# Personal Expense Forecasting System

A professional end-to-end Machine Learning pipeline for predicting and optimising personal finances through intelligent automation

Presented by – Sakshi Birajdar

# Presentation Overview

01

## The Problem

Understanding the core challenge in modern personal finance management

02

## Our Solution

A data-driven forecasting approach that transforms financial planning

03

## System Architecture

How our intelligent pipeline processes and predicts expenses

04

## Data Science

Deep dive into preprocessing, feature engineering and analysis

05

## Model Selection

Rigorous benchmarking to identify the champion algorithm

06

## Key Features & Capabilities

Key Features & Capabilities of project

07

## Future Vision

Roadmap for enhanced features and broader deployment

# The Challenge: Moving Beyond Reactive Budgeting

## Traditional Budgeting Limitations

Most people manage finances through manual, backward-looking tracking. This reactive approach only identifies overspending after it occurs, creating unnecessary financial stress and making effective saving difficult.

Static budgets fail to adapt to seasonal changes—festival spending, holiday expenses, or lifestyle shifts catch us unprepared. The result? Constant financial anxiety and missed savings opportunities.

### Static Planning

Budgets don't adapt to life changes

### Backward Looking

Identifies problems too late

### Manual Tracking

Time-consuming and error-prone

> **The key question:** How can we leverage machine learning to shift from reactive expense tracking to proactive, forward-looking financial planning?

# Our Solution: Intelligent Automation Meets User Insight

### Raw Transaction Data

User's historical spending records

### AI Processing Pipeline

Automated cleaning, feature engineering and model training

### Interactive Dashboard

Personalised forecasts and actionable recommendations

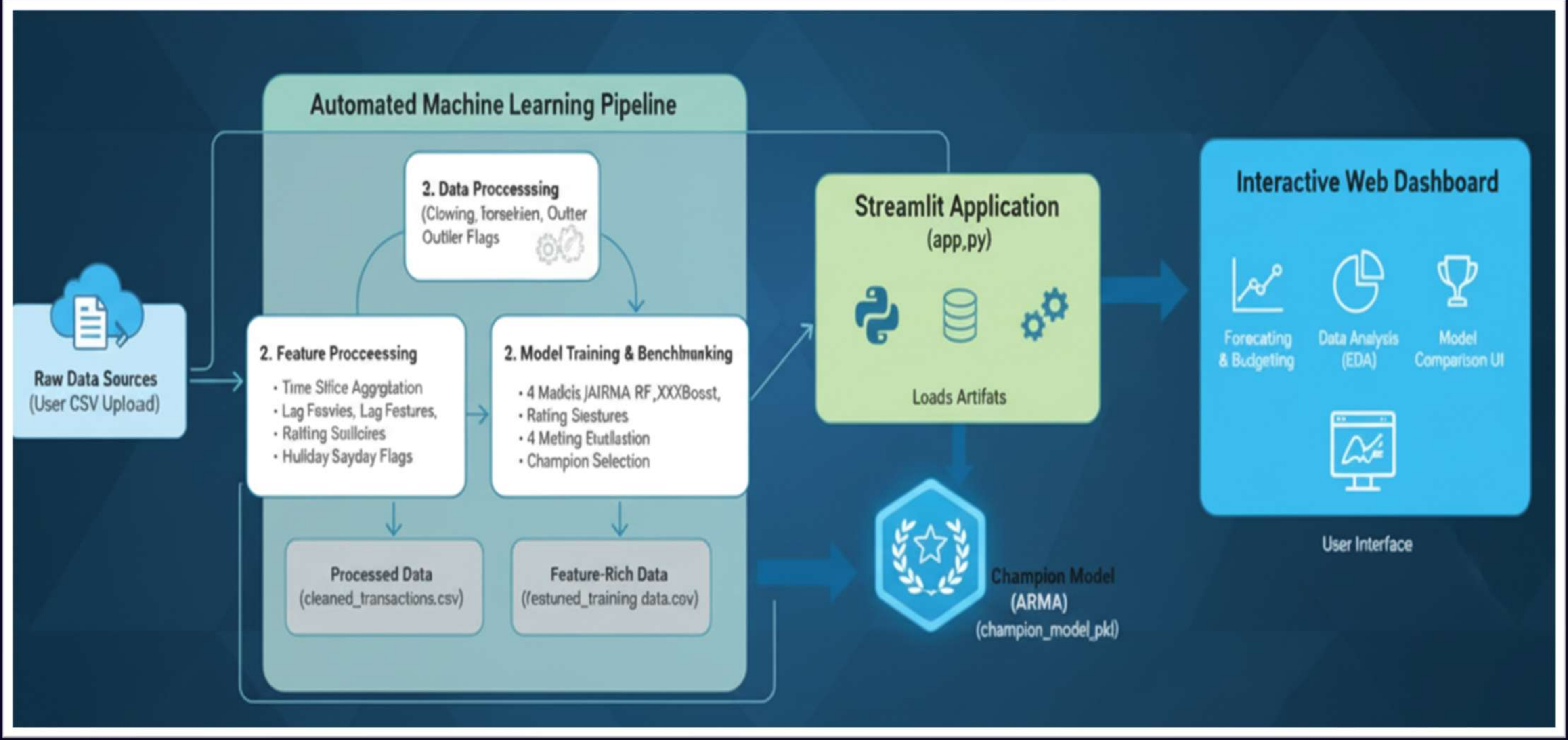
## Automated Machine Learning Pipeline

Professional Python scripts transform messy transaction data into clean, feature-rich datasets. The system automatically benchmarks multiple ML models, identifying the champion for accurate predictions.

## Interactive Streamlit Dashboard

A clean, intuitive web application delivers three core capabilities: personalised expense forecasts, deep spending habit analysis, and actionable budget recommendations aligned with savings goals.

# System Architecture: Separation of Concerns



System Architecture

# The Data Science Pipeline: Three-Stage Automation

### Stage 1: Data Preprocessing

`01_preprocess_data.py`

Deep cleaning transforms raw, inconsistent data into a reliable foundation. The script standardises date formats, removes currency symbols, normalises categories and adds quality flags like `is_outlier` to create a complete audit trail.

### Stage 2: Feature Engineering

`02_create_features.py`

Aggregates cleaned data into daily time series and constructs over 10 predictive features. These include lagged values, rolling statistics, and advanced indicators like holiday flags and payday detection—critical for model accuracy.

### Stage 3: Model Training

`03_train_models.py`

Runs the complete modelling gauntlet. Loads feature-rich data, trains four candidate algorithms, evaluates performance across multiple metrics, and automatically identifies and saves the champion model. Full automation of the research process.

This modular, scripted approach ensures reproducibility, simplifies debugging, and represents professional-grade data science engineering rather than ad-hoc notebook exploration.

# Model Benchmarking: Finding the Champion

We conducted a comprehensive evaluation of four models from different algorithm families, measuring performance across multiple dimensions to identify the optimal solution.

| Model | MAE (₹) | RMSE (₹) | MAPE (%) | Direction (%) |
|---|---|---|---|---|
| ARIMA | 34,127 | 44,893 | 50.2 | 63.8 |
| Random Forest | 47,892 | 63,241 | 68.4 | **69.5** |
| XGBoost | 49,156 | 64,873 | 70.1 | 67.2 |
| Linear Regression | 52,334 | 68,912 | 75.3 | 61.4 |

## ARIMA: The Clear Winner

ARIMA achieved a Mean Absolute Error of just ₹34,127—dramatically outperforming tree-based models. For financial applications, accuracy in actual currency values matters most, making MAE our primary decision metric.

Whilst Random Forest showed superior directional accuracy (69.5%), predicting whether spending would increase or decrease, ARIMA's precision in forecasting exact amounts makes it the champion for budgeting applications.

34K

Best MAE

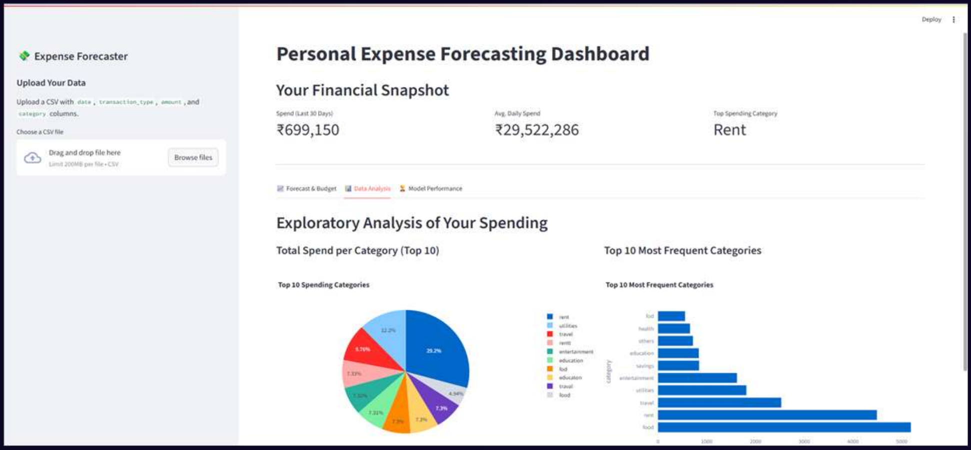
Average error in rupees

44%

Lower Error

vs. nearest competitor

# Results:



Expense Forecaster

Upload Your Data

Upload a CSV with `date`, `transaction_type`, `amount`, and `category` columns.

Choose a CSV file

Drag and drop file here
Limit 200MB per file • CSV

## Personal Expense Forecasting Dashboard

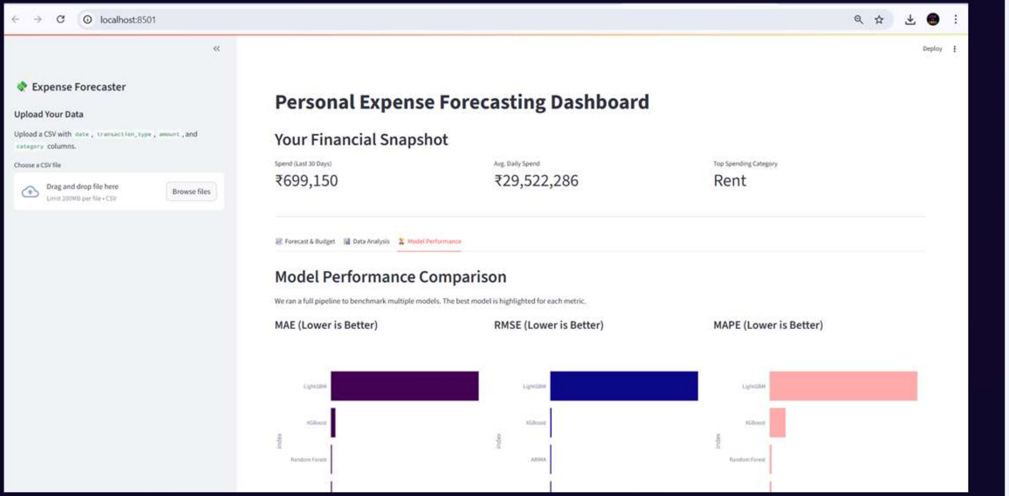
### Your Financial Snapshot

| Spend (Last 30 Days) | Avg. Daily Spend | Top Spending Category |
|---|---|---|
| ₹699,150 | ₹29,522,286 | Rent |

### Exploratory Analysis of Your Spending

Total Spend per Category (Top 10)

Top 10 Most Frequent Categories



localhost:8501

Expense Forecaster

Upload Your Data

Upload a CSV with `date`, `transaction_type`, `amount`, and `category` columns.

Choose a CSV file

Drag and drop file here
Limit 200MB per file • CSV

## Personal Expense Forecasting Dashboard

### Your Financial Snapshot

| Spend (Last 30 Days) | Avg. Daily Spend | Top Spending Category |
|---|---|---|
| ₹699,150 | ₹29,522,286 | Rent |

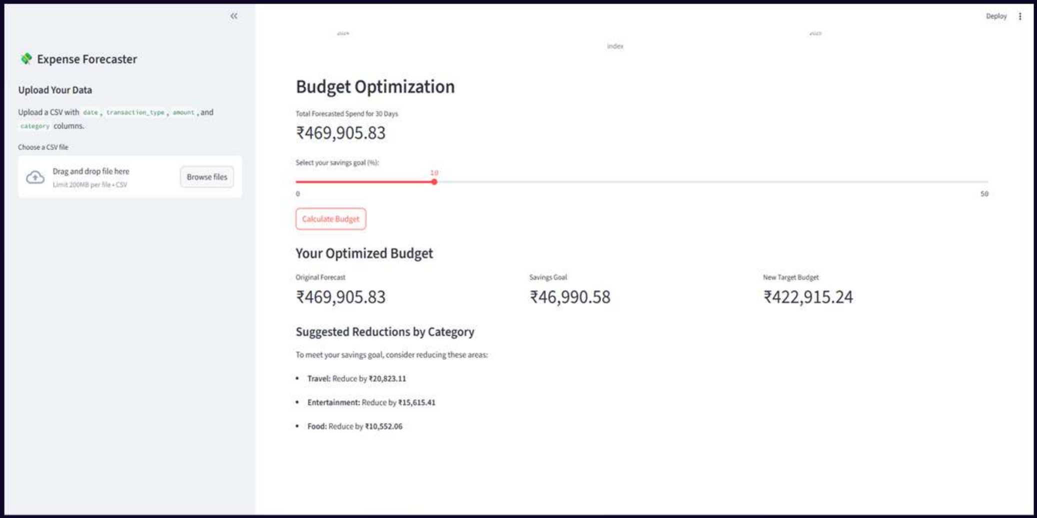
### Model Performance Comparison

We ran a full pipeline to benchmark multiple models. The best model is highlighted for each metric.

MAE (Lower is Better)

RMSE (Lower is Better)

MAPE (Lower is Better)



Expense Forecaster

Upload Your Data

Upload a CSV with `date`, `transaction_type`, `amount`, and `category` columns.

Choose a CSV file

Drag and drop file here
Limit 200MB per file • CSV

### Budget Optimization

Total Forecasted Spend for 30 Days

₹469,905.83

Select your savings goal (%): 10

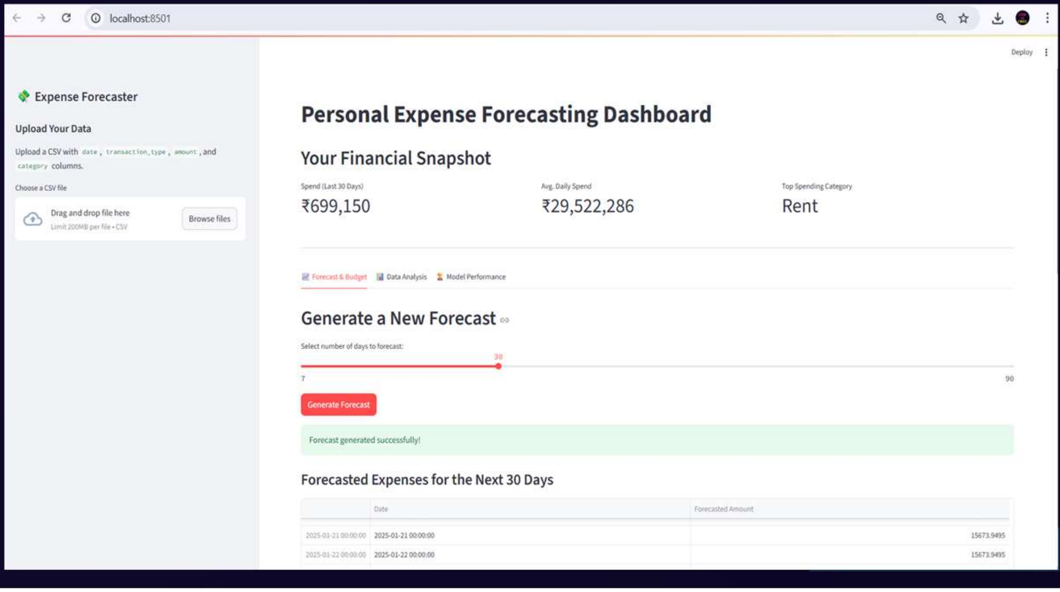
### Your Optimized Budget

| Original Forecast | Savings Goal | New Target Budget |
|---|---|---|
| ₹469,905.83 | ₹46,990.58 | ₹422,915.24 |

#### Suggested Reductions by Category

To meet your savings goal, consider reducing these areas:

- **Travel:** Reduce by **₹20,823.11**
- **Entertainment:** Reduce by **₹15,615.41**
- **Food:** Reduce by **₹10,552.06**



localhost:8501

Expense Forecaster

Upload Your Data

Upload a CSV with `date`, `transaction_type`, `amount`, and `category` columns.

Choose a CSV file

Drag and drop file here
Limit 200MB per file • CSV

## Personal Expense Forecasting Dashboard

### Your Financial Snapshot

| Spend (Last 30 Days) | Avg. Daily Spend | Top Spending Category |
|---|---|---|
| ₹699,150 | ₹29,522,286 | Rent |

### Generate a New Forecast

Select number of days to forecast: 30

Forecast generated successfully!

### Forecasted Expenses for the Next 30 Days

| | Date | Forecasted Amount |
|---|---|---|
| 2025-01-21 00:00:00 | 2025-01-21 00:00:00 | 15673.9495 |
| 2025-01-22 00:00:00 | 2025-01-22 00:00:00 | 15673.9495 |

# Key Features & Capabilities

### Budget Optimization Engine

Users set savings goals with an intuitive slider, receiving concrete, actionable advice on where to reduce spending to meet financial targets effectively.

### Dynamic File Upload

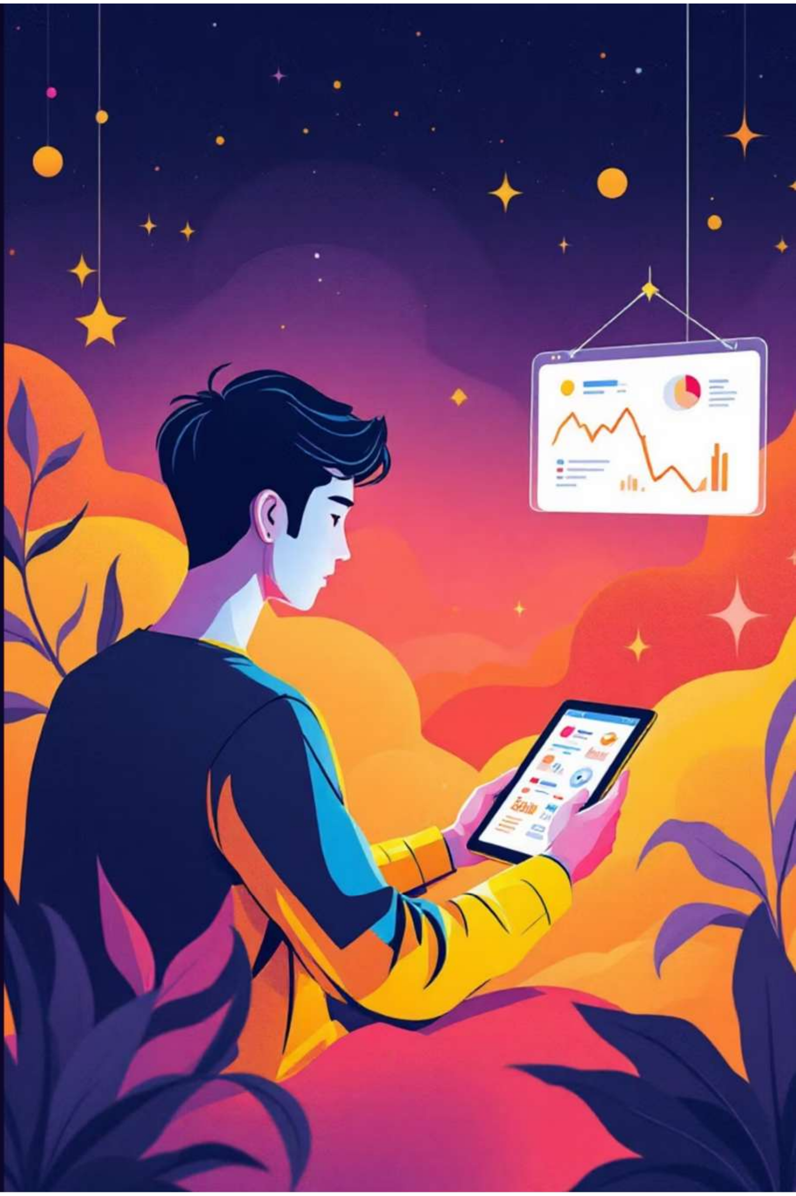
Support for personalized analysis through the sidebar, allowing any user to upload their own transaction data and receive customized forecasts instantly.

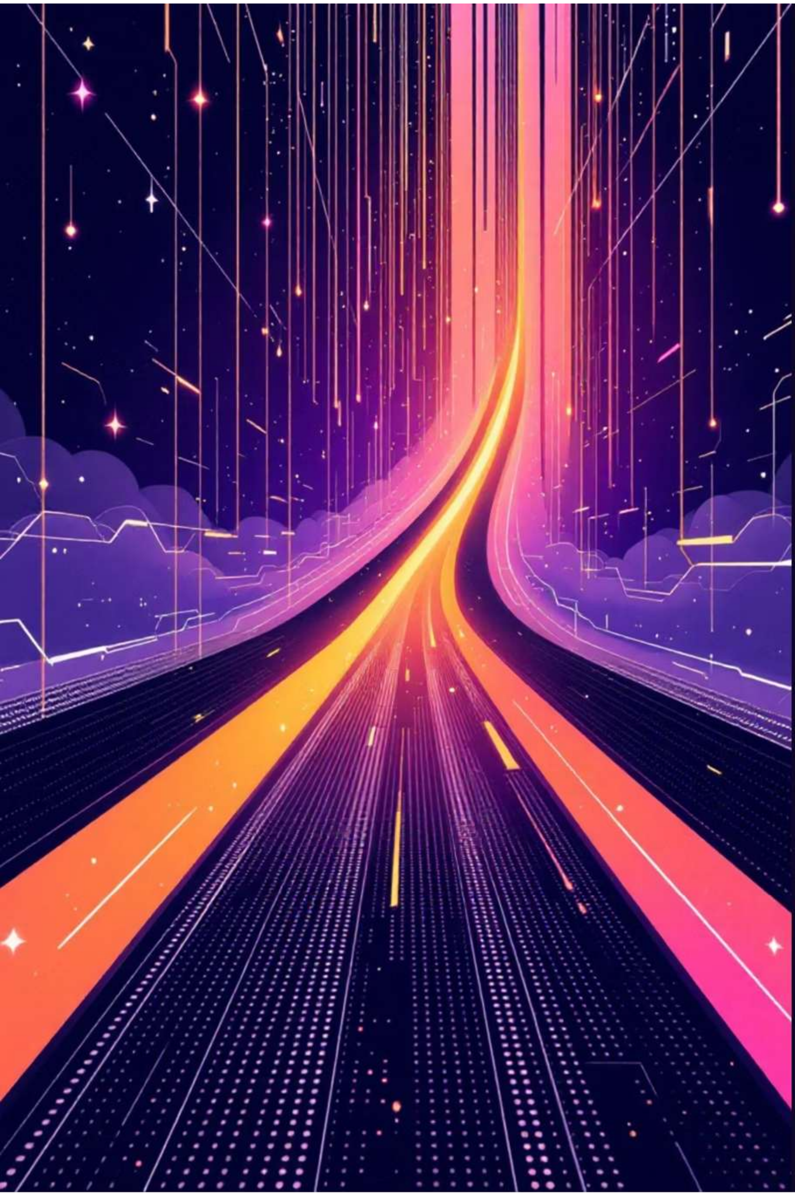
### Interactive Visualizations

Powered by Plotly for rich interactivity — zoom, hover for details, and explore data dynamically to gain deeper insights into spending patterns.

### Model Performance Transparency

Dedicated tab showcasing benchmark results, visually explaining why the ARIMA model was selected as champion, building user trust through transparency.





# Conclusion & Future Roadmap

We've successfully designed and built a complete end-to-end data science product, transforming raw, messy transaction data into a polished, interactive web application driven by robust, automated machine learning.

## What We've Achieved

- Fully automated, three-stage data pipeline
- Rigorous benchmarking across four model families
- Champion ARIMA model with ₹34K MAE
- Professional Streamlit web application
- Actionable budget optimization engine
- Complete project documentation

## Future Enhancements

1 Cloud Deployment

Make the application publicly accessible via AWS, Azure, or Google Cloud

2 Hyperparameter Tuning

Implement automated optimization to further refine champion model performance

3 External Data Integration

Incorporate inflation rates, economic indicators to boost prediction accuracy