

REAL TIME FRAUD DETECTION USING PYSPARK AND MACHINE LEARNING TECHNIQUES

Antony Satya Vivek Vardhan Akisetty¹, Priyank Mohan², Phanindra Kumar³, Niharika Singh⁴, Prof. (Dr) Punit Goel⁵,
& Om Goel⁶

¹Southern New Hampshire University, Manchester NH, US

²Scholar, Seattle University, Dwarka, New Delhi, India

³Kankanampati, Binghamton University, USA

⁴ABES Engineering College Ghaziabad, India

⁵Maharaja Agrasen Himalayan Garhwal University, Uttarakhand, India

⁶ABES Engineering College Ghaziabad, India

ABSTRACT

Real-time fraud detection has emerged as a critical necessity in today's digital economy, where financial transactions are increasingly conducted online. This paper explores the implementation of PySpark, a powerful big data processing framework, in conjunction with machine learning techniques to enhance the accuracy and speed of fraud detection systems. By leveraging the distributed computing capabilities of PySpark, organizations can process vast amounts of transaction data in real time, allowing for immediate detection of potentially fraudulent activities. The study analyzes various machine learning algorithms, including decision trees, logistic regression, and ensemble methods, to assess their performance in identifying fraudulent transactions. Evaluation metrics such as precision, recall, and F1-score are utilized to compare the effectiveness of these algorithms. Furthermore, the paper highlights the challenges faced in the domain, including data imbalance and the evolving nature of fraudulent schemes. Through the integration of PySpark and machine learning, this research demonstrates a scalable solution that not only improves detection rates but also reduces false positives, thereby enhancing the overall security of financial transactions. This study aims to provide insights into developing robust real-time fraud detection systems that can adapt to the dynamic landscape of online fraud.

KEYWORDS: Real-Time Fraud Detection, Pyspark, Machine Learning, Big Data, Financial Transactions, Decision Trees, Logistic Regression, Ensemble Methods

Article History

Received: 10 Nov 2022 / Revised: 12 Nov 2022 / Accepted: 18 Nov 2022
