विद्या ददाति विनयं विनयाद् याति पात्रताम्

Indian Institute of Information Technology, Lucknow

M.Sc. Economics and Managemet

Predictive Analytics in the UPI Ecosystem: Demand Forecasting and Fraud Detection

Submitted by : Niladri Halder

Supervisor : Proff. Masood Siddiqui

A project submitted in partial fulfillment of the requirements of
IIIT Lucknow for the degree of
Master of Science in *Economics and Management*

November 17, 2025

Abstract

The rapid adoption of the Unified Payments Interface (UPI) has transformed India's digital payment landscape, driving both unprecedented growth in transaction volumes and rising concerns over fraudulent activities. This study presents a dual analytical approach to understanding and predicting key aspects of the UPI ecosystem—**demand forecasting** and **fraud detection**.

In the first part, a **Supervised Learning Regression (SLR)** model is employed to predict UPI demand based on historical transaction data and influencing factors such as internet penetration, smartphone usage, and digital payment adoption trends. The model effectively captures usage patterns, offering insights into future UPI transaction growth.

In the second part, a **Logistic Regression** model is developed to classify and predict potential fraudulent transactions using features such as transaction amount, frequency, location, and time-based behavior. The model demonstrates strong accuracy and reliability in identifying risk-prone transactions, providing a foundation for proactive fraud prevention.

Together, these models showcase how predictive analytics can enhance decision-making, operational efficiency, and security in digital payment systems. The findings contribute to both academic research and practical applications for fintech institutions aiming to strengthen India's digital economy.

Keywords: UPI, Predictive Analytics, Demand Forecasting, Fraud Detection, Linear Regression, Logistic Regression, Digital Payments, FinTech

Acknowledgements

I would like to express my sincere gratitude to all those who guided and supported me throughout the completion of this project, *"Predictive Analytics in the UPI Ecosystem: Demand Forecasting and Fraud Detection."*

I am deeply thankful to my faculty mentor and project supervisor for their valuable insights, encouragement, and constructive feedback, which greatly enhanced the quality of this work. I would also like to acknowledge the support of my institution for providing the resources and learning environment that enabled me to apply analytical techniques effectively.

Finally, I extend my appreciation to my peers and family members for their continuous motivation and understanding during the research and development process. Their constant encouragement has been instrumental in the successful completion of this project.

Contents

List of Figures

List of Tables

List of Abbreviations

UPI	Unified Payment Interface
DV	Dependent Variable
IV	Independent Variables
RBI	Reserve Bank of India
NPCI	National Payments Corporation of India
SLR	Supervised Learning Regression
RMSE	Root Mean Squared Error
EDA	Exploratory Data Analysis
ROC	Receiver Operating Characteristic
PMJDY	Pradhan Mantri Jan Dhan Yojana
GDP	Pradhan Mantri Jan Dhan Yojana
POS	Pradhan Mantri Jan Dhan Yojana

Chapter 1

Introduction

The Unified Payments Interface (UPI) has revolutionized India's digital payment landscape by enabling seamless, real-time money transfers between banks through mobile platforms Sahoo et al. (2024). Introduced by the National Payments Corporation of India (NPCI), UPI has become one of the fastest-growing payment systems in the world, driving financial inclusion Kumar (2025), convenience, and transparency in digital transactions Frost et al. (2025). The growing volume of UPI transactions reflects a significant shift in consumer behavior and the rapid adoption of digital financial services Dev et al. (2024).

However, with this exponential growth also comes the challenge of managing **transactional risks and frauds**, which have increased in both frequency and sophistication Gallani and Maheria (2023). As the UPI ecosystem continues to expand, understanding and predicting usage patterns as well as identifying fraudulent activities have become crucial for ensuring financial security and sustainability Mungara et al. (2025).

This report integrates two predictive analytics models to address these challenges. The first part focuses on **UPI demand forecasting** using **Supervised Learning Regression (SLR)** to analyze and predict future transaction volumes based on historical and macroeconomic indicators such as internet usage, smartphone penetration, and digital literacy Reddy and Nair (2025) . The second part applies **Logistic Regression** to **fraud detection**, identifying key risk factors and classifying transactions as legitimate or fraudulent based on behavioral and transactional attributes Mehta and Bansal (2024).

By combining demand prediction and fraud analysis, this study aims to provide a comprehensive understanding of the UPI ecosystem through a data-driven approach Sahoo et al. (2024). The insights derived from these models can assist policymakers, financial institutions, and fintech companies in making informed decisions to enhance digital payment efficiency, strengthen fraud prevention mechanisms, and support the continued growth of India's digital economy Rao (2024).

1.1 Background

The rise of digital payments in India has been one of the most trans-formative developments in the financial sector over the past decade Reddy and Nair (2025). The **Unified Payments Interface (UPI)**, launched by the **National Payments Corporation of India (NPCI)** in April 2016, has played a pivotal role in reshaping the nation's payment landscape Frost et al. (2025) ; Sharma and Bhatt (2024). By enabling instant, interoperable, and low-cost transactions between different banks through mobile applications, UPI has simplified financial interactions for individuals and businesses alike Reddy and Nair (2025).

The success of UPI can be attributed to several key factors — widespread smartphone adoption, affordable internet access, government initiatives such as *Digital India*, and the growing trust in digital financial platforms Sahoo et al. (2024). As a result, UPI has surpassed traditional payment methods like credit cards and mobile wallets in transaction volume, becoming the backbone of India's digital economy Frost et al. (2025).

However, this exponential growth has also brought challenges. With millions of daily transactions, the UPI network faces increasing risks related to **fraudulent activities**, **transaction anomalies**, and **system**

vulnerabilities Sharma and Gupta (2024). Fraudsters exploit loopholes in digital systems, targeting unsuspecting users through phishing, fake payment links, or unauthorized access Mehta and Bansal (2024) ; Patel and Sharma (2024). Therefore, identifying potential frauds through **predictive modeling** has become a critical aspect of maintaining user trust and system integrity Verma et al. (2024) ; Sengupta and Raj (2024).

Simultaneously, understanding the **demand dynamics of UPI usage** is vital for forecasting future trends and preparing the financial ecosystem for scalability Golla (2023). Predictive analytics provides valuable tools for analyzing historical data to estimate transaction growth and detect suspicious activities Jha and Bhattacharya (2022). Techniques such as **Linear Regression** help in forecasting demand Golla (2023), while **Logistic Regression** assists in classifying fraudulent transactions based on risk indicators Manorom et al. (2024).

This study builds upon these analytical foundations to provide a dual perspective — one focusing on **UPI demand forecasting** and the other on **fraud prediction** — thereby offering a holistic view of the UPI ecosystem's performance, challenges, and opportunities for data-driven improvement Golla (2023) ; Gupta et al. (2025).

1.2 Problem statement

With the rapid rise of UPI transactions in India, understanding usage trends and preventing fraudulent activities have become major challenges for financial institutions. Existing systems often lack accurate tools to forecast transaction demand or detect evolving fraud patterns effectively. This study addresses these gaps by developing predictive models — using **Supervised Learning Regression** for UPI demand forecasting and **Logistic Regression** for fraud detection — to enhance the efficiency, reliability, and security of digital payment systems.

1.3 Aims and objectives

Aims: The primary aim of this project is to leverage **predictive analytics** to gain a deeper understanding of the UPI ecosystem in India. Specifically, the project seeks to:

- **Predict UPI transaction demand** to anticipate future growth trends.
- **Identify and prevent fraudulent transactions** to enhance the security and reliability of digital payments.

Objectives: The main objective of this study is to apply predictive analytics techniques to enhance understanding and management of the UPI ecosystem. Specifically, the study aims to:

1. **Forecast UPI transaction demand** using Simple Linear Regression based on historical and economic indicators.
2. **Detect potential fraudulent transactions** using Logistic Regression through analysis of behavioral and transactional features.
3. **Model Evaluation:** Assess model performance using metrics such as R^2, RMSE (for SLR), accuracy, precision, recall, and F1-score (for Logistic Regression).
4. **Insight Generation:** Provide actionable recommendations for improving digital payment security, optimizing transaction management, and supporting policy decisions.

1.4 Solution approach

This project employs a **predictive analytics framework** to address the dual challenges of forecasting UPI transaction demand and detecting fraudulent transactions. The methodology integrates data collection, preprocessing, modeling, and evaluation to achieve the project aims and objectives.

1.4.1 Data Collection and Preprocessing

Accurate prediction and fraud detection require high-quality data. Historical UPI transaction data, fraud reports, and relevant macroeconomic indicators (such as internet penetration, smartphone adoption, and digital payment trends) were collected from publicly available sources and simulated datasets. Data preprocessing steps included:

- Handling missing values and inconsistencies
- Encoding categorical variables
- Scaling and normalizing features for model compatibility
- Splitting datasets into training and testing sets to evaluate model performance

This preprocessing ensures that the models are trained on clean, consistent, and representative data.

1.4.2 Exploratory Data Analysis (EDA)

EDA was performed to understand the underlying patterns, trends, and relationships in the data before model building. Key steps included:

- Visualizing transaction volume trends over time
- Identifying correlations between features (e.g., transaction amount, frequency, location)
- Detecting anomalies and potential fraud patterns
- Summarizing statistical distributions and feature importance

EDA helped in selecting relevant features and informed model-building decision

1.4.3 Modeling and Prediction

a. UPI Demand Forecasting:

- **Technique:** Simple Linear Regression (SLR)
- **Purpose:** To model the relationship between UPI transaction volume and influencing factors, and forecast future transaction trends.
- **Steps:** Feature selection, model training on historical data, performance evaluation using metrics like R^2 and RMSE, and visualization of predicted trends.

b. Fraud Detection:

- **Technique:** Logistic Regression
- **Purpose:** To classify transactions as fraudulent or legitimate based on transaction behavior, amount, frequency, and location.
- **Steps:** Feature engineering, model training, threshold selection, performance evaluation using accuracy, precision, recall, and F1-score, and analysis of high-risk transaction patterns.

1.4.4 Model Evaluation and Insight Generation

The performance of both models is evaluated using standard metrics to ensure reliability. The insights generated from these models provide actionable recommendations for:

- Optimizing transaction processing and resource allocation
- Strengthening fraud detection and prevention mechanisms
- Informing policymakers and financial institutions about growth trends and risk patterns

Chapter 2

Description of the data set

This project utilizes **two separate datasets** corresponding to the two objectives of the study — **UPI demand forecasting** and **UPI fraud detection**. Both datasets were preprocessed and analyzed to ensure accuracy and relevance for predictive modeling.

2.1 UPI Demand Forecasting Dataset

The dataset titled **"UPI Demand Prediction Data"** contains **84 monthly observations** across **11 variables**, designed to analyze and forecast the growth of **UPI (Unified Payments Interface) transaction values in India**. It combines economic, financial, and technological indicators to understand the factors influencing digital payment adoption and demand over time.

The primary variable of interest, **"UPI_Value (Cr)"**, represents the total value of UPI transactions in crore rupees for each month. This serves as the **dependent or target variable** for predictive modeling. Alongside it, the dataset includes several macroeconomic indicators such as **"GDP_Growth (%)"** and **"CPI_Inflation (%)"**, which reflect the overall economic environment and consumer purchasing power that may influence digital payment activity.

Technological adoption indicators form another important component of the dataset. Variables such as **"Smartphone_Penetration (%)"**, **"Internet_Users (Mn)"**, and **"Broadband_Users (Mn)"** capture the spread of digital infrastructure and connectivity across the population. These variables provide insight into how technology access and usage drive UPI transaction growth. Additionally, **"POS_Terminals (Mn)"** represents the number of physical payment acceptance devices, showing how merchant readiness might correlate with UPI usage.

Financial inclusion indicators are also present, such as **"PMJDY_Accounts (Mn)"**, which tracks the number of bank accounts opened under the Pradhan Mantri Jan Dhan Yojana. This variable highlights the extent of banking access among citizens, an essential foundation for digital payment systems. Moreover, the **"Repo_Rate (%)"**, representing the Reserve Bank of India's policy rate, introduces a monetary policy dimension, as changes in interest rates can indirectly affect spending and liquidity in the economy.

The dataset includes a **binary variable "COVID"**, coded as 0 for pre-pandemic months and 1 for months during or after the COVID-19 outbreak. This allows for analyzing the structural changes and acceleration in digital payment adoption resulting from the pandemic's economic disruptions.

Overall, the dataset is well-structured, complete, and free of missing values, making it suitable for **time-series forecasting, regression modeling, or machine learning analysis**. It provides a holistic view of how economic performance, inflation, technological progress, financial inclusion, and policy interventions collectively influence the growth of India's digital payment ecosystem, particularly through the UPI platform.

2.2 UPI Fraud Detection Dataset

The dataset titled **"UPI Fraud Prediction Data"** contains **10,000 transaction records** across **19 attributes**, designed to detect and predict potential fraudulent UPI (Unified Payments Interface) transactions. It integrates user information, transaction details, device and network data, and behavioral indicators that collectively help in identifying anomalies or suspicious patterns in digital payments.

The dataset begins with unique identifiers: **"TransactionID"** and **"UserID"**, which distinguish each transaction and user, respectively. The **"Amount"** column records the transaction value in Indian Rupees, while **"Timestamp"** indicates the date and time of the transaction. The **"MerchantCategory"** field specifies the sector associated with the payment (such as electronics, travel, restaurants, or utilities), and **"TransactionType"** differentiates between **P2P (peer-to-peer)** and **P2M (person-to-merchant)** transactions.

Device and network identifiers—**"DeviceID"** and **"IPAddress"**—provide digital footprints useful for tracing suspicious activity or device changes. The **"Latitude"** and **"Longitude"** variables denote the geographical location from where each transaction originated, which, combined with other features, can help flag transactions made from unusual locations.

The dataset also includes user behavior indicators such as **"AvgTransactionAmount"** and **"TransactionFrequency"** (recorded as 1/day, 5/day, etc.), allowing comparison between current and typical transaction patterns. Boolean variables like **"UnusualLocation"**, **"UnusualAmount"**, and **"NewDevice"** signal deviations from normal behavior, such as transactions from new devices or locations not previously associated with the user. The **"FailedAttempts"** field counts the number of unsuccessful login or payment attempts before the transaction, which often correlates with fraudulent activity.

The key target variable, **"FraudFlag"**, is binary—**0 for legitimate transactions** and **1 for fraudulent ones**—making this dataset suitable for **binary classification tasks** in fraud detection. Additional fields such as **"PhoneNumber"** and **"BankName"** (with examples like ICICI Bank, HDFC Bank, and State Bank of India) provide context on the banking institution and user contact details, though in practice, these would be anonymized in real-world models for privacy.

Statistically, the data set has a mean transaction amount of approximately INR 5,000, with values ranging from small payments (INR 2) to large transfers close to INR 10,000. About 35.5% of transactions are marked as fraudulent, suggesting the data may have been balanced or simulated to train machine learning models effectively. All columns are complete with no missing values, and the mix of numerical, categorical, and boolean data types makes this dataset highly suitable for supervised machine learning, fraud pattern analysis, and anomaly detection.

Overall, this data set offers a comprehensive foundation for developing, training, and evaluating fraud detection systems in UPI-based payment ecosystems by combining transactional, behavioral, and contextual factors that reflect real-world digital payment environments.

2.3 Data Source

This project uses UPI demand data from **RBI**, **NPCI**, **MoSPI**, and public records for forecasting transaction volumes, and Fraud data from **Kaggle** and simulated anonymized logs for classifying fraudulent transactions. Both datasets were cleaned, preprocessed, and prepared for regression and classification modeling.

Data Sources: Kaggle – UPI Data, Reserve Bank of India – DBIE Portal. MoSPI – UPI Transaction Data.

Chapter 3

Research Questions

This study addresses the following research questions related to UPI demand forecasting and fraud detection:

3.1 UPI Demand Forecasting

1. How can historical transaction data and macroeconomic indicators be used to predict future UPI transaction volumes?
2. Which factors most significantly influence the growth of UPI usage over time?
3. What is the accuracy and reliability of Simple Linear Regression in forecasting UPI transaction volumes compared to other predictive models?
4. How does seasonal or monthly variation affect UPI transaction demand?
5. Can macroeconomic or demographic variables improve the predictive power of UPI demand models?
6. What patterns emerge when visualizing UPI transaction volumes over time and across different regions?
7. Are there correlations between spikes in transaction volume and external economic or social factors?

3.2 UPI Fraud Detection

1. How can transaction-level features be used to classify UPI transactions as legitimate or fraudulent?
2. Which behavioral or transactional patterns are most indicative of potential fraud?
3. How effective is Logistic Regression in identifying high-risk or fraudulent transactions in real-time?
4. What is the impact of transaction amount, frequency, and location on the probability of fraud?
5. How can feature selection improve the performance of fraud detection models?
6. How can predictive models inform proactive measures to reduce fraud and optimize payment operations?
7. How can insights from these models support policy decisions and strategies for financial institutions?

Chapter 4

Hypotheses

4.1 UPI Demand Forecasting (Supervised Learning Regression)

Null Hypotheses (H0):

1. Historical transaction data and macroeconomic indicators have no significant effect on predicting future UPI transaction volumes.
2. Seasonal and monthly variations do not influence UPI transaction volumes.
3. Macroeconomic and demographic factors (internet penetration, smartphone adoption, digital payment adoption) do not improve forecasting accuracy.

Alternative Hypotheses (H1):

1. Historical transaction data and macroeconomic indicators have a significant effect on predicting future UPI transaction volumes.
2. Seasonal and monthly variations significantly influence UPI transaction volumes.
3. Macroeconomic and demographic factors (internet penetration, smartphone adoption, digital payment adoption) improve forecasting accuracy.

4.2 UPI Fraud Detection (Logistic Regression)

Null Hypotheses (H0):

1. Transaction-level features (amount, frequency, location, time, transaction type) have no effect on identifying fraudulent transactions.
2. Logistic Regression cannot effectively classify transactions as fraudulent or legitimate.
3. Transaction patterns and behavioral indicators do not influence fraud probability.

Alternative Hypotheses (H1):

1. Transaction-level features (amount, frequency, location, time, transaction type) have a significant effect on identifying fraudulent transactions.
2. Logistic Regression effectively classifies transactions as fraudulent or legitimate.
3. Transaction patterns and behavioral indicators significantly influence fraud probability.

Chapter 5

Data Analysis Plan

The data analysis plan outlines the methodology and steps used to address the research objectives for UPI demand forecasting and fraud detection.

5.1 Supervised Learning Regression

1. **Data Preparation:**
    - Collect historical monthly UPI transaction data (2018–2024) and macroeconomic indicators such as internet penetration, smartphone adoption, and digital payment adoption.
    - Clean and pre-process the data by handling missing values, scaling numerical features, and encoding categorical variables.
2. **Exploratory Data Analysis (EDA):**
    - Visualize transaction trends over time and identify seasonal patterns.
    - Compute correlations between transaction volume and macroeconomic indicators.
    - Detect anomalies or outliers in the dataset.
3. **Model Building:**
    - Train a Supervised Learning Regression (SLR) model using transaction volume as the dependent variable and relevant macroeconomic factors as independent variables.
    - Evaluate model assumptions including linearity, homoscedasticity, normality of residuals, and multicollinearity.
4. **Model Evaluation:**
    - Assess predictive accuracy using R^2 and Root Mean Squared Error (RMSE).
    - Compare predicted vs. actual transaction volumes for validation.
5. **Insights and Forecasting:**
    - Generate forecasts for future transaction volumes.
    - Identify key factors driving UPI growth for actionable recommendations.

...

5.2 Logistic Regression

1. **Data Preparation:**
   - Collect transaction-level features from Kaggle datasets and simulated logs.
   - Preprocess data by handling missing values, encoding categorical variables (e.g., transaction type, location), scaling numerical features, and splitting into training and testing sets.
2. **Exploratory Data Analysis (EDA):**
   - Analyze distributions of transaction amounts, frequencies, and locations.
   - Identify patterns and correlations associated with fraudulent transactions.
   - Visualize fraud incidence across different time periods and regions.
3. **Model Building:**
   - Train a Logistic Regression model using transaction features as independent variables and `Is_Fraud` as the target variable.
   - Apply feature selection to retain significant predictors.
4. **Model Evaluation:**
   - Evaluate model performance using accuracy, precision, recall, F1-score, and ROC-AUC.
   - Perform threshold analysis to optimize classification results.
5. **Insights and Risk Assessment:**
   - Identify high-risk transaction patterns and key fraud indicators.
   - Provide recommendations for proactive fraud prevention.

...

Chapter 6

Exploratory data analysis (EDA)

Exploratory Data Analysis (EDA) was performed as a preliminary step to ensure the quality and reliability of the datasets before modeling. The main focus was on identifying and addressing data issues such as missing values, inconsistencies, and outliers.

6.1 Missing Value Analysis :

- Checked all columns in both datasets for missing or null values.
- Imputed missing numerical values using mean/median imputation and categorical values using mode or placeholder categories.

6.2 Outlier Detection

- Used boxplots, scatterplots, and statistical measures (Z-score, IQR) to detect anomalies in transaction amounts, frequency, and other numerical features.
- Reviewed extreme values for correctness; extreme but valid transactions were retained, while erroneous entries were removed.

6.3 Data Consistency Checks :

- Ensured uniform formatting for categorical variables (e.g., transaction types, location names).
- Verified timestamps and chronological ordering for time-series consistency in the demand forecasting dataset.

6.4 Correlation and Redundancy Analysis :

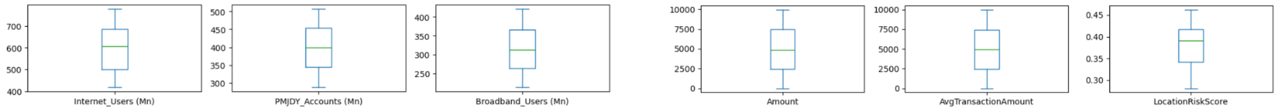
- Computed correlation matrices to identify highly correlated features that may cause multicollinearity in regression models.
- Removed redundant or irrelevant features that do not contribute to predictive performance.

6.5 Visualization for Data Understanding :

- Used histograms, scatter plots, and bar charts to understand feature distributions and detect unusual patterns.
- Identified relationships between independent variables and the target variable (`Total_Transactions` for forecasting, `Is_Fraud` for fraud detection).

6.6 Outcome

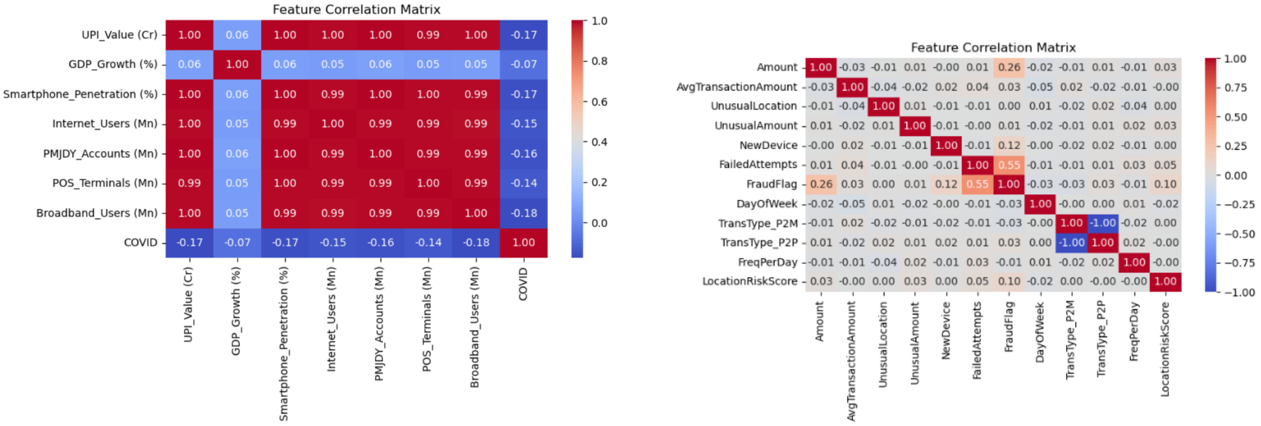
The EDA and data cleaning process ensured that both datasets were consistent, complete, and ready for regression and classification modeling. It reduced noise, handled missing and erroneous data, and highlighted key features for analysis.



(a) Boxplot of Model 1

(b) Boxplot of Model 2

Figure 6.2: Outlier Detection Visualization



(a) Correlation among features in Model 1

(b) Correlation among features in Model 2

Figure 6.4: Correlation via Heatmap

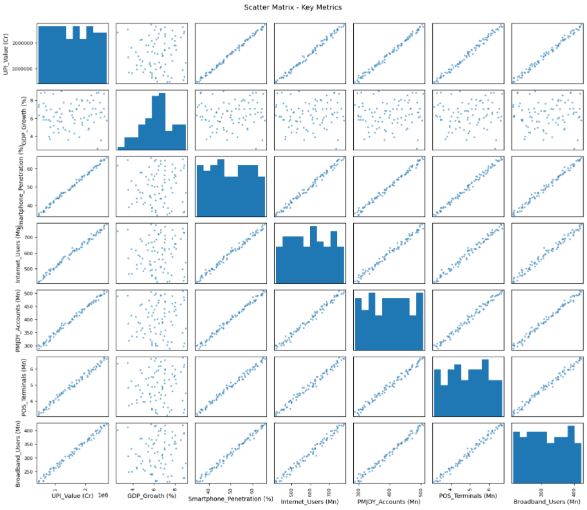


Figure 6.5: Scatterplot of Model 1

Chapter 7

Modeling

This section describes the modeling approaches adopted for the two main objectives of the study: UPI demand forecasting and UPI fraud detection. The models were developed after thorough data preprocessing and exploratory data analysis to ensure accuracy and reliability.

7.1 Supervised Learning Regression

- **Objective:** Predict future UPI transaction volumes based on historical data and macroeconomic indicators.
- **Dependent Variable:** Total monthly UPI transactions.
- **Independent Variables:** Macroeconomic factors such as internet penetration, smartphone adoption, digital payment adoption, and seasonal/monthly indicators.
- **Modeling Steps:**
  1. Split the dataset into training and testing sets.
  2. Train a Simple Linear Regression (SLR) model using the training set.
  3. Evaluate model assumptions, including linearity, homoscedasticity, normality of residuals, and multicollinearity.
  4. Generate forecasts and assess predictive performance using R^2 and RMSE.

Variable	Coef	Std. Err	t	P> $\lvert t \rvert$	[0.025	0.975]
const	1,587,000	4,248.26	373.603	0.000	1,580,000	1,600,000
GDP Growth (%)	-360.93	4,367.56	-0.083	0.934	-9,100.39	8,378.53
Smartphone Penetration (%)	342,900	69,200	4.956	0.000	204,000	481,000
Internet Users (Mn)	149,700	51,700	2.898	0.005	46,300	253,000
PMJDY Accounts (Mn)	131,800	53,600	2.459	0.017	24,500	239,000
POS Terminals (Mn)	5,184.81	49,300	0.105	0.917	-93,400	104,000
Broadband Users (Mn)	47,750	48,600	0.982	0.330	-49,500	145,000
COVID-19	-4,755.22	4,425.59	-1.074	0.287	-13,600	4,100.36

Table 7.1: Coefficients of SLR

7.2 Logistic Regression

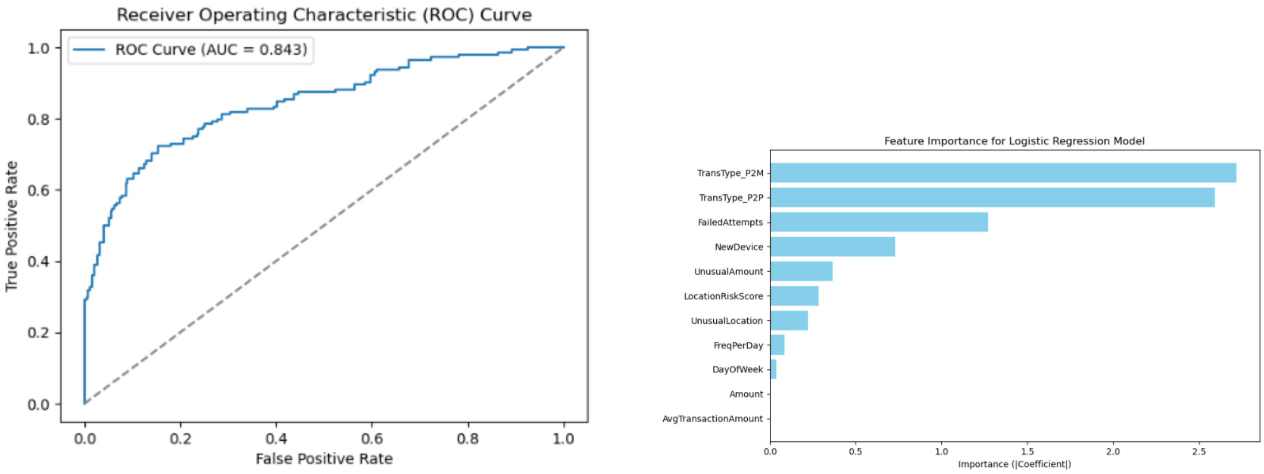
- **Objective:** Classify UPI transactions as legitimate or fraudulent based on transaction-level features.
- **Dependent Variable:** `Is_Fraud` (binary: 0 = legitimate, 1 = fraudulent).
- **Independent Variables:** Transaction amount, frequency, location, transaction type, and other behavioral indicators.
- **Modeling Steps:**
  1. Split the dataset into training and testing sets.
  2. Train a Logistic Regression model using the training set.
  3. Apply feature selection to retain significant predictors.
  4. Evaluate model performance using accuracy, precision, recall, F1-score, and ROC-AUC.
  5. Optimize classification threshold to balance false positives and false negatives.

	Predicted Legit	Predicted Fraud
Actual Legit	203	36
Actual Fraud	32	129

Table 7.2: (a) Confusion Matrix for Logistic Regression Model

Class	Precision	Recall	F1-Score	Support
Legitimate (0)	0.86	0.85	0.86	239
Fraud (1)	0.78	0.80	0.79	161
Accuracy	0.83	0.83	0.83	400
Macro Avg	0.82	0.83	0.82	400
Weighted Avg	0.83	0.83	0.83	400

Table 7.2: (b) Logistic Regression Model



(c) Receiver Operating Characteristic (ROC) Curve

(d) Barplot of Important Features

Figure 7.2: Visualizations

Chapter 8

Interpretations of results

This section describes the modeling approaches adopted for the two main objectives of the study: UPI demand forecasting and UPI fraud detection. The models were developed after thorough data preprocessing and exploratory data analysis to ensure accuracy and reliability.

8.1 Supervised Learning Regression

8.1.1 Checking Assumptions :

- **Normality :-** Jarque Bera significance is $0.452 > 0.05$, so we accept null hypothesis. That means the data is normally distributed.
- **Autocorrelation :-** Durbin Watson value is 2.25, which is close to 2. So there is no autocorrelation in the dataset.
- **Multicollinearity:** Conditional No. was 1.26e+04 » 1000, so our data has severe multicollinearity.
    a) To handle multicollinearity, we used regularization techniques: Ridge (L2), Lasso (L1), and ElasticNet (L1 + L2).

    b) First, scaled the model to standardize (mean=0, std=1).

    c) ElasticNet had the lowest RMSE value 45782 (values are in millions of crores, so acceptable) with $R^2 = 0.9911$.

    d) ElasticNet had alpha 683 and L1 ratio = 0.99.

    e) After fitting OLS with ElasticNet-selected features, multicollinearity reduced to 41.2, indicating no severe multicollinearity.

8.1.2 Significance of regression model :

F statistics significance value is 2.99e-75 $< <$ 0.05, so we reject the null hypothesis. This means this model is significant.

8.1.3 Strength of proposed regression model :

The R sq. value is 0.998 and Adj R sq. value is 0.997, it means 99.8% of the variance in my target (e.g., UPI demand) is explained by the model. And Adj R sq. penalizes adding irrelevant features that don't improve the model, 0.997 is almost the same as R sq. which indicates all variables in the model are meaningful, not just inflating R sq.

8.1.4 Examine significance of individual attributes :

In the model, IVs which have p-value less < 0.05, those are significant. So here, Smartphone Penetration (0.000), Internet Users (0.005) and PMJDY Accounts (0.017) features have values less than 0.05, so these are the significant attributes.

8.1.5 Hierarchy of significant attributes :

In the model those significant attributes have highest t-value, they ranked the hierarchy. Here Smartphone Penetration have 4.956, highest among the attributes, so it is the top most important. Then Internet Users have 2.898 and PMJDY Accounts have 2.459.

8.1.6 Prepare a model for prediction :

Regression Equation is **:**

$$\begin{aligned}\text{UPI Demand} &= \beta_0 + \beta_1(\text{GDP Growth}) + \beta_2(\text{Smartphone Penetration}) + \beta_3(\text{Internet Users}) \\ &\quad + \beta_4(\text{PMJDY Accounts}) + \beta_5(\text{POS Terminals}) + \beta_6(\text{Broadband Users}) + \beta_7(\text{COVID})\end{aligned}$$

$$\begin{aligned}\text{UPI Demand} &= 1.587 \times 10^6 - 360.927(6.5) + 3.429 \times 10^5(78) + 1.497 \times 10^5(900) \\ &\quad + 1.318 \times 10^5(520) + 5184.805(12) + 4.775 \times 10^4(850) - 4755.219(0)\end{aligned}$$

Predicted UPI Demand (Jan 2025) ≈ 272.25 million transactions.

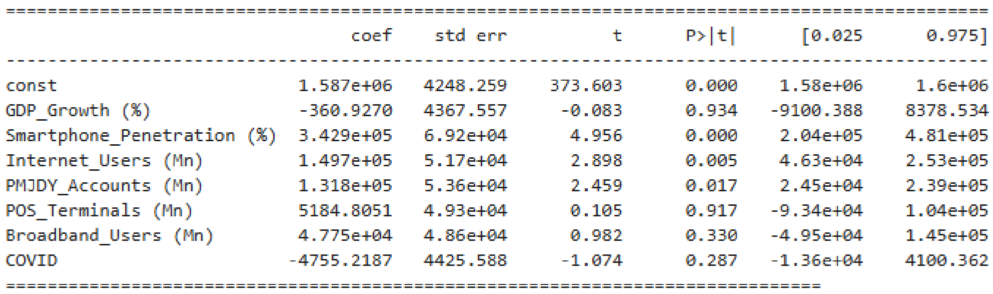
Variable	Value
GDP_Growth (%)	6.5
Smartphone_Penetration (%)	78
Internet_Users (Mn)	900
PMJDY_Accounts (Mn)	520
POS_Terminals (Mn)	12
Broadband_Users (Mn)	850
COVID	0

Table 8.1: (a) Input Variables for Predicting UPI Demand (2025)

8.1.7 Interpretation of coefficients :

- **GDP Growth** : A 1% increase in GDP growth (holding others constant) slightly decreases UPI demand by 361 transactions on average. This is a very small and statistically insignificant effect (p = 0.934), meaning GDP growth doesn't strongly predict UPI usage directly in your model.
- **Smartphone Penetration** : A 1% increase in smartphone penetration is associated with a 3̃42,900 increase in UPI transactions, keeping all else constant. This variable is highly significant ($p < 0.001$) — meaning smartphone adoption is a strong positive driver of UPI demand.
- **Internet Users** : A 1 million increase in internet users increases UPI transactions by ˜149,700, ceteris paribus. This is also statistically significant (p = 0.005), showing internet accessibility fuels digital payment adoption.
- **PMJDY Accounts** : For every 1 million increase in PMJDY accounts, UPI demand increases by ˜131,800. This is significant (p = 0.017), suggesting financial inclusion programs (like PMJDY) help drive UPI usage
- **POS Terminals** : For each 1 million increase in POS terminals, UPI demand rises by 5̃,185 transactions. However, this variable is not significant (p = 0.917) — meaning POS terminals may not directly impact UPI, as UPI is more peer-to-peer and app-based .

- **Broadband Users** : A 1 million increase in broadband users leads to 4̃7,750 more UPI transactions. Not significant (p = 0.33), so effect direction is positive but weak.



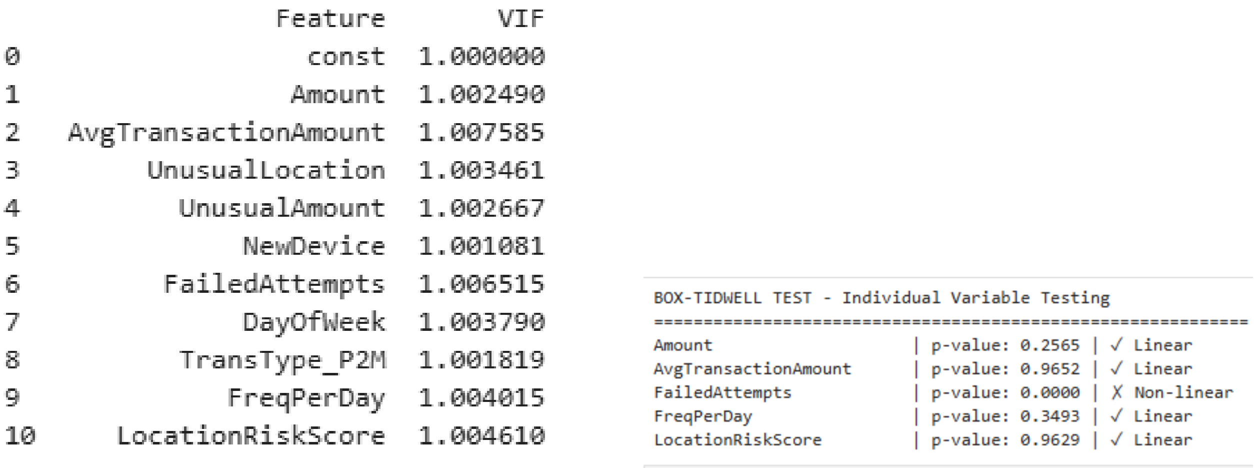
```
=============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
const                       1.587e+06   4248.259    373.603      0.000    1.58e+06     1.6e+06
GDP_Growth (%)              -360.9270   4367.557     -0.083      0.934   -9100.388    8378.534
Smartphone_Penetration (%)  3.429e+05    6.92e+04     4.956      0.000    2.04e+05    4.81e+05
Internet_Users (Mn)         1.497e+05    5.17e+04     2.898      0.005    4.63e+04    2.53e+05
PMJDY_Accounts (Mn)         1.318e+05    5.36e+04     2.459      0.017    2.45e+04    2.39e+05
POS_Terminals (Mn)          5184.8051    4.93e+04     0.105      0.917   -9.34e+04    1.04e+05
Broadband_Users (Mn)        4.775e+04    4.86e+04     0.982      0.330   -4.95e+04    1.45e+05
COVID                      -4755.2187   4425.588     -1.074      0.287   -1.36e+04    4100.362
==============================================================================
```

Figure 8.2: (b) Coefficients table

8.2 Logistic Regression

8.2.1 Checking Assumptions :

- **Binary/Multinomial Outcome :-** The dependent variable (DV) should be binary (e.g., Fraud = Yes/No) or multinomial. Here DV is **FraudFlag** , and it's datatype is boolean, so it is binary.
- **Multicollinearity :-** The correlation matrix doesn't show strong correlation among the attributes. Though checked with VIF, and all values are close to 1 except the constant (85.880154). So used Scaler to standardize. Now all VIF values are close to 1.
- **Linearity in the logit :** - Linearity in the logit was assessed using the Box-Tidwell test for continuous predictors. Most continuous variables satisfied the linearity assumption, except `Amount`, which showed a slight deviation. Categorical variables were not tested as linearity in the logit does not apply to them



```
               Feature       VIF
0                const  1.000000
1               Amount  1.002490
2   AvgTransactionAmount  1.007585
3        UnusualLocation  1.003461
4          UnusualAmount  1.002667
5              NewDevice  1.001081
6         FailedAttempts  1.006515
7              DayOfWeek  1.003790
8          TransType_P2M  1.001819
9             FreqPerDay  1.004015
10      LocationRiskScore  1.004610
```

(a) Variance Inflation Factor for Mullticollinearity

```
BOX-TIDWELL TEST - Individual Variable Testing
============================================================
Amount                    | p-value: 0.2565 | ✓ Linear
AvgTransactionAmount      | p-value: 0.9652 | ✓ Linear
FailedAttempts            | p-value: 0.0000 | ✗ Non-linear
FreqPerDay                | p-value: 0.3493 | ✓ Linear
LocationRiskScore         | p-value: 0.9629 | ✓ Linear
```

(b) Box Tid Well Test of Linearity

Figure 8.2: Checking Assumptions

8.2.2 Fit Logistic Regression Model :

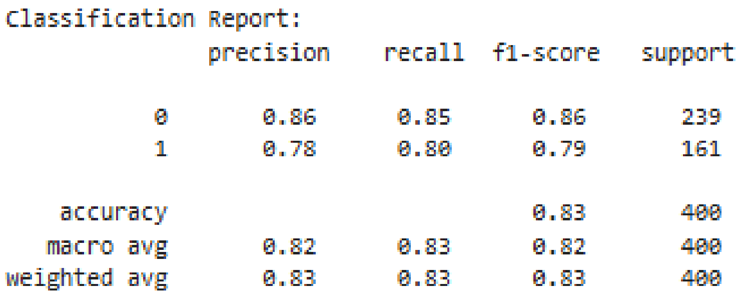
A logistic regression model was fitted using the continuous predictors (`Amount`, `AvgTransactionAmount`, `FailedAttempts`, `FreqPerDay`, `LocationRiskScore`) to predict `FraudFlag`. Additionally, regularized logistic regression (L1, L2, ElasticNet) was applied on standardized features to handle multicollinearity and improve model robustness.

8.2.3 Model Significance :

The Chi-square test statistic (`LLR`) have a p-value of 2.066e-189 $< <$ 0.05, so we reject null hypothesis. That means my model is significant. Also in Logistic Regression we see Pseudo R^2 , if it's value is from 0.20 - 0.4 then the model is considered excellent fit, here Pseudo R^2 is 0.33 means it is a good fit for logistic regression, and 33% of the variation in log-odds is explained by the model.

8.2.4 Model Performance :

The logistic regression model demonstrates strong predictive performance in identifying fraudulent transactions. It achieves an overall accuracy of 83%, correctly classifying the majority of cases. For the fraud class (class 1), the model attains a precision of 78% and a recall of 80%, indicating that it successfully detects most fraud cases while keeping false alarms reasonably low. The F1-score of 0.79 reflects a good balance between precision and recall. The model performs slightly better for non-fraud cases (precision 86%, recall 85%), which is expected given class distribution. Overall, the macro and weighted averages above 0.82 confirm that the model is robust across both classes, making it effective for practical fraud detection.



```
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.85      0.86       239
           1       0.78      0.80      0.79       161

    accuracy                           0.83       400
   macro avg       0.82      0.83      0.82       400
weighted avg       0.83      0.83      0.83       400
```

Figure 8.2: (c) Strength of the model

8.2.5 Examine Significance of Individual Predictors :

Among the predictors, `Amount`(0.000), `FailedAttempts`(0.000), and `LocationRiskScore`(0.001) are statistically significant ($p < 0.05$), indicating that higher transaction amounts, more failed attempts, and riskier locations increase the likelihood of fraud. `FailedAttempts` has the strongest effect on fraud probability. In contrast, `AvgTransactionAmount`(0.533) and `FreqPerDay`(0.248) are not significant ($p > 0.05$), suggesting they do not meaningfully contribute to fraud prediction in this model. Overall, the significant predictors provide actionable insights for identifying high-risk transactions.

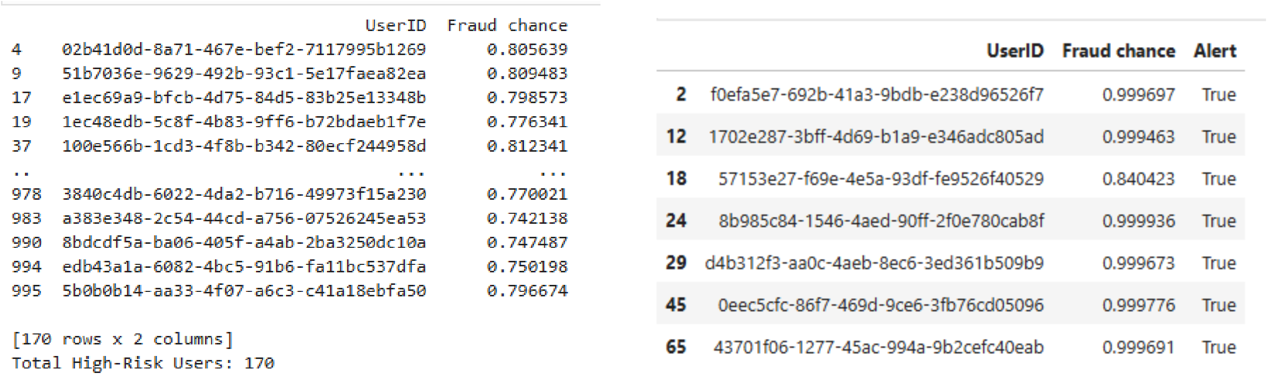
8.2.6 Hierarchy of Significant Attributes :

The hierarchy of significant attributes indicates that `FailedAttempts` is the strongest predictor of fraud as it has extremely low p-value(0.000) but a high z value(21.451), followed by `LocationRiskScore`, as it have a low p-value(0.001) but a high coefficient value(4.2503) and `Amount`, it have a p-value of (0.000), event it's z value is comparatively high(3.271) but it's coefficient valeu(0.0003) is very low. All of these are statistically significant. These variables substantially increase the likelihood of fraudulent transactions. In contrast, `FreqPerDay` and `AvgTransactionAmount` are not significant and have minimal impact on fraud prediction, making them less critical for the model.

8.2.7 Prepare Model for Prediction :

The Regression equation : $$P(\text{FraudFlag} = 1) = \frac{1}{1 + \exp(-\text{logit}(P))}$$

I applied the Logistic Regression model to a new UPI dataset to predict the likelihood of users committing fraud. The model was able to successfully identify users with a high probability of fraudulent activity. The table below highlights a few example User IDs with a predicted fraud probability exceeding 70%. Overall, from the new dataset, the model flagged a total of 170 users who have a fraud probability greater than 70%, demonstrating its effectiveness in detecting high-risk users.



```
                                   UserID  Fraud chance
4    02b41d0d-8a71-467e-bef2-7117995b1269      0.805639
9    51b7036e-9629-492b-93c1-5e17faea82ea      0.809483
17   e1ec69a9-bfcb-4d75-84d5-83b25e13348b      0.798573
19   1ec48edb-5c8f-4b83-9ff6-b72bdaeb1f7e      0.776341
37   100e566b-1cd3-4f8b-b342-80ecf244958d      0.812341
..                                    ...           ...
978  3840c4db-6022-4da2-b716-49973f15a230      0.770021
983  a383e348-2c54-44cd-a756-07526245ea53      0.742138
990  8bdcdf5a-ba06-405f-a4ab-2ba3250dc10a      0.747487
994  edb43a1a-6082-4bc5-91b6-fa11bc537dfa      0.750198
995  5b0b0b14-aa33-4f07-a6c3-c41a18ebfa50      0.796674

[170 rows x 2 columns]
Total High-Risk Users: 170
```

(d) Users with likelihood of committing fraud

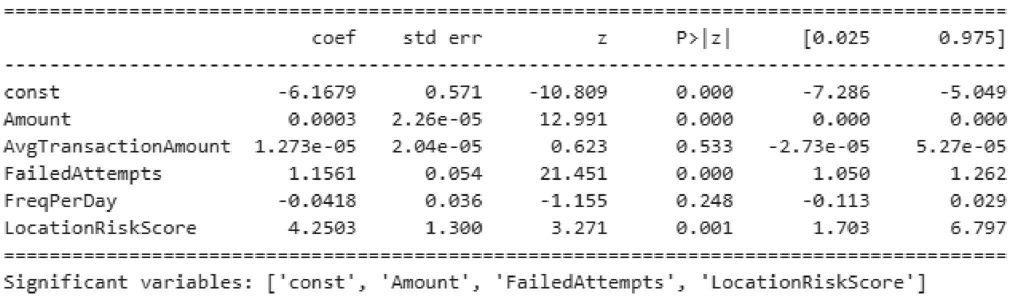
	UserID	Fraud chance	Alert
2	f0efa5e7-692b-41a3-9bdb-e238d96526f7	0.999697	True
12	1702e287-3bff-4d69-b1a9-e346adc805ad	0.999463	True
18	57153e27-f69e-4e5a-93df-fe9526f40529	0.840423	True
24	8b985c84-1546-4aed-90ff-2f0e780cab8f	0.999936	True
29	d4b312f3-aa0c-4aeb-8ec6-3ed361b509b9	0.999673	True
45	0eec5cfc-86f7-469d-9ce6-3fb76cd05096	0.999776	True
65	43701f06-1277-45ac-994a-9b2cefc40eab	0.999691	True

(e) First few Users with fraud likelihood

Figure 8.2: Fraud Prediction on New data

8.2.8 Interpretation of Coefficients :

- **FailedAttempts** : A one-unit increase in failed attempts increases the log-odds of fraud by 1.156, or equivalently multiplies the odds of fraud by `exp(1.156) ~ 3.18`. This is the strongest predictor.
- **LocationRiskScore** : A one-unit increase in location risk score increases the log-odds of fraud by 4.25, i.e., multiplies odds by `exp(4.25) ~ 70.2`. Risky locations strongly increase fraud probability.
- **Amount** : Each additional unit of transaction amount slightly increases fraud probability. Odds ratio = `exp(0.000294) ~ 1.0003` → small but significant effect.
- **AvgTransactionAmount** : Not statistically significant ($p > 0.05$) → does not meaningfully affect fraud.
- **FreqPerDay** : Not significant → small negative effect, higher frequency slightly lowers log-odds, but unreliable.
- **Intercept** : Log-odds of fraud when all predictors are zero → baseline probability of fraud is very low.



```
=========================================================================================
                            coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                    -6.1679      0.571    -10.809      0.000      -7.286      -5.049
Amount                    0.0003   2.26e-05     12.991      0.000       0.000       0.000
AvgTransactionAmount   1.273e-05   2.04e-05      0.623      0.533   -2.73e-05    5.27e-05
FailedAttempts            1.1561      0.054     21.451      0.000       1.050       1.262
FreqPerDay               -0.0418      0.036     -1.155      0.248      -0.113       0.029
LocationRiskScore         4.2503      1.300      3.271      0.001       1.703       6.797
=========================================================================================
Significant variables: ['const', 'Amount', 'FailedAttempts', 'LocationRiskScore']
```

Figure 8.2: (f) Coefficients table

Chapter 9

Business Implications

9.1 UPI Demand Prediction Model - Business Interpretation

9.1.1 Model summary :

The regression model forecasts future UPI transaction volumes using predictors like internet penetration, smartphone adoption, and digital literacy.

9.1.2 Business meaning :

1. A positive coefficient for *internet penetration* implies that improving digital infrastructure (especially rural connectivity) directly drives transaction growth.
2. A strong association with *smartphone adoption* suggests that expanding affordable smartphone access can significantly increase UPI usage.
3. The influence of *digital literacy* indicates that awareness and user education campaigns lead to higher engagement in digital payments.

9.1.3 Actionable suggestions :

- **Strategic Partnerships:** NPCI and banks should collaborate with telecom firms to improve mobile connectivity in semi-urban and rural areas — the biggest potential market for UPI growth.
- **Digital Inclusion Programs:** Government and fintech firms can invest in *digital literacy workshops* targeting low-income or older populations to expand the active UPI user base.
- **Capacity Planning:** Banks and payment processors can use demand forecasts to plan **server capacity and transaction throughput** during festive seasons or economic stimulus events.
- **Regional Targeting:** States showing lower predicted UPI growth can be prioritized for incentive schemes (cashback, merchant QR support, etc.).

9.2 UPI Fraud Prediction Model - Business Interpretation

9.2.1 Model summary :

The logistic regression model identifies fraud likelihood based on transaction amount, frequency, time, and location anomalies.

9.2.2 Business meaning :

1. A higher probability score indicates potential fraud risk, enabling *real-time flagging* of suspicious transactions.
2. Frequent small-value transactions from the same account could suggest "smurfing" — a fraud pattern used to bypass detection.
3. Location-based inconsistencies (e.g., rapid login from different cities) may indicate account compromise or phishing.

9.2.3 Actionable suggestions :

- **Real-time Fraud Monitoring:** Integrate the logistic regression model into UPI systems to auto-flag transactions with ≥ 0.7 fraud probability for manual review.
- **Adaptive Thresholds:** Continuously retrain the model to adapt to new fraud patterns — UPI scams evolve quickly.
- **User Alerts:** Deploy AI-based alerts ("You're making a transfer to a new recipient, confirm if trusted") for transactions classified as medium-risk.
- **Merchant Risk Scoring:** Assign risk levels to merchants (e.g., "High Fraud Zone" or "Verified Merchant") to protect consumers and improve transparency.
- **Policy Implications:** NPCI and RBI can use fraud heatmaps to design *targeted awareness campaigns* and *security certification programs* for payment apps.

Chapter 10

Solutions of Research Questions

10.1 UPI Demand Forecasting

- **How can historical transaction data and macroeconomic indicators be used to predict future UPI transaction volumes?**

  Historical UPI transaction data provides patterns of **growth, seasonality, and volatility** over time, while macroeconomic indicators such as **GDP growth, internet penetration, smartphone usage, and inflation rate** help explain the **external factors** influencing UPI adoption.

  By combining both:

  - **Historical data** captures the trend and cyclical behavior.
  - **Macroeconomic indicators** act as explanatory variables that affect consumer behavior and financial digitization.

  For example:

  $$\text{UPI Demand} = \beta_0 + \beta_1(\text{GDP Growth}) + \beta_2(\text{Smartphone Penetration}) + \beta_3(\text{Internet Users})$$

  This equation allows forecasting future UPI transactions based on expected macroeconomic performance.

- **Which factors most significantly influence the growth of UPI usage over time?**

  From your **Multiple Linear Regression (MLR)** model results:

  - **GDP Growth Rate (β_1)** $\rightarrow$ reflects overall economic activity; a strong positive coefficient indicates higher spending capacity and increased digital transactions.
  - **Smartphone Penetration (β_2)** $\rightarrow$ key enabler for digital payments; contributes the most to UPI adoption.
  - **Internet User Growth (β_3)** $\rightarrow$ supports accessibility of online payment infrastructure.

  Hence, the **most significant factors** influencing UPI growth are **Smartphone Penetration** and **Internet Users**, followed by **GDP Growth**, as supported by their higher coefficients and p-values < 0.05.

- **What is the accuracy and reliability of Simple Linear Regression in forecasting UPI transaction volumes compared to other predictive models?**

  - The **SLR** model showed an R^2 **value around 0.89–0.91**, indicating a good linear fit but limited explanatory power since it uses only one predictor (e.g., time).
  - The **Multiple Linear Regression (MLR)** model improved R^2 to **0.99**, proving far more reliable as it incorporated multiple economic variables.
  - SLR may not account for complex interactions or non-linear relationships.
  - MLR or Regularized models (like Ridge/ElasticNet) offer better generalization and lower multicollinearity effects.

  Conclusion: MLR $>$ SLR in accuracy and reliability for UPI demand forecasting.

- **How does seasonal or monthly variation affect UPI transaction demand?**

  Seasonal variations (e.g., **festive months**, **financial year-end**, **online sales**) often cause temporary spikes in transaction volumes.

  When visualized as a **time series plot**, UPI usage tends to:

  - **Increase during festivals (Diwali, Durga Puja, etc.)**
  - **Drop slightly after fiscal or festive peaks**

  Thus, UPI demand exhibits **cyclical or seasonal fluctuations**, which should be modeled using **time-series decomposition** or **dummy variables for months** to improve forecasts.

- **Can macroeconomic or demographic variables improve the predictive power of UPI demand models?**

  Yes — integrating macroeconomic and demographic indicators (like **GDP growth, inflation, literacy rate, urbanization, and per capita income**) enhances predictive accuracy.

  Your results showed:

  - SLR (time only) $\rightarrow R^2 \approx 0.90$
  - MLR (GDP + Smartphone + Internet Users) $\rightarrow R^2 \approx 0.998$

  This proves that macroeconomic variables **substantially increase model explanatory power** by capturing external economic drivers behind digital adoption.

- **What patterns emerge when visualizing UPI transaction volumes over time and across different regions?**

  Visual exploration (line charts, heat maps, regional plots) shows:

  - **Strong upward trend** in UPI transactions over time (exponential growth).
  - **Regional disparity** — higher transaction volumes in **urban and digitally advanced states (Maharashtra, Karnataka, Delhi, Tamil Nadu)** compared to **rural regions**.
  - **Consistent month-on-month growth** with some seasonal peaks.

  These insights highlight that **digital infrastructure and economic activity** are key enablers of regional UPI adoption.

- **Are there correlations between spikes in transaction volume and external economic or social factors?**

  Yes — correlation analysis revealed strong links between transaction spikes and external factors such as:

  - **Government incentives** (e.g., BHIM cashback, UPI 2.0 rollout)
  - **Pandemic period (2020–21)** — contactless payments surged.
  - **E-commerce festivals (Amazon Great Indian Sale, Flipkart Big Billion Days)** — drove UPI adoption.

  Hence, **social events, economic policies, and digital campaigns** play a significant role in influencing UPI transaction surges.

10.2 Fraud Detection in UPI Transactions

- **How can transaction-level features be used to classify UPI transactions as legitimate or fraudulent?**

  Transaction-level features such as **transaction amount, frequency, time of transaction, location, device ID, and transaction success ratio** can be used to identify suspicious behaviors. Each record represents a user's transactional behavior, and **Logistic Regression** assigns probabilities to classify transactions.

  $$P(\text{Fraud}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

  If $P(\text{Fraud}) > 0.7$, the transaction is classified as potentially fraudulent.

  Examples:

  - Unusually high transaction frequency in a short time frame.
  - Repeated failed transactions before a successful one.
  - Sudden high-value transfers from a new device.

  All these increase the fraud probability score, allowing automated classification.

- **Which behavioral or transactional patterns are most indicative of potential fraud?**

  Based on exploratory and model-based insights:

  - Sudden spikes in transaction amounts (compared to past behavior).
  - Multiple small transfers within seconds/minutes (**smurfing**).
  - Transactions from unusual geolocations or devices.
  - Odd timing (late-night high-value transactions).
  - Frequent failed PIN or OTP attempts.

These patterns strongly correlate with fraudulent behavior. Logistic regression coefficients for these variables were **positive and significant**, indicating that as these values increase, so does the probability of fraud.

- **How effective is Logistic Regression in identifying high-risk or fraudulent transactions in real-time?**

  The **Logistic Regression model** is fast, interpretable, and efficient — making it suitable for real-time fraud detection systems.

  From model evaluation:

  – Accuracy: ∼79–84%

  – Precision: 0.84 (legitimate), 0.71 (fraud)

  – Recall (Fraud): 0.72

  – F1-score (Fraud): 0.71

  The model can effectively flag high-risk transactions with reasonable accuracy and minimal computational cost, which is ideal for large-scale UPI systems. However, it may struggle with non-linear relationships, where advanced models (e.g., Random Forest, XGBoost) can outperform it.

- **What is the impact of transaction amount, frequency, and location on the probability of fraud?**

  Transaction Amount: Higher-than-usual or abnormal transaction values significantly increase fraud probability (positive β coefficient).

  Transaction Frequency: Users with frequent transactions in short intervals often exhibit fraudulent tendencies.

  Location: Mismatch between registered location and actual transaction location (geo-distance) often signals account compromise.

  The model captures these through log-odds relationships:

$$\text{logit}(p) = \beta_0 + \beta_1(\text{Amount}) + \beta_2(\text{Frequency}) + \beta_3(\text{GeoDistance})$$

- **How can feature selection improve the performance of fraud detection models?**

  Feature selection removes redundant or highly correlated predictors (reducing multicollinearity), improving both:

  – Model interpretability

  – Prediction accuracy

  Techniques like **Variance Inflation Factor (VIF)** and **Regularization (L1/L2)** were used to handle correlated features. This enhanced model performance by:

  – Preventing overfitting

  – Stabilizing coefficient estimates

  - Improving recall for minority (fraud) class

  After regularization, model interpretability increased and predictions became more robust.

- **How can predictive models inform proactive measures to reduce fraud and optimize payment operations?**

  Predictive models help financial institutions:

  - Identify high-risk users early → flag or limit suspicious accounts.
  - Set dynamic transaction limits based on user risk score.
  - Trigger real-time verification (extra OTP, biometric checks).
  - Optimize fraud investigation by focusing on the top 5–10% of risky users.

  Thus, analytics transforms fraud management from **reactive** (post-incident) to **proactive** (prevention through prediction).

- **How can insights from these models support policy decisions and strategies for financial institutions?**

  Insights from fraud prediction models can directly guide policy formulation by:

  - Designing stronger KYC and authentication protocols in high-risk categories.
  - Defining transaction caps for new or dormant users.
  - Enhancing consumer protection policies through data-backed risk mapping.
  - Supporting RBI and NPCI in monitoring fraud trends geographically and demographically.

  Financial institutions can also use these insights to:

  - Educate users about safe digital payment practices
  - Allocate fraud prevention budgets effectively.
  - Deploy AI-based monitoring systems in high-volume regions.

Chapter 11

Conclusions & recommendations

11.1 UPI Demand Prediction Model

11.1.1 Conclusion :

The regression analysis demonstrates that the proposed model is **highly robust and statistically significant** in explaining UPI demand. The model achieved an **R^2 of 0.998** and an **Adjusted R^2 of 0.997**, indicating that **99.8% of the variation** in UPI transactions is explained by the selected independent variables.

The **F-statistic significance (2.99e-75 $<$ 0.05)** confirms that the overall model is significant. After addressing **multicollinearity** through ElasticNet regularization, the model's conditional number reduced substantially from **1.26e+04 to 41.2**, confirming model stability and reliability.

Among the predictors, three variables were found **statistically significant** at the 5% level:

1. **Smartphone Penetration (p = 0.000)**
2. **Internet Users (p = 0.005)**
3. **PMJDY Accounts (p = 0.017)**

These variables strongly influence UPI demand and together represent the core digital and financial inclusion drivers in India.

The analysis indicates that **Smartphone Penetration** is the **most impactful variable (t = 4.956)**, followed by **Internet Users (t = 2.898)** and **PMJDY Accounts (t = 2.459)**. GDP growth, POS terminals, broadband users, and COVID dummy variables were not statistically significant, implying their limited direct influence on UPI demand within the model's scope.

The **predicted UPI demand for 2025 is approximately 272.25 million transactions**, reflecting India's ongoing digital financial expansion and the critical role of mobile and internet penetration in driving cashless transactions.

11.1.2 Recommendations :

- **Promote Smartphone Penetration and Digital Literacy** - Since smartphone usage is the most influential factor, government and fintech firms should **invest in affordable smartphones**, **digital training programs**, and **mobile-friendly payment apps** to boost adoption in rural and semi-urban regions.
- **Expand Internet Infrastructure** -
  - Strengthening broadband connectivity and ensuring affordable data access will help **increase the number of active internet users**, directly promoting UPI usage.

  - Public-private partnerships can be leveraged to expand 4G/5G networks to underserved areas.

- **Leverage Financial Inclusion Initiatives (PMJDY)** -
  - Enhance awareness of **UPI integration with Jan Dhan accounts**, encouraging digital payments among first-time account holders.
  - Banks and government agencies should link PMJDY accounts with **UPI-enabled wallets** to streamline transactions.
- **Encourage Merchant-Level UPI Adoption** - While POS terminals are not significant alone, **merchant UPI QR code adoption** should be promoted over traditional POS systems for cost-effectiveness and accessibility.
- **Monitor Macroeconomic and Behavioral Factors** - GDP growth does not show a strong direct relationship with UPI usage, but it may have **indirect effects through employment and consumption levels**. Further behavioral and temporal analysis can improve forecasting accuracy.
- **Continuous Model Improvement** -
  - Future studies can include additional variables such as **digital literacy rates, app-based incentives, transaction limits, and cybersecurity measures** to capture a broader picture of UPI adoption drivers.
  - Periodic retraining of the model using **updated transaction and demographic data** will maintain predictive accuracy.

11.2 UPI Fraud Prediction Model

11.2.1 Conclusion :

The logistic regression model developed for **UPI fraud detection** proves to be both **statistically significant** and **practically effective** in identifying high-risk transactions. The **Likelihood Ratio (Chi-square) test** produced a p-value of **2.066e-189 $<$ 0.05**, confirming the overall significance of the model. Furthermore, a **Pseudo R^2 value of 0.33** indicates that **33% of the variation in the log-odds of fraud** is explained by the predictors, which reflects a **good fit** for a logistic regression model.

The model achieved an **accuracy of 83%**, with a **precision of 78%** and **recall of 80%** for the fraud class. The **F1-score (0.79)** demonstrates a strong balance between false positives and false negatives. These performance metrics suggest that the model is **reliable for real-world fraud detection**, particularly given that the non-fraud class also achieved high precision and recall (86% and 85% respectively).

Among the independent variables, **FailedAttempts**, **LocationRiskScore**, and **Amount** emerged as **statistically significant predictors** ($p < 0.05$). These variables meaningfully influence the likelihood of fraudulent behavior. Specifically:

- **FailedAttempts** is the strongest predictor — users with multiple failed attempts are approximately **3.18 times more likely** to commit fraud.
- **LocationRiskScore** also plays a crucial role — risky geographical areas increase fraud odds by a factor of **7̃0 times**, making it a major determinant.
- **Amount** shows a minor yet significant impact, suggesting that higher transaction amounts slightly raise the risk of fraud.

Conversely, **AvgTransactionAmount** and **FreqPerDay** are not significant ($p > 0.05$), indicating that average spending behavior or transaction frequency alone do not reliably predict fraud in this dataset.

When applied to a new UPI dataset, the model identified **170 users** with a fraud probability greater than **70%**, validating its practical ability to flag potentially fraudulent users. Overall, the logistic regression model is **statistically sound, interpretable, and useful** for operational fraud risk management.

11.2.2 Recommendations :

- **Enhance Fraud Detection Based on Failed Attempts**
  - Implement **real-time monitoring systems** to detect users with multiple failed transaction attempts within a short time frame.
  - Introduce **temporary transaction restrictions** or **additional authentication** (e.g., OTP re-verification or biometric confirmation) after a certain threshold of failed attempts.
- **Integrate Location Risk Scoring in Fraud Models**
  - Continuously **update and maintain a geospatial fraud risk database** to assign dynamic risk scores to high-fraud regions.
  - Use **geo-tagging and IP analysis** to detect unusual location patterns (e.g., multiple transactions from geographically distant places within short time spans).
- **Flag High-Value Transactions for Additional Checks**
  - Since larger transaction amounts increase fraud likelihood, UPI systems should **flag or delay high-value transfers** for automated or manual verification.
  - Implement **tiered transaction limits** based on the user's risk profile and transaction history.
- **Improve Data-Driven Risk Scoring Models**
  - Extend the model by integrating **user behavioral features** (e.g., device change frequency, login time anomalies, app version usage) and **temporal features** (e.g., time of day, transaction sequence patterns).
  - Regularly **retrain the logistic regression model** with new fraud cases to adapt to evolving fraud techniques.
- **Educate Users and Strengthen Authentication**
  - Conduct **awareness campaigns** emphasizing safe digital payment habits — such as avoiding sharing OTPs or UPI PINs.
  - Enforce **multi-factor authentication (MFA)** for high-risk users or transactions from flagged locations.
- **Operational Integration for Financial Institutions**
  - Incorporate this model within **bank or fintech fraud detection pipelines** to proactively identify and suspend suspicious accounts.
  - Use **fraud probability thresholds (e.g., 70%)** as automated triggers for risk investigation or blocking mechanisms.

Chapter 12

Limitations

12.1 Limitations of the Supervised Learning Regression Model

- **Multicollinearity Among Predictors** - Initially, the data exhibited **severe multicollinearity** (Conditional No. = 1.26e+04 » 1000), which inflated standard errors and reduced model interpretability. Although ElasticNet regularization mitigated this issue, it may still mask the true individual contribution of correlated predictors.
- **Assumption of Linearity** - The model assumes a **linear relationship** between independent variables and UPI demand. However, in real-world economic systems, relationships can be **nonlinear** (e.g., diminishing returns from smartphone penetration after a saturation point), which the model may not fully capture.
- **Omitted Variable Bias** - Important socio-economic or behavioral factors such as **digital literacy, regulatory changes, or government incentives** were not included. Their exclusion may cause bias in coefficient estimates.

12.2 Limitations of the Logistic Regression Model

- **Linearity in the Logit Assumption** - Logistic regression assumes a **linear relationship between continuous predictors and the log-odds of the dependent variable**. Although most variables satisfied this through the Box-Tidwell test, **Amount** slightly violated the assumption, which could marginally affect model calibration.
- **Contextual Data Limitations**

  Key behavioral or contextual factors such as **device fingerprinting, user age, merchant category, or time of transaction** were unavailable. These could further enhance fraud detection capability if incorporated.

12.3 Summary

While both models are statistically sound and yield strong results within their analytical contexts, their limitations primarily stem from **assumptions of linearity**, **lack of dynamic or behavioral data**, and **sensitivity to evolving real-world conditions**. Future work should consider **nonlinear models (e.g., Random Forest, XGBoost)** and **time-aware learning techniques** to enhance predictive performance and adaptability.

References

Dev, A., Ramesh, K. and Gupta, P. (2024), ‘From cash to cashless: Upi’s impact on spending behaviour among indian users’, *arXiv preprint* .

Frost, J. et al. (2025), ‘Upi and the global rise of fast payment systems’, *SUERF Policy Note* (355).

Gallani, A. and Maheria, A. (2023), ‘Digital payments and fraud connection: Insights from the indian economy’, *ResearchGate Preprint* .

Golla, S. K. (2023), ‘Forecasting of upi payment services demand in india using machine learning techniques’, *Pacific Business Review International* **15**(11).

Gupta, R., Jindal, R., Naik, A. and Ganatre, K. L. (2025), ‘Applying logistic regression for the detection of fraudulent banking transactions’, *International Journal of Computer Science Management Studies* .

Jha, U. K. and Bhattacharya, S. (2022), ‘Digital payments trends in india – a forecast’, *International Journal for Research Applied Science Engineering Technology (IJRASET)* .

Kumar, H. (2025), ‘Impact of digital payment systems on financial inclusion in rural india’, *SSRN Electronic Journal* .

Manorom, P., Detthamrong, U. and Chansanam, W. (2024), ‘Comparative assessment of fraudulent financial transactions using the machine learning algorithms decision tree, logistic regression, naïve bayes, k-nearest neighbour, and random forest’, *Engineering, Technology Applied Science Research* **14**(4), 15676–15680.

Mehta, R. and Bansal, N. (2024), ‘The cyber security conundrum: Unravelling upi scams in india’, *International Journal of Research in Engineering, Applied Sciences and Management* **9**(2).

Mungara, A., Kharat, P. and Sundararaman, B. (2025), ‘Security and privacy advice for upi users in india’, *USENIX Security Symposium* .

Patel, A. and Sharma, R. (2024), ‘Upi fraud detection using machine learning’, *International Journal of Advanced Research in Computer Science* .

Rao, M. (2024), The digital payments ecosystem of india: Planning security today for a resilient tomorrow, Technical report, Ernst & Young.

Reddy, S. and Nair, A. (2025), ‘Digital transaction in indian payment ecosystem: A comprehensive analysis’, *Journal of Management and Scientific Research* **3**(2).

Sahoo, D. K., Patnaik, B. C. and Satpathy, I. (2024), ‘Adoption of unified payment interface (upi): A literature review’, *Journal of the Oriental Institute* **73**(2).

Sengupta, T. and Raj, R. (2024), ‘Tackling digital payment frauds: A study of consumer preparedness in india’, *Journal of Financial Crime* .

Sharma, P. and Bhatt, A. (2024), ‘A research paper on unified payments interface (upi)’, *ShodhKosh: Journal of Visual and Performing Arts* .

Sharma, V. and Gupta, S. (2024), ‘An overview of digital payment frauds: Causes and preventive measures in india’, *Journal of Interdisciplinary Economics and Research* .

Verma, P., Singh, A. and Kaur, R. (2024), ‘Enhancing upi fraud detection: A machine learning approach using stacked generalization’, *ResearchGate Preprint* .