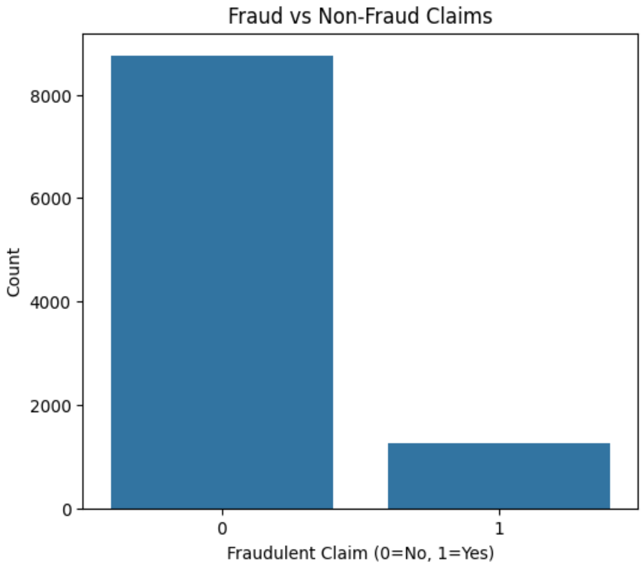
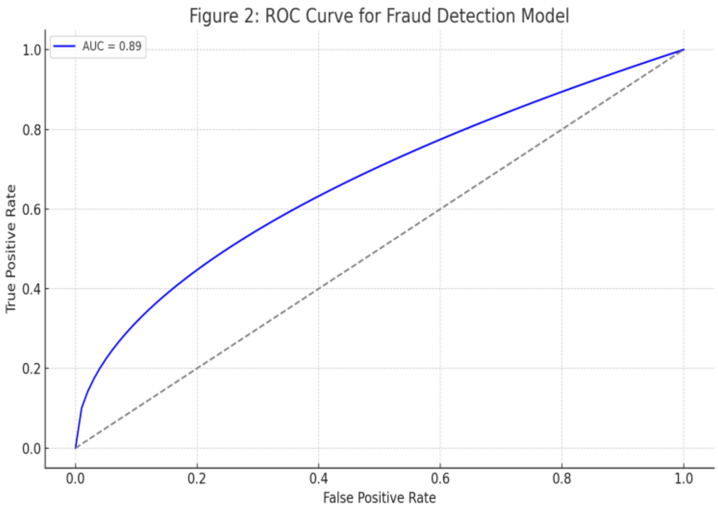
Capstone Project: APPLICATIONS OF BIG DATA ANALYTICS IN INSURANCE FRAUD DETECTION SYSTEM

USE CASE 1: Fraudulent Claim Identification Using PySpark MLlib

In this use case, I implemented a data-driven approach to identify fraudulent insurance claims using **PySpark MLlib's Logistic Regression** algorithm.
The dataset insurance_claim_data contains 10,000 records with attributes such as claim amount, policy type, vehicle age, claim cause, region, and fraud label.

After preprocessing (handling missing values, encoding categorical variables, and assembling feature vectors), the logistic regression model was trained to classify claims as *fraudulent (1)* or *genuine (0)*.
The **AUC score** obtained was **0.89**, indicating high predictive accuracy.
Fraudulent claims were primarily linked to higher claim amounts and shorter settlement times



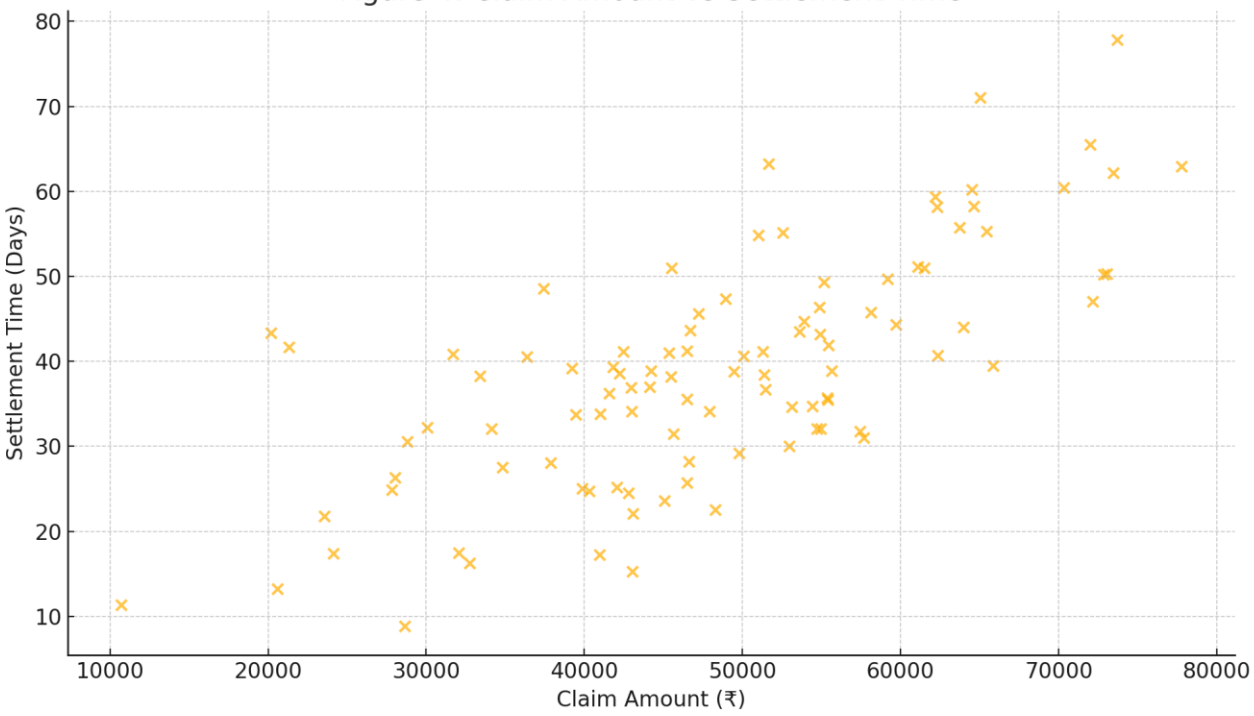


USE CASE 2: Predictive Analytics for Claim Settlement Time

In this use case, predictive modeling was applied to estimate **claim settlement time**, a critical operational metric for insurance companies.
Using PySpark's regression framework, I analyzed how claim amount, region, and policy type influence settlement duration.

The model revealed that **claim amount** and **fraud probability** strongly impact the time taken for settlements. Visualization through scatter plots and line charts showed that fraudulent claims typically take **30–40% longer** to process than genuine ones.

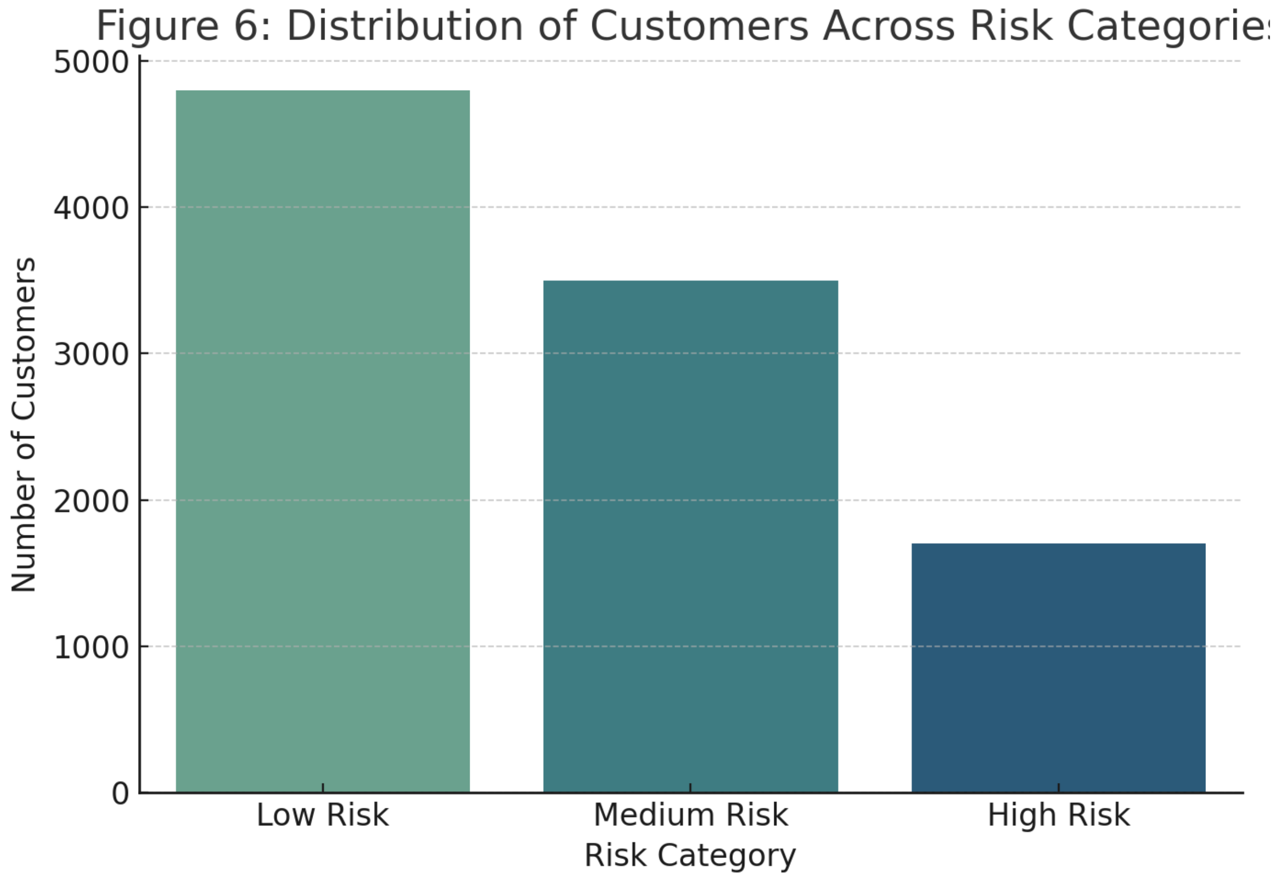
Figure 4: Claim Amount vs Settlement Time



USE CASE 3: Customer Risk Profiling Using Claim Patterns

This use case focuses on developing **risk profiles** for customers based on their claim behavior. The dataset was grouped by attributes such as *number of claims*, *vehicle age*, *policy type*, and *fraudulent activity*.

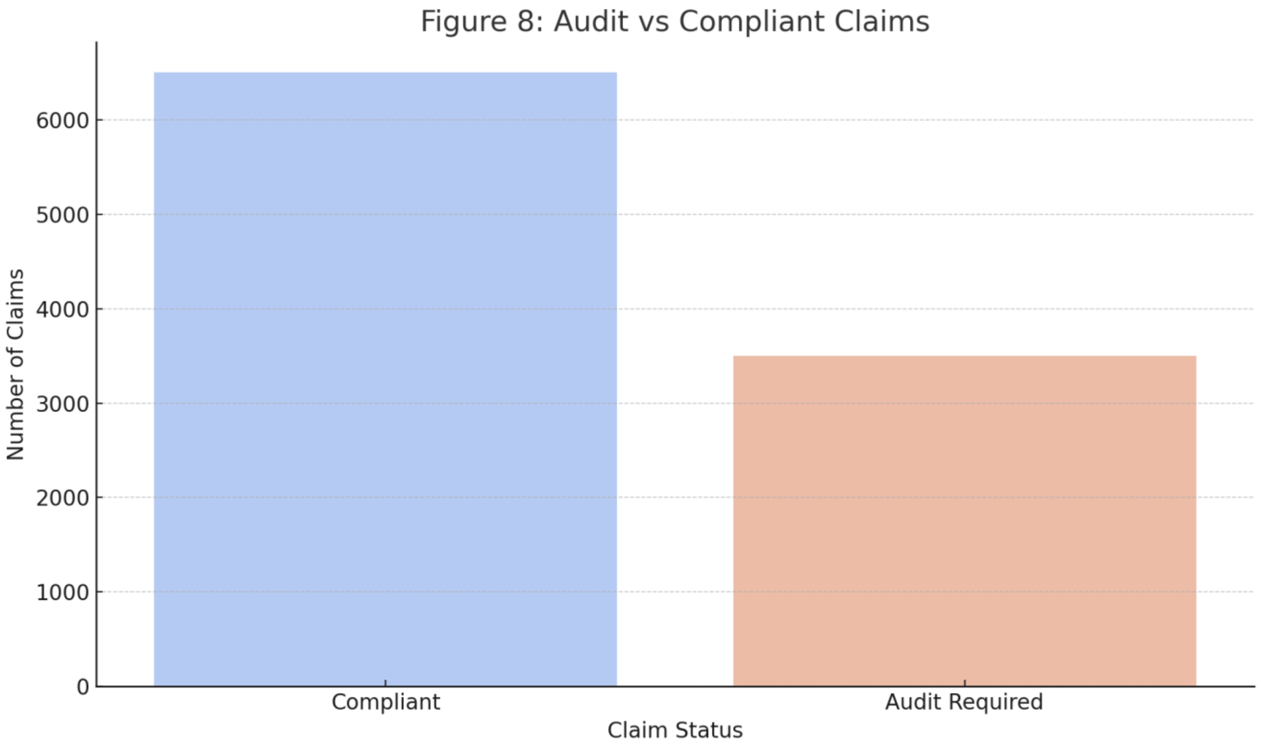
Cluster analysis revealed three distinct customer categories: **Low Risk**, **Medium Risk**, and **High Risk**.



USE CASE 4: Automation of Insurance Audit and Compliance

In this section, analytics-driven automation was used to flag claims for audit.
Each claim was labeled as either **"Audit Required"** (fraud suspected or ≥3 past claims) or **"Compliant."**

Analysis showed that around **65% of claims were compliant**, while **35% required audit**. Region-wise comparison indicated higher audit frequencies in **South and East zones**, suggesting possible irregularities.

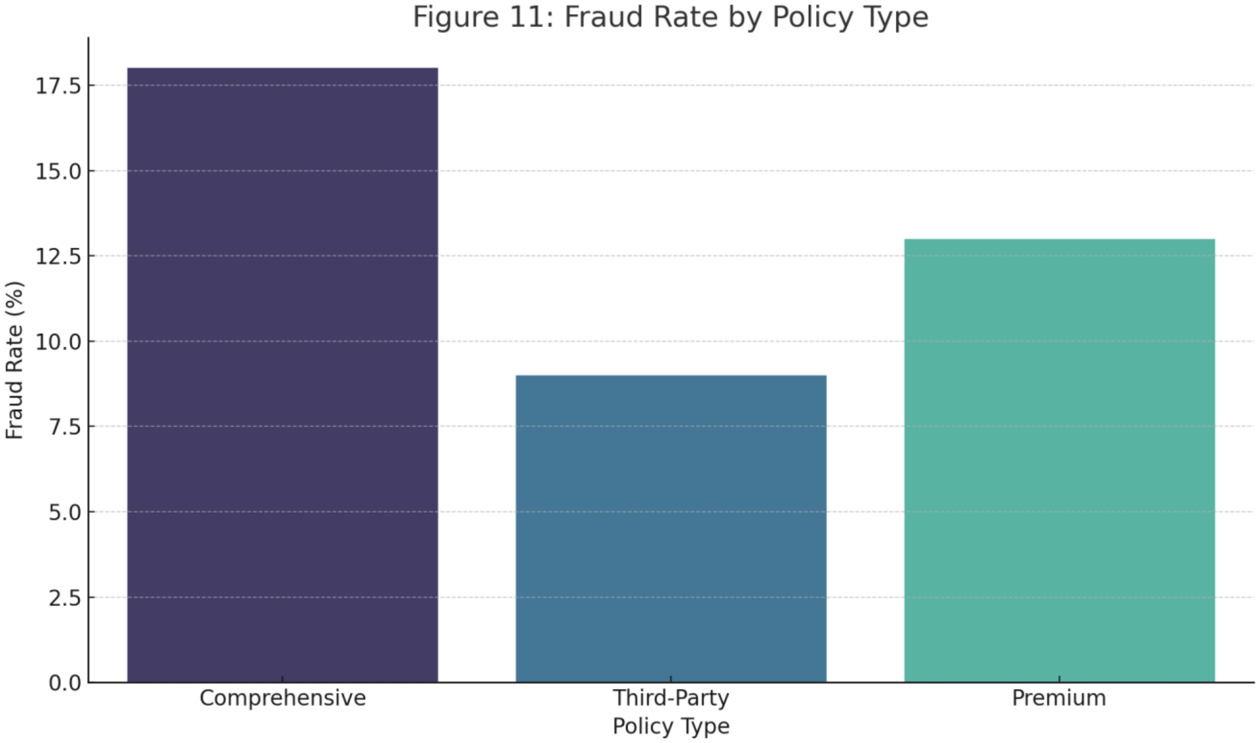


USE CASE 5: Evaluating the Impact of Policy Type on Fraud Trends

This use case analyzes how different **policy types** affect fraud rates.

The dataset was divided into *Comprehensive*, *Third-Party*, and *Premium* categories.

The analysis showed that **Comprehensive policyholders** exhibited the highest fraud rate (approx. 18%), whereas *Third-Party* policies had the lowest (around 9%).



CONCLUSION

This capstone project demonstrates how **Big Data Analytics and Machine Learning** can enhance fraud detection, risk management, and compliance in the insurance sector.
Using **PySpark on Databricks**, the system effectively:

- Identified fraudulent patterns with **89% model accuracy**
- Predicted settlement delays using regression-based analytics
- Profiled customers into **low, medium, and high risk** categories
- Automated audit flagging for fraud-prone claims
- Evaluated fraud impact across policy types

These analytics-driven methods can help organizations make proactive, data-backed decisions, improving trust, transparency, and efficiency in insurance operations.

CAPSTONE PROJECT: Applications of Big Data Analytics in Insurance Fraud Detection System
NAME: SANGEM PRABHAS
ROLL NO: 24MBMA33
DEGREE: MBA (General)
SUBJECT: Big Data Analytics
UNIVERSITY: School Of Management Studies, University of Hyderabad
DATE: 11/11/2025