Forecasting Farmer Contributions to PMFBY (Pradhan Mantri Fasal Bima Yojana) Using Historical Data

Submitted by
Jyothy Das, Swaraj Gopal, and Vishnu Premji

25/01/2025

A project report submitted to ICT Academy of Kerala
in partial fulfillment of the requirements
for the certification of

CERTIFIED SPECIALIST IN DATA SCIENCE & ANALYTICS

Contents

List of Figures

List of Tables

1 Abstract

This research aims to develop a predictive model for farmers' contributions to the Pradhan Mantri Fasal Bima Yojana (PMFBY), an agricultural insurance scheme in India. Utilizing historical data on crop insurance coverage, the study will identify patterns and trends in farmer contributions, considering factors like crop type, state policies, and premium rates. The analysis will inform the optimization of government subsidies to maximize coverage while ensuring equitable access. Furthermore, the research will enhance transparency and informed decision-making in policy design. The final model will be deployed as an interactive web application using Flask, enabling stakeholders to explore predictions and their implications dynamically.

2 Problem Definition

2.1 Overview

The Pradhan Mantri Fasal Bima Yojana (PMFBY) is a flagship agricultural insurance scheme introduced by the Government of India to protect farmers from crop losses due to natural calamities. While the scheme has positively impacted agricultural risk management, several challenges remain, particularly in understanding and optimizing farmers' contributions. With participation varying widely across regions and crop types, policymakers face difficulties in ensuring balanced subsidy allocation and equitable access for farmers. This research leverages data-driven methods to explore patterns in farmers' contributions and to provide actionable insights for improving the scheme's efficiency. By analyzing historical data on crop insurance coverage, the study aims to uncover key trends and dependencies, which will aid in designing policies that better align with farmers' financial capacity and agricultural risk profiles. Through the development of a predictive model, this research seeks to bridge the gap between data insights and practical policy-making for PMFBY.

2.2 Problem statement

This project aims to predict the farmers' contribution to PMFBY using historical data on crop insurance coverage. The outcome of this analysis will

provide insights into patterns in farmers' contributions and help optimize subsidy allocation policies.

3 Introduction

Agriculture is the backbone of India's economy, supporting nearly half of the nation's population. However, farmers in India face significant risks due to unpredictable weather patterns, natural calamities, and market fluctuations, which threaten both their livelihoods and national food security. To address these challenges, the Government of India introduced the Pradhan Mantri Fasal Bima Yojana (PMFBY) in 2016, an ambitious crop insurance scheme aimed at protecting farmers from financial losses caused by crop failures. Under this scheme, farmers contribute a portion of the insurance premium, with the remainder subsidized by the government, ensuring affordability and accessibility.

While the PMFBY has brought relief to many farmers, its implementation has faced several challenges. Participation rates vary widely across states, crop types, and farming communities, influenced by factors such as economic constraints, varying premium rates, and regional policy frameworks. These disparities have created barriers to equitable access, limiting the scheme's potential to provide widespread protection. Furthermore, the government faces challenges in optimizing subsidy allocations to ensure effective coverage without overburdening public finances.
Given the scale and complexity of PMFBY, there is a pressing need for data-driven solutions to enhance its efficiency and transparency. By analyzing historical data on crop insurance coverage, this research aims to identify key patterns and trends in farmers' contributions to the scheme. These insights will help policymakers design targeted interventions to improve participation and optimize subsidy distribution.

The primary objective of this study is to develop a predictive model for farmers' contributions under PMFBY. The model will take into account factors such as crop type, regional policies, and premium rates to forecast future trends. The ultimate goal is to support policymakers in making informed decisions that balance the scheme's financial sustainability with equitable access for farmers. Additionally, the research will culminate in the development of an interactive web application using Flask, enabling stakeholders to explore

predictions dynamically and understand their policy implications.

This study not only contributes to the optimization of PMFBY but also demonstrates the potential of data analytics in addressing critical challenges in agriculture and policymaking.

4 Literature Survey

The literature surrounding the Pradhan Mantri Fasal Bima Yojana (PMFBY) and related crop insurance schemes highlights the critical need for effective risk mitigation strategies in the Indian agricultural sector. PMFBY, launched in 2016, aims to provide financial support to farmers facing crop loss due to unforeseen events such as droughts, floods, and pests. By offering subsidized insurance premiums, the scheme has played a pivotal role in stabilizing farmers' incomes and reducing their vulnerability to climatic and market uncertainties.

4.1 Key Areas of Study

- Insurance Premium Models and Farmer Participation: Research on crop insurance schemes has consistently emphasized the importance of premium rates in influencing farmers' participation. For PMFBY, farmers contribute a portion of the sum insured, depending on the crop type and season. Several studies have explored the relationship between these premiums, government subsidies, and regional agricultural practices to understand their impact on coverage and equity.

- Role of Data Analytics in Agricultural Risk Management: The increasing availability of agricultural data has led to the development of predictive models for insurance coverage and risk assessment. Techniques such as regression analysis- Random Forest, and Support Vector Machines (SVMs)- have been employed to forecast farmers' contributions and optimize subsidy distribution. These models utilize diverse features, including crop type, state policies, indemnity levels, and sum insured values, to predict trends and enhance decision-making.

- Challenges in Crop Insurance Implementation: Despite its benefits, PMFBY has faced implementation challenges, such as low participation in certain regions and inequities in subsidy allocation. Research

has also identified issues related to skewed data distributions, multicollinearity among features, and missing or irrelevant data. Addressing these challenges requires advanced preprocessing techniques, feature engineering, and robust validation methods like k-fold cross-validation.

- Technology and Process Innovations: The operational guidelines for PMFBY advocate the use of modern technologies like drones, satellites, and geospatial mapping to enhance accuracy in damage assessment and claim processing. These innovations also contribute to better premium rate calculations and the identification of high-risk areas, enabling targeted interventions.

- Alternative Problem Formulations: Studies have proposed alternative objectives within the PMFBY framework, such as classifying high-risk states, forecasting optimal premium rates, and analyzing the impact of indemnity levels on claim payouts. These approaches highlight the flexibility and potential of data-driven insights in improving scheme implementation.

4.2 Key Findings and Gaps

While existing literature provides a strong foundation for understanding PMFBY's mechanisms, gaps remain in creating accessible tools for stakeholders to explore predictions and policy implications. Few studies have focused on integrating predictive models into interactive platforms, which could significantly enhance transparency and engagement among policymakers, insurers, and farmers.

4.3 Relevance to Current Study

This research aims to address these gaps by developing a predictive model for farmers' contributions to PMFBY. By leveraging historical data on crop insurance coverage and deploying the model through a Flask-based web application, the study seeks to provide actionable insights for subsidy optimization and equitable policy design. Furthermore, the use of advanced machine learning techniques and robust evaluation metrics ensures the reliability and applicability of the findings to real-world scenarios.

5 Methodology

5.1 Data Collection

The dataset used in this study was sourced from Kaggle, a popular platform for open-source datasets. It contains historical records related to farmer contributions under the Pradhan Mantri Fasal Bima Yojana (PMFBY), including details on premium rates, crop types, insured areas, and state-wise participation. After an extensive search, no prior studies or projects were found using this specific dataset, making this research a novel contribution to understanding farmer participation trends in PMFBY. The dataset was carefully examined for completeness and consistency before proceeding with further analysis.

5.2 Data preprocessing & EDA(Exploratory Data Analysis)

5.2.1 Data Overview

The dataset provides comprehensive information on the agricultural insurance scheme, covering details about policies, premiums, shares paid by various stakeholders, and other key attributes. The dataset provides insights into farmers' enrollment and insurance coverage under two key agricultural insurance schemes in India: the Pradhan Mantri Fasal Bima Yojana (PMFBY) and the Weather-Based Crop Insurance Scheme (WBCIS). The data spans from the inception of PMFBY in 2018 up to the Rabi 2022-23 season and includes district-wise statistics on enrollment, coverage, and financial details.

- Scheme and seasonal details
  - sssyName.Year : Represents the year in which the farmer joined the scheme, combining State, Season, Scheme, and Year identifiers.
  - sssyName.seasonName : Denotes the season (e.g., Kharif, Rabi) during which the scheme was active.
  - sssyName.schemeName : Specifies the name of the insurance scheme under which the farmer is enrolled.

  - sssyName.stateName : Indicates the state in which the farmer is enrolled in the scheme.

- Regional and Crop Information
  - level3Name : Refers to the district or local administrative region involved in the insurance scheme.
  - cropName : Indicates the type of crop insured under the scheme.

- Financial Details
  - sumInsured : The total value of the crop insured against potential losses. This serves as the basis for calculating premium amounts.
  - premiumRate : The rate at which the insurance premium is calculated as a percentage of the sum insured.
  - stateShare : The percentage of the sum insured that the state government contributes towards the premium.
  - goiShare : The percentage of the sum insured that the Government of India contributes towards the premium.
  - farmerShare : The percentage of the sum insured that the farmer needs to pay as their share of the premium.
  - farmerShareValue : The monetary value of the premium amount paid by the farmer.
  - goiShareValue : The monetary value of the premium amount contributed by the Government of India.
  - stateShareValue : The monetary value of the premium amount contributed by the state government.

- Policy and Administrative Details
  - sssyID : A unique identifier for the State-Season-Scheme-Year combination.
  - seasonID and seasonCode : Unique identifiers and codes representing specific seasons.
  - schemeID and schemeCode : Unique identifiers and codes for the insurance scheme.

  - stateID and stateCode : Unique identifiers and codes for states.
  - level3ID and level3Code : Unique identifiers and codes for districts.
  - cropID and cropCode : Unique identifiers and codes for crops.

- Temporal Information
  - Year : The specific year when the scheme or policy was active..
  - startDate and endDate : The start and end dates of the scheme or policy period.
  - policyStartDate and policyEndDate : The start and end dates of the insurance policy coverage period.
  - yieldEndDate : The date on which the yield estimation for the insured crop concludes

- Indemnity Details
  - IndemnityLevel : Indemnity Level refers to the percentage of the sum insured that farmers are eligible to receive as compensation in case of crop loss due to insured perils. Specifies the percentage of the sum insured that farmers are eligible to receive as compensation in case of crop loss.

    In the PMGBY dataset there are only 3 IndemnityLevels; 70%, 80%, and 90%.
  - thresholdYield : The benchmark yield level linked to the indemnity level, below which compensation is provided.

    Farmers are eligible for compensation when the actual yield falls below the Threshold Yield. The amount of compensation is determined based on the extent of the yield shortfall and the applicable indemnity level.

- Insurance Company Information
  - insuranceCompanyName : The name of the insurance company providing coverage under the scheme.
  - insuranceCompany.insuranceCompanyID and insuranceCompany.insuranceCompanyCo : Unique identifiers and codes for the insurance company.

  - headQuaterAddress, headQuaterEmail, and tollFreeNumber : Contact details of the insurance company's headquarters.
  - websiteLink : Official website link of the insurance company for additional information.

- Policy Details
  - isOpen : Indicates whether the scheme is still active or closed.
  - cnStarted : Denotes the year in which the scheme commenced.
  - isPreviousSeasonYearInSubsidy : A Yes/No field indicating whether subsidies for the previous season/year are included.
  - Policy : Details of the specific insurance policy issued.
  - Notification : Notifications associated with the policy or scheme.
  - isOfflineChallan : Indicates whether the payment or subsidy process is done offline.
  - firstGoiSubsidy : Represents the first subsidy amount released by the Government of India.
  - goiOfflineChallan and stateOfflineChallan : Specific details about offline payments made by the Government of India and state governments.

- Miscellaneous Details
  - categoryName: Represents different categories (14 in total) to which crops or policies belong.
  - cropType: The type of crop insured (e.g., Kharif or Rabi crops).
  - Unit: Represents the insurance coverage unit. Most rows have a value of 1, with only 2 rows having a value of 2.
  - pickingType: Represents specific collection or grouping mechanisms within the dataset.
  - type: Classification or categorization associated with the data record.

5.2.2 Data Characteristics

Size: The dataset includes district-level enrollment and coverage data, which covers a large number of records across various districts, crops, and states. The data spans several years, beginning from 2018, and continues up to the Rabi 2022-23 season. There are 29999 rows and 62 columns. Thus, the dataset implies data records of 29999 incidents w.r.t 62 features which are taken.

Time span: The data covers the period from the introduction of the PMFBY scheme in 2018 up to February 2023, ensuring a wide temporal range for analysis. This includes multiple agricultural seasons (Kharif, Rabi), allowing for seasonal comparisons.

Data types:

- Integer (int64): Used for categorical variables like IDs (e.g., sssyID, seasonCode, stateCode) and numerical values (e.g., year, schemeCode).
- Float (float64): Used for numerical variables such as sumInsured, premiumRate, stateShare, and goiShare.
- Object (object): Used for textual data, including categorical variables like seasonName, stateName, cropName, and insurance company details.
- Boolean (bool): Used for flags, such as whether a record has specific characteristics (e.g., isOfflineChallan).

Non-null Counts: All columns in the dataset have 29999 non-null entries, indicating that the dataset does not have missing values for the variables in question. This suggests that the data has been adequately prepared and cleaned for analysis.

Handling Missing Values: There are no missing values, as all 29,999 rows are complete.

5.3 Data Transformation

5.3.1 Data Synchronization:

Efforts have been made to synchronize the state and district names throughout the dataset, ensuring a uniform format.

5.3.2 Column Removal

As part of the data preprocessing, some columns were renamed for clarity and consistency, while irrelevant or redundant features were removed to streamline the dataset for analysis.

To streamline the dataset and focus on relevant variables for the analysis, several columns were removed, including those related to administrative IDs, season names, and insurance company details. This transformation resulted in a reduced set of features that better align with the goals of the project.

The following feature columns were dropped:

- sssyName.seasonName
- sssyName.schemeName
- seasonID
- schemeID
- schemeCode
- level3Name
- stateID
- stateCode
- level3ID
- level3
- level3
- Code
- cropName
- cropID
- cropCode
- pickingType
- sssyID
- year
- policyStartDate
- policyEndDate
- isOfflineChallan
- goiOfflineChallan

- stateOfflineChallan
- yieldEndDate
- currentTime
- default
- insuranceCompanyName
- cutOfDate
- tollFreeNumber
- headQuaterAddress
- headQuaterEmail
- websiteLink
- insuranceCompany.insuranceCompanyCode
- insuranceCompany.insuranceCompanyID
- isOpen
- cnStarted
- unit
- ayTy
- Scheme
- Start

5.3.3 Column Rename

To enhance clarity and improve readability, several columns were renamed to more intuitive and consistent names.

5.3.4

5.3.5

5.3.6

5.3.7

5.4 Model Development

Insert Model Development details here.

6 Results and Discussion

Insert Results here.

7 Conclusion

Insert Conclusion here.

8 References

Insert References here.