CREDIT RISK PREDICTION

By Koyalkar Chetan(24MBMB20)

Abstract:

This project focuses on predicting the likelihood of credit card default using machine learning models built in Databricks with PySpark. It applies the Medallion architecture to process large-scale financial data efficiently and generate actionable insights for risk management.

PROBLEM STATEMENT:

Financial institutions struggle to identify customers who may default on payments, leading to financial losses.

OBJECTIVES/USE CASES:

- Predict credit default probability.
- Segment customers based on repayment patterns.
- Feature Importance and Explainability
- Generate alerts for high-risk customers.
- Recommend optimal credit limits.

METHODOLOGY

Databricks Medallion (Bronze → Silver → Gold)

Layer	Function	Tools
Bronze	Ingest raw CSV data	PySpark
Silver	Clean & transform data	Feature engineering, scaling
Gold	ML modeling & analytics	PySpark MLlib

ML Models Used:

- Logistic Regression
- Decision Tree
- Random Forest
- PCA
- K-Means Clustering

RESULTS & INSIGHTS

Gboost gave the best AUC score (~ 0.766).

Top influencing features: PAY_0(the repayment status in September, 2005), BILL_AMT1(Amount of bill statement), LIMIT_BAL(Amount of the given credit).

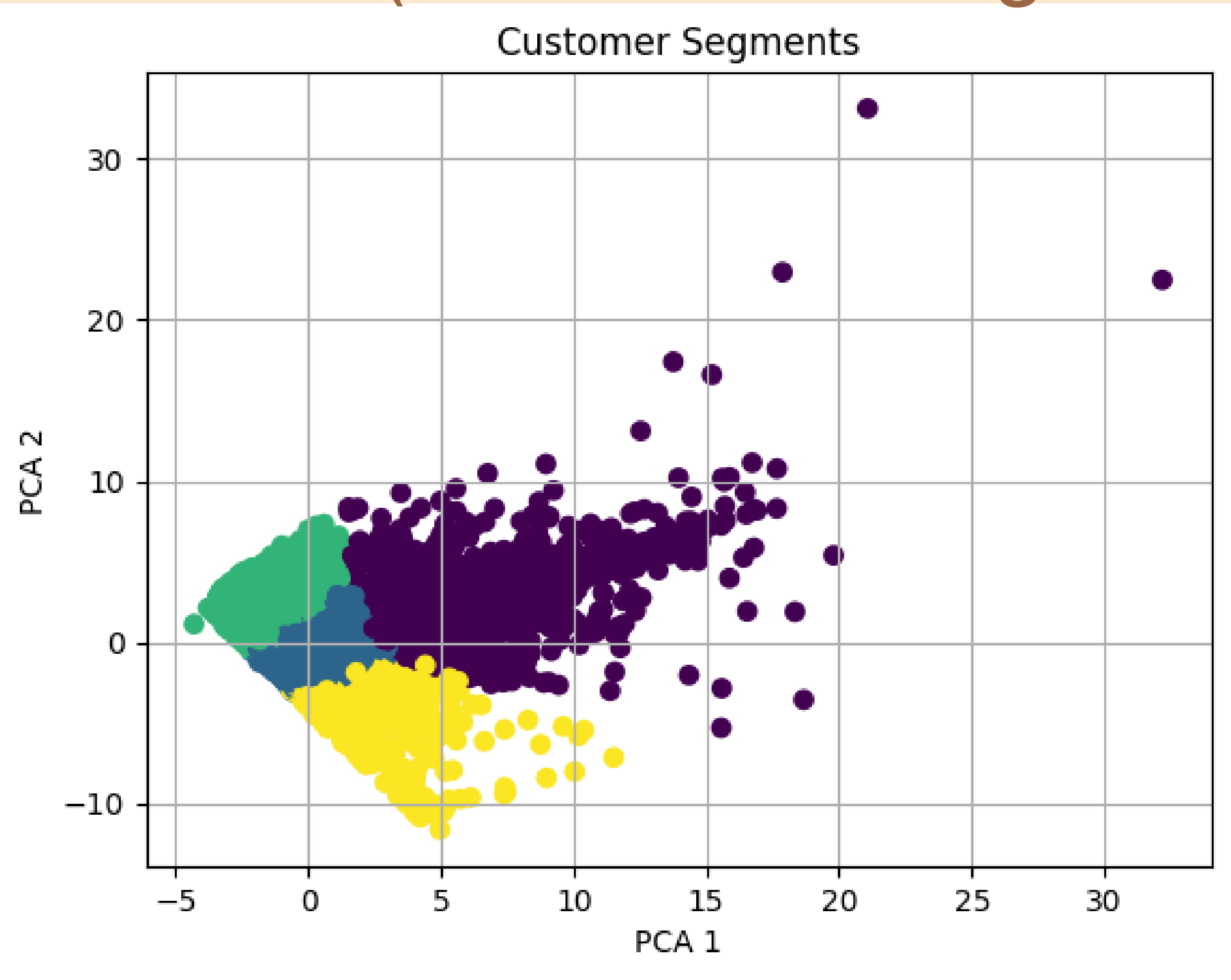
K-Means created 4 customer segments:

Reliable low-risk

Medium spenders

Risky defaulters

High-value customers



CONCLUSION & FUTURE WORK

The project successfully demonstrates an end-to-end big data ML pipelir risk prediction using Databricks.

Future Enhancements:

Real-time streaming using Spark Structured Streaming
Integration with Kafka for live updates
Deep learning models for improved accuracy
API integration for financial dashboards

THANK YOU