ONLINE PAYMENTS FRAUD DETECTION USING WITH MACHINE LEARNING:

To build an application that can detect the legitimacy of the transaction in real-time and increase the security to prevent fraud.

By

(Marri yashmitha)

(Manthina raja rishika)

(Kutagula safa)

Guided by

Prof. Ms swetha raj

A Dissertation Submitted to
SRI VENKATESWARA COLLEGE OF ENGINEERING AND TECHNOLOGY, An Autonomous Institution affiliated to 'JNTU Ananthapur' in Partial Fulfilment of the Bachelor of Technology branch of ***Computer science and Engineering***

May 2024

SRI VENKATESWARA COLLEGE OF ENGINEERING AND TECHNOLOGY

R.V.S. Nagar Tirupathi Road, Andhra Pradesh– 517127

DATA QUALITY:

Creating a data quality report for an online fraud detection system using machine learning involves several steps. This report will highlight the integrity, completeness, accuracy, consistency, and validity of the data. Below is a detailed template you can follow:

1. Introduction

Objective: To evaluate the quality of data used for online fraud detection.
Scope: The report covers data integrity, completeness, accuracy, consistency, and validity.

2. Data Overview

Data Sources: List the data sources (e.g., transaction logs, user account details, third-party risk scores).
Data Collection Period: Specify the time frame of the data.
Number of Records: Total number of records in the dataset.
Key Variables: Highlight the key variables used in fraud detection (e.g., transaction amount, user ID, transaction timestamp, IP address, device ID).

3. Data Quality Dimensions

Integrity

Primary Key Violations: Check for unique constraints on primary keys (e.g., transaction ID).
Foreign Key Violations: Ensure foreign key relationships are maintained (e.g., user ID in transactions table matches user ID in users table).

Completeness

Missing Values: Report on missing values for each variable.

Number and percentage of missing values.

Critical fields with missing values (e.g., transaction amount, IP address).

Imputation Methods: Describe the methods used to handle missing values (e.g., mean imputation, deletion, interpolation).

Accuracy

Outliers Detection: Identify outliers in numerical variables (e.g., unusually high transaction amounts).

Data Entry Errors: Check for erroneous data entries (e.g., negative transaction amounts, impossible timestamps).

Verification Against External Sources: Compare data with external benchmarks (e.g., comparing transaction data with known fraud databases).

Consistency

Format Consistency:

Ensure uniform data formats (e.g., date formats, currency formats).

Duplication:

Identify and report duplicate records.

Logical Consistency:

Ensure logical relationships are maintained (e.g., transaction date should not precede account creation date).

Validity

Range Checks:

Ensure numerical values fall within acceptable ranges (e.g., transaction amounts should be positive and within expected limits).

Referential Integrity:

Verify that all references are valid (e.g., all user IDs in transactions should exist in the users table).

Domain Constraints:

Validate categorical variables against predefined domains (e.g., transaction status should be one of 'approved', 'pending', 'declined').

4. Data Quality Metrics

Completeness Rate:

Calculate the percentage of complete records.

Accuracy Rate:

Measure the rate of error-free records.

Consistency Rate:

Evaluate the consistency of data entries.

Validity Rate:

Determine the percentage of valid records.

5. Data Quality Issues and Recommendations

Identified Issues:

List and describe significant data quality issues found.

Impact Assessment:

Assess the impact of these issues on the fraud detection model.

Recommendations:

Provide recommendations to address data quality issues (e.g., improving data entry processes, implementing more rigorous validation checks).

6. Conclusion

Summary of Findings:

Summarize the key findings from the data quality analysis.

Next Steps:

Outline the next steps for improving data quality and further analysis.

7. Appendices

Appendix A: Detailed Data Quality Metrics

Appendix B: Data Dictionary

Appendix C: Methodology and Tools Used for Data Quality Assessment

This template provides a structured approach to evaluate and report on data quality for an online fraud detection system using machine learning. You can tailor the specifics based on your dataset and organizational requirements.