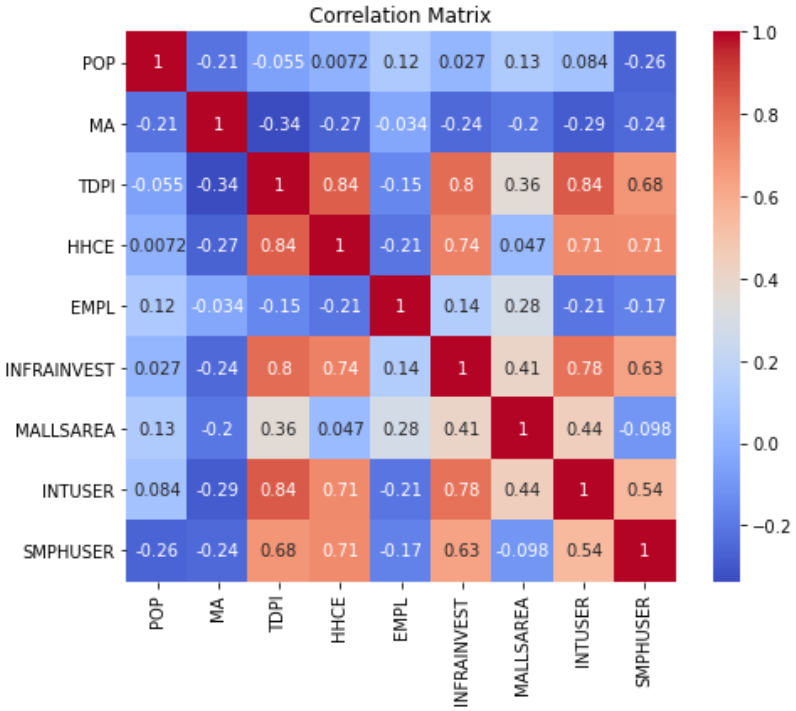
Table 4: VIFs with first-differenced independent data

Variable	*VIF*
POP	4.881
MA	1.592
TDPI	9.773
HHCE	13.168
EMPL	2.085
INFRAINVEST	10.377
MALLSAREA	6.622
INTUSER	13.809
SMPHUSER	4.500

Source: Researchers' analysis.

The correlation heatmap for the variables is presented in Exhibit 2. The correlation coefficients in the heatmap reduces substantially from higher values, shown in the previous heatmap (Table 3). Due to higher variance inflation factors (VIF), HHCE (13.17), and INTUSER (13.81) were excluded. The time series stationarity test was conducted with the remaining seven variables such as, POP, MA, TDPI, EMPL, INFRAINVEST, MALLSAREA and SMPHUSER using Augmented Dickey-Fuller (ADF) Test.

Exhibit 2: Heatmap of Correlation Matrix (First differenced values)



Source: Researchers' analysis.

5.2 Stationarity Test Using Augmented Dickey-Fuller (ADF)

The Augmented Dickey-Fuller (ADF) test was employed to assess the stationarity of the chosen time series variables. Stationarity, characterized by constant mean and variance over time, is crucial for statistical modeling: Non-stationary series exhibits time varying properties often necessitates data transformation like differencing.

5.2.1 Methodology

The ADF test was applied to each variable to assess stationarity. ADF Statistic is a test statistic used to determine stationarity. If a p-value of ADF statistic is less than the critical value at 5% of 0.05, then the null hypothesis is rejected, confirming that the time series is stationary, otherwise the series is non-stationary. Below are the ADF test results for each variable:

(a) **Population (POP) -** ADF Statistic: -1.603581; p-value: 0.481795
Conclusion: Failed to reject Ho; the time series is non-stationary.

(b) **Median age (MA) -** ADF Statistic: -1.859835; p-value: 0.351186
Conclusion: Failed to reject Ho; the time series is non-stationary.

(c)Total disposable personal income (TDPI) - ADF Statistic: 0.065348; p-value: 0.963614
Conclusion: Failed to reject Ho; the time series is non-stationary.

(d)Employment (EMPL) - ADF Statistic: 4.410679; p-value: 0.000284
Conclusion: Reject Ho; the time series is stationary.

(e) Malls Area (MALLSAREA) - ADF Statistic: -4.563708; p-value: 0.000732

Conclusion: Reject Ho; the time series is stationary.
(f) Infrastructure Investment (INFRAINVEST) - ADF Statistic: -4.189689; p-value: 0.039740
Conclusion: Reject Ho; the time series is stationary.
(g) Smartphone Users (SMPHUSER) - ADF Statistic: 8.406330; p-value: 1.000000
Conclusion: Failed to reject Ho; the time series is non-stationary.
The Augmented Dickey-Fuller (ADF) test results indicate that four variables -- Population (POP), Median Age (MA), Total Domestic Personal Income (TDPI), and Smartphone Users (SMPHUSER) - are non-stationary. The remaining three variables --Employment (EMPL), Malls Area (MALLSAREA), and Infrastructure Investment (INFRAINVEST) are stationary.
The ADF test results for seven variables are summarized in Table 5.

Table 5: Results summary of the ADF test

Variable	ADF Statistic	p-value	Conclusions
Population (POP)	-1.604	0.482	Non-stationary
Median Age (MA)	-1.860	0.351	Non-stationary
Total Domestic Personal Income (TDPI)	0.065	0.964	Non-stationary
Employment (EMPL)	4.411	0.000	Stationary
Malls Area (MALLSAREA)	-4.564	0.001	Stationary
(INFRAINVEST)	-4.190	0.040	Stationary
Smartphone Users (SMPHUSER)	8.406	1.000	Non-stationary

Source: Researchers' analysis.
The ADF test selected three variables – EMPL, INFRAINVEST and MALLSAREA which were subsequently used in various regression methods to develop models. The selected variables are presented in Table 6.

Table 6: ADF test Selected variables for developing models

Variable Name	Abbreviation	Variable type	Description
Organized retail penetration (INR billion)	ORP	Dependent	Organized retail sales value (INR billion)
Employment (million)	EMPL	Independent	Number of employment in India
Infrastructure investment (INR billion)	INFRAINVEST	Independent	Investment in infrastructue in India
Mall Space (million Sq. ft.)	MALLSAREA	Independent	Mall Space in India

Source: Researchers' analysis.

5.3 Various linear and nonlinear models
5.3.1 Multiple linear regression model (Ordinary least square)

Ordinary Least Squares (OLS) regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. In this analysis, we estimate the impact of employment (EMPL), infrastructure investment (INFRAINVEST), and mall area (MALLSAREA) on the dependent variable, organized retail penetration (assumed to revenue performance). The independent variables are extracted from the dataset and converted into a NumPy array. A constant term is added to the independent variable matrix to account for the intercept. The stats model's library is used to fit an OLS regression model. The model's summary statistics are examined to evaluate performance, shown in Table 7.

Table 7: MLR (OLS) Model statistics

Statistic	Value
R-squared	0.458
Adjusted R-squared	0.363
F-statistic	4.792
Prob (F-statistic)	0.0135
Log-Likelihood	-156.34
AIC	320.7
BIC	324.9
Observations (n)	21
Degrees of Freedom (Residual)	17
Degrees of Freedom (Model)	3

Source: Researchers' analysis.

R-squared (0.458): The model explains approximately 45.8% of the variance in ORP, suggesting a moderate fit.
Adjusted R-squared (0.363): Adjusting for the number of predictors, the model explains about 36.3% of the variance. Adding more predictors may not increase model's explainability power.
F-statistic (4.792, p = 0.0135): The F-test suggests that at least one of the independent variables significantly contributes to explaining ORP. Table value of F(3,17) = 3.20. Since the estimated F-statistic is larger than the critical value of 3.20, the independent variables have significant relationship with the dependent variable, ORP. organized retail penetration
AIC (320.7) & BIC (324.9): These values are useful for model comparison; lower values suggest better model fit. The coefficients of the multiple linear regression are presented in Table 8.

Table 8: Coefficients and Statistical Significance for linear regression

Variable	Coefficient	Std. Error	t-Statistic	p-Value	Confidence Interval (95%)
Constant	121.6846	169.287	0.719	0.482	(-235.481, 478.850)
EMPL	1.1186	3.748	-0.298	0.047	(-5.927, 6.790)
INFRAINVEST	0.0069	0.002	3.758	0.002	(0.003, 0.011)
MALLSAREA	-33.1728	33.688	-0.985	0.339	(-104.248, 37.903)

Source: Researchers' analysis.

EMPL and INFRAINVEST show a statistically significant ($p<0.05$), impact on the organized retail penetration (ORP), indicating a strong effect, while MALLSAREA does not, suggesting it has no substantial impact (p-value > 0.05).
Diagnostic Tests were conducted to test autocorrelations and normality of the residuals.
Durbin-Watson (1.325): This indicates some level of positive autocorrelation in residuals.
Jarque-Bera (p = 0.183): Suggests residuals are approximately normal.
Table 9 presents the results of autocorrelation and normality tests of residuals.

Table 9: Diagnostic Tests (MLR)

Durbin-Watson = 1.325	Jarque-Bera (p=0.183)
Positive autocorrelation in residuals.	Residuals are approximately normal

Source: Researchers' analysis.
Therefore, the multiple linear equation takes the form,
ORP (Y) = 121.684+1.1186(EMPL)+0.0069(INFRAINVEST)

The multiple linear regression analysis reveals that the independent variables explain 45.8% of the variation in organized retail penetration (ORP), with a statistically significant model ($p = 0.0135$). Specifically, EMPL and INFRAINVEST significantly affect ORP, while MALLSAREA does not.

5.3.2 Polynomial Regression

Polynomial regression is a form of regression analysis in which the relationship between the independent variable(s) and the dependent variable is modeled as an nth-degree polynomial. The model's performance is evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2).

Model Fitting (Polynomial): For our data, the model is given by the equation,
$ORP (Y) = b_0+b_1(EMPL) +b_2(INFRAINVEST) + b_3(MALLSAREA) + b_4(EMPL.INFRAINVEST) +b_5(EMPL.MALLSAREA) +b_6(INFRAINVEST.MALLSAREA) + EMPL^2 + INFRAINVEST^2 + MALLSAREA^2$

Table 10 presents the polynomial regression statistics below.

Table 10: Polynomial regression (OLS) statistics

Metrics	Value
R-squared	0.577
Adj. R-squared	0.395
F-statistic	3.179 (p = 0.0350)

Source: Researchers' analysis.

Comparing R^2 values of multiple linear regression and an initial OLS polynomial model (Tables 7 and Table 10), polynomial model demonstrates improved explanatory power (57.7%) compared to

multiple linear regression (45.8%) in explaining the variance of the dependent variable, ORP. Additionally, a p-value of 0.0350, being less than 0.05, indicates a statistical significance, suggesting a significant relationship between independent variables and organized retail penetration. Table 11 presents the metrics of second-degree polynomial model.

Table 11: Second-degree polynomial metrics

Metrics	Value
R-squared	0.7059
MSE	93056.34
MAE	229.95

Source: Researchers' analysis.

The second-degree polynomial regression explains 71% of the variance in the dependent variable. With an MSE of 93056.34 and an MAE of 229.95, indicating modest prediction error. It outperforms both multiple liner and initial OLS polynomial models, as reflected by a higher adjusted R-squared (0.395 vs.0.363). However, the relatively high error values suggest potential overfitting, which could be mitigated using regularization techniques like Ridge or Lasso.

5.3.3 Ridge Regression

Ridge and Lasso regression techniques used in machine learning to prevent overfitting by adding regularization terms to the cost function. Ridge (L2 regularization) penalizes the sum of the squared regression coefficients, helping manage multicollinearity and improve generalization.

The dataset used for this analysis was loaded from a CSV file and preprocessed as follows:
The dependent variable (target): ORP
Independent variables: EMPL, INFRAINVEST, and MALLSAREA
Data was split into training (80%) and testing (20%) sets using train test split from sklearn. The model was trained with an alpha value of 1.0.

Table 12 presents the Ridge coefficients, and Table 13 the error matrix.

Table 12: Coefficients of Ridge regression

Ridge Variables	Coefficients
Constant	142.5968
EMPL	1.1308
INFRAINVEST	0.007
MALLSAREA	-35.5719

Source: Researchers' analysis.

Table 13: Ridge metrics

Ridge metrics	Value
R-squared	0.5941
MSE	27037.7143
MAE	106.8326

Source: Researchers' analysis.

The ridge regression model demonstrates moderate predictive power with an R^2 of 0.5941, lower than the R^2 of 0.7059 from the second-degree polynomial model. MSE and MAE values are much lower than polynomial model. Overall, the ridge model effectively captures key trends between ORP and the independent variables, EMPL and INFRANET.

5.3.4 LASSO Regression

LASSO regression uses L1 regularization to reduce model complexity and prevent overfitting by shrinking some coefficients to zero, aiding feature selection and interpretability. The model was trained using an alpha value of 0.1. Table 15 shows the metrics of the Lasso Regression.

The dataset is split into training (80%) and testing (20%) sets: Training set - 16 observations and Testing set: 5 observations. The LASSO regression model was initialized with an alpha value of 0.1 and trained on the dataset. Table 14 presents the LASSO coefficients.

Table 14: LASSO coefficients

LASSO Variables	Coefficients
Constant	142.5968
EMPL	1.186
INFRAINVEST	0.007
MALLSAREA	-33.1728

Source: Researchers' analysis.

The presence of coefficients with values close to zero suggests that LASSO has performed some shrinkage, but no independent variables were entirely eliminated from the model. Table 15 presents the error of LASSO coefficients.

Table 15: Lasso metrics

LASSO Metrics	Value
R-squared	0.59
MSE	27092.33
MAE	107.19

Source: Researchers' analysis.

LASSO regression achieved a moderate fit with an R^2 of 0.59 effectively reducing coefficient magnitudes for better generalization. Table 16 presents a comprehensive metric comparison.

Table16: Comparative Analysis of Regression Models

Models	R^2 Scroe	MSE	MAE
Linear OLS	0.458	AIC-320.7	BIC-324.9
Polynomial	0.71	93056.34	229.95
Ridge	0.59	106.73	27037.71
LASSO	0.59	107.19	27092.33

Source: Researchers' analysis.

7. Conclusion

The analysis indicates that multiple linear regression model is the least effective, explaining only 45.8% of the variance in organized retail penetration. Its inability to capture non-linear relationship and the presence of some auto-correlation (Durbin-Watson = 1.325) limit the predictive capability, despite residuals appearing approximately normal (Jarque-Bera = 0.183).

The second-degree of polynomial Regression demonstrate the highest explanatory power (R^2 = 0.71), effectively capturing nonlinear trends. However, its high MSE and MAE suggest a risk of overfitting.

Ridge and Lasso regressions both achieve an R^2 of 0.59, but Ridge performs slightly better with lower MSE and MAE, indicating more stable and accurate predictions. Lasso, while similar in predictive performance, excels in automatic feature selection by shrinking irrelevant predictions to zero.

Future studies should incorporate cross-validation and hyperparameter tuning to optimize model performance. Exploring alternative models like Elastic Net could better balance regularization effects. Adding relevant variables and a large sample size is recommended. Where extended time series data is unavailable, panel cross sectional data can be used to improve model robustness.

References:

1. Abdul-Mumuni, A., & Quaidoo, C. (2016). Effect of international remittances on inflation in Ghana using the bounds testing approach. *Business and Economic Research*, *6*(1), 192-209.
2. Amral, N., Ozveren, C. S., & King, D. (2007, September). Short term load forecasting using multiple linear regression. In *2007 42nd International universities power engineering conference* (pp. 1192-1198). IEEE.
3. Anitha, R. (2012). Foreign direct investment and economic growth in India. *International Journal of Marketing, Financial Services & Management Research*, *1*(8), 108-125.
4. Bianco, V., Manca, O., & Nardini, S. (2009). Electricity consumption forecasting in Italy using linear regression models. *Energy*, *34*(9), 1413-1421.
5. CARE Ratings (2019). Indian Readymade Garments Apparel Industry Overview, April 2019.
6. Chaiyasoonthorn, W., & Suksa-ngiam, W. (2011). Factors influencing store patronage: A study of modern retailers in Bangkok Thailand. *International Journal of Trade, Economics and Finance*, *2*(6), 520.

7. Chattopadhyay, P. (2018). A study on promising trends in e-commerce with reference to e-retailing in Indian context: a conjectural approach, *International Journal of Research and Development,* 3(12), 81-86.
8. Chowdappa, V. A (2020). Comparative study of organized and Unorganized Retail sector in Karnataka state- a case study, *JETIR,* 7(4), 470-484.
9. Dai, D., Javed, F., Karlsson, P., & Månsson, K. (2025). Nonlinear forecasting with many predictors using mixed data sampling kernel ridge regression models. *Annals of Operations Research*, 1-20.
10. Gandhi, A., & Shankar, R. (2014). Efficiency measurement of Indian retailers using data envelopment analysis. *International Journal of Retail & Distribution Management*, *42*(6), 500-520.
11. Jhamb, D. and Kiran, R. (2012). Organized Retail in India – Drivers facilitator and SWOT Analysis, sian Journal of Management Research, Online Open Access publishing platform of Management Research, Issue 1, pp. 264 – 273.
12. Lalou, P., Ponis, S. T., & Efthymiou, O. K. (2020). Demand forecasting of retail sales using data analytics and statistical programming. *Management & Marketing*, *15*(2), 186-202.
13. Manoj,P.K.,(2013). Problems and Prospects of Retailing Industry in India: A Macro Perspective. *CLEAR International Journal of Research in Commerce & Management*, *3*(5).
14. Mittal, A. (2020). Trends and drivers of growth of organized retail industry in India. *IOSR Journal of Humanities and Social Science*, *25*(10), 23-33.
15. Roy, A., Mazumder, S., & Maity, S. K. (2018). Implications Of foreign direct investment on GDP and Indian retail business: opportunities for domestic entrepreneurial ventures. *International Journal on Recent Trends in Business and Tourism (IJRTBT)*, *2*(3), 48-59.
16. Upadhyay, A., Bandyopadhyay, G., & Dutta, A. (2012). Forecasting stock performance in indian market using multinomial logistic regression. *Journal of Business Studies Quarterly*, *3*(3), 16.
17. Yang, Z. (2020, October). Machine learning methods on COVID-19 situation prediction. In *2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE)* (pp. 78-83). IEEE.
18. Wang, T. (2016). Forecast of Economic Growth by Time Series and Scenario Planning Method – A case Study by Shenzhen, Modern Economy, 2016, 212-222.