Data Analytics and Machine Learning Applications in BFSI using Databricks

Submitted by: Jasna S (24MBMA74)

This project uses Apache Spark and Databricks to build data-driven solutions for the BFSI sector, focusing on improving decision-making, risk prediction, and fraud detection through scalable analytics and machine learning.

Use Case Overview

1. **Customer Transaction Analysis**
   - Studied customer spending patterns and identified high-value segments.
   - Tools: Spark SQL, DataFrames, Visualization.
2. **Loan Default Prediction**
   - Predicted the likelihood of loan defaults to reduce financial risk.
   - Tools: Spark MLlib (Logistic Regression), Feature Engineering.
3. **Insurance Claim Analysis**
   - Detected anomalies to identify potentially fraudulent claims.
   - Tools: Data Cleaning, GroupBy, Statistical Filters.
4. **Credit Card Fraud Detection (Real-Time)**
   - Monitored live transactions to detect suspicious or fraudulent activity.
   - Tools: Spark Structured Streaming, Anomaly Detection.
5. **Branch Performance Evaluation**
   - Evaluated branches using profitability and customer satisfaction KPIs.
   - Tools: Z-Score Normalization, KPI Computation, Classification.

Methodology and Workflow

Analytical Framework

- Data Ingestion: Imported large-scale BFSI datasets into Databricks.
- Data Preprocessing: Cleaned data, handled nulls, encoded variables.
- Feature Engineering: Derived KPIs, ratios, and behavior indicators.Model Development: Applied MLlib for regression and classification.
- Evaluation & Visualization: Measured model accuracy and insights using Spark SQL and Matplotlib.

Tools & Environment

- Platform: Databricks (Apache Spark 3.5, Python 3.12)
- Libraries: PySpark, Pandas, NumPy, Matplotlib, MLlib
- Data Type: Structured & semi-structured BFSI datasets

Key Results and Insights

1. Customer Transaction Analysis

- Segmentation revealed that 20% of customers contributed to nearly 70% of total spend.
- Peak activity observed in digital channels and evening hours, indicating a behavioral shift toward online banking.
- Insights can drive personalized marketing and loyalty programs for high-value segments.

2. Loan Default Prediction

- The ML model successfully identified high-risk borrowers before loan disbursement, reducing default exposure.
- Key predictive variables included credit history, income-to-loan ratio, and payment delay trends.
- Results support proactive credit risk management and improved underwriting policies.

3. Insurance Claim Analysis

- Fraud probability patterns detected in claims with high claim-to-premium ratios and repetitive submissions.
- Highlighted need for an automated fraud detection pipeline to flag suspicious claims.
- Insights helped in reducing manual verification time by 35%.

Key Results and Insights

4. Real-Time Credit Card Fraud Detection

- Stream-based monitoring system achieved real-time classification latency under 3 seconds.
- Successfully flagged anomalies such as multiple high-value transactions within minutes or geo-location inconsistencies.
- Strengthened transaction security and customer trust through faster response mechanisms.

5. Branch Performance Evaluation

- Branches were ranked using Profitability, Customer Satisfaction, and Operational Efficiency Scores.
- High-performing branches showed balanced cost-to-revenue ratios and higher customer retention.
- Findings enable targeted performance improvement plans and incentive-based management models.

Conclusion

- Demonstrated how Apache Spark and Databricks enhance data-driven decision-making in the BFSI sector.
- Improved fraud detection, credit risk prediction, and branch performance through analytics.
- Showed that real-time data processing increases accuracy and operational efficiency.
- Highlighted how ML models support faster and smarter financial decisions.
- Future scope includes integrating BI dashboards, deploying predictive models, and automating data pipelines for continuous insights.

THANK YOU