Capgemini

Tuesday, August 22, 2023 9:17 PM

Project Outline

#	Stage	Key Activities	Example Techniques/Tools/Methods
1	Problem Statement	Define the problem that needs to be solved.	Identify potentially fraudulent and circular transactions based on historical data, user history, and profiles.
2	Data Collection	Gather all the necessary data.	Transaction Details, User History, User Profiles
3	EDA	Conduct preliminary data analysis.	Missing Values, Distribution Analysis, Correlation Analysis, Trend Analysis
4	Data Preprocessing	Prepare the data for model training.	Handling Missing Values, Data Normalization, Data Transformation, Encoding Categorical Variables
5	Feature Engineering	Create new features or modify existing ones to improve model performance.	Time-Based Features, Statistical Features, Text Features (if applicable)
6	Model Selection	Choose appropriate machine learning models.	K-means for Cold Start Problems, Isolation Forest
7	Training	Train the selected models on the preprocessed data.	Use K-means to classify new transactions initially, then train Isolation Forest.
8	Evaluation	Assess model performance using various metrics.	Metrics: F1-score, Precision, Recall, AUC-ROC; Cross-Validation: K-Fold
9	Improvement	Fine-tune the model for better performance.	Hyperparameter Tuning, Feature Selection
10	Deployment	Deploy the model to a live environment once satisfactory performance is achieved.	Deploy once accuracy reaches 92% or better

What are the types of Fraud that can happen with a financial firm using Finacle, and some examples

Functionality	Type of Fraudulent Activity	Description
Core Banking	Account Takeover	Unauthorized users gaining control over a legitimate user's account.
	Fake Account Creation	Creating accounts with fraudulent identification to funnel illicit funds.
	Internal Fraud	Bank employees manipulating data for personal gains.
Online Banking	Phishing Attacks	Using fake websites to gain confidential user info.
	Man-in-the-Middle Attacks	Unauthorized interception of user data during a transaction.
	Transaction Reversal Fraud	A user maliciously reverses a completed online transaction.
Mobile Banking	SIM Swap Fraud	Unauthorized users gaining control over a user's mobile SIM and receiving OTPs.
	Mobile Malware	Malware on a user's device capturing sensitive banking details.
	QR Code Tampering	Scanning a tampered QR code which redirects payment to a fraudulent account.
Payments	Duplicate Payments	The same transaction gets processed multiple times fraudulently.
	Salary Fraud	Manipulating the payroll to disburse salaries to fake or terminated employee accounts.
	Cross-border Fraud	Illicit funds being transferred to offshore accounts to avoid detection and taxation.
Common Types	Circular Transactions	Quick back-and-forth transferring of money between accounts to obfuscate the money trail.
	Mismatched Locations/IP	Login or transaction requests coming from locations not usually associated with the account holder.
	Anomalies in Transaction Frequency/Amount	Unusually large or frequent transactions that deviate from normal behavior patterns.

Dataset with this features are used

Feature Category	Feature	Description	Data Type
User Demographics	Age	Age of the user	Numerical
	Gender	Gender of the user (Male, Female, Other)	Categorical
	Location	Location where the user resides (Urban, Suburban, Rural)	Categorical
	Marital_Status	Marital status (Single, Married, Divorced)	Categorical
User Profile	Employment	Employment status (Employed, Unemployed, Self-Employed)	Categorical
	AccountCreationDate	Date when the user account was created	Date
	SavingsAmount	Amount of money saved in the user's account	Numerical
User Transaction History	AvgTransactionValue	Average value of transactions	Numerical
	TransactionCountLast30Days	Number of transactions in the last 30 days	Numerical
	AvgTransactionPerMonth	Average number of transactions per month	Numerical
	AvgTransactionAmount	Average amount per transaction	Numerical
Device and Application Usage (Supply Side)	DeviceType	Type of device used for transactions (Mobile, Desktop, Tablet)	Categorical
	AppUsageTime	Time spent on the application in minutes	Numerical
	AppUsageDuration	Number of days the user has used the app	Numerical
	Browser	Browser used for transactions	Categorical
	ChatbotUsage	Whether the user has interacted with the chatbot	Binary (0 or 1)

App/Software (Demand Side)	ActiveUsers	Number of users currently active on the app	Numerical
	DailyNewUsers	Number of new users acquired daily	Numerical
	AverageSessionLength	Average time a user spends per session on the app	Numerical
	MonthlyRevenue	Total revenue generated by the app in a month	Numerical
Label	IsFraud	Whether the transaction is fraudulent or not	Binary (0 or 1)

Additional Features for Improvement

Time-Series Features	Trends over time, seasonality components, time-series analysis methods
Geographical Features	More granular geographical data like city, latitude, longitude
Network Features	Relationships between users, who sends money to whom
Boolean Flags	Multi-factor authentication status, login from multiple locations, etc.
Text-based Features	Transaction descriptions or notes, processed using natural language processing
Device-Related Features	Operating system, screen resolution, device model

Types of Algorithms used:

Category	Algorithm	Description
Statistical Methods		
	Z-Score Method	Computes Z-scores and flags points beyond a certain threshold as anomalies.
	MAD (Median Absolute	Similar to Z-score but based on median,

	Deviation)	making it more robust to outliers.
Distance-Based Methods		
	K-Nearest Neighbors (K-NN)	Flags observations as anomalies if their distance to their k nearest neighbors exceeds a predefined threshold.
	DBSCAN	Identifies low-density regions as anomalies.
	LOF (Local Outlier Factor)	Considers the local density deviation of a data point with respect to its neighbors.
Ensemble Methods		
	Random Forest for Anomaly Detection	Trained to perform anomaly detection by fitting it to the inliers and using it to score both inliers and outliers.
	Gradient Boosting for Anomaly Detection	Similar to the Random Forest but using Gradient Boosting Trees.
	Isolation Forest	An ensemble-based method specifically designed for anomaly detection. Isolates anomalies instead of constructing normal profiles.

- Numerical Features
- Categorical Features

- Z-scores can be applied to only numerical features
- But the numerical features are supposed to follow a NORMAL distribution
  - How can we check for normality?
    - Visually
      - Histogram
      - QQ plot

        - Box Plot
        - IQR plot
        - KDE plot
    - Formula based
        - Shapiro-Wilk Test
        - Kolmogorov-Smirnov Test
        - Anderson-Darling Test
        - D'Agostino and Pearson's Test
        - Skewness and Kurtosis
- If not, we should try converting them to NORMAL using Data Transformation techniques
    - Log Transformation
    - Square Root Transformation
    - Box-Cox Transformation
    - Yeo-Johnson Transformation
    - Z-score Transformation
    - Quantile Transformation
    - Min-Max Scaling
- What if we cannot transform them at any cost? Should we ignore them?
    - We cannot achieve 100% normality
    - Also, NOT all methods need NORMALITY
- Methods that ASSUME normality:

Method/Technique	Purpose/Usage	Assumes Normality Of
Parametric Tests		
t-Test	Used to compare the means of two groups.	Data (but robust to violations with large samples)
ANOVA	Used to compare the means of more than two groups.	Data (but robust to violations with large samples)
Linear Regression	Predictive modeling, especially for continuous outcomes.	Residuals
Pearson's Correlation	To measure the strength and direction of the relationship between two variables.	Each of the two variables

Quality Control		
Control Charts	Monitoring the quality of processes.	Data Points
Other Methods		
Principal Component Analysis	Dimensionality reduction, feature extraction, data visualization.	Original Variables
Discriminant Analysis	Classifying a set of observations into predefined classes.	Each group
Statistical Intervals		
Confidence Intervals	To estimate the range in which a population parameter lies with a certain level of confidence.	For means and other statistics (not strictly required)
Prediction Intervals	In regression analysis, to predict an individual outcome within a range.	Residuals (but can be relaxed with large samples)
	AppUsageDuration	Number of days the user has used the app
	Browser	Browser used for transactions
	ChatbotUsage	Whether the user has interacted with the chatbot

- Methods that does NOT ASSUME normality

Method/Technique	Purpose/Usage	Assumes Normality Of
Non-Parametric Tests		
Mann-Whitney U	Used to compare the distributions of two groups.	No

Test		
Kruskal-Wallis H Test	Used to compare the distributions of more than two groups.	No
Spearman's Rank Correlation	Measures the strength and direction of association between two ranked variables.	No
Wilcoxon Signed-Rank Test	Used to compare two related samples, paired samples, or repeated measurements on a single sample.	No
Machine Learning Methods		
Decision Trees	Classification and Regression	No
k-Nearest Neighbors	Classification and Regression	No
Random Forest	Classification and Regression	No
Gradient Boosting Machines	Classification and Regression	No
Clustering Methods		
k-Means Clustering	Partitioning a dataset into clusters based on similarity	No (but sensitive to outliers)
Hierarchical Clustering	Produces a tree of clusters, useful for understanding hierarchical relationships.	No
DBSCAN	Density-based clustering algorithm	No
Other Methods		
Chi-Square Test	Used to test the independence of two categorical variables.	No (but assumes large sample sizes for accuracy)

Fisher's Exact Test	Used to examine the significance of the association between two categorical variables (for small samples).	No
Bootstrap Methods	Used to estimate the sampling distribution of an estimator by resampling with replacement.	No