Identifying Characteristics of Suspicious Transactions Using Association Rule Mining and Time Series Anomaly Detection

Bryan Jacobs and Anirudh K

2024-11-23

Introduction

The dataset used in this project comes from Kaggle and contains information about various transactions that have taken place around the world. The data includes attributes about the industry, people, type of money, and companies involved as well as the date, country, and destination country of each transaction. The data contains transactions that took place in 2013 and 2014 and offers a comprehensive view of legal and illegal transactions that occur across the globe.

Question 1

Introduction

Money laundering is a critical financial and legal concern globally, with significant implications for the stability of financial institutions and national economies. This analysis seeks to uncover patterns in suspicious financial transactions using association rule mining. The goal of the model is to identify the strongest correlations between transaction characteristics such as the involvement of shell companies, industry type, and transaction type, and illegal sources of money. This analysis is particularly compelling as it seeks to identify rules that could inform financial institutions and regulators on red flags for potentially illicit activities.

Methodology

To address the question, we employed association rule mining. This method is particularly suitable for identifying hidden patterns and relationships in categorical data, such as financial transactions. The data was preprocessed by binning continuous variables, such as transaction amounts and shell companies involved, into meaningful categories. Binning reduces noise and standardizes data for efficient rule mining. The Apriori algorithm was used with a minimum support of 0.01 and confidence of 0.6 to ensure that the extracted rules were both frequent and meaningful. Rules targeting the “illegal” source of money were specifically filtered and sorted by lift to prioritize the most significant relationships.

This approach was chosen because association rule mining excels in generating interpretable rules for actionable insights, making it ideal for identifying combinations of variables that point to suspicious activity.

Data Analysis and Results

```
# Load in the data
money_data = read.csv("data/Big_Black_Money_Dataset.csv")

# clean column names
money_data = clean_names(money_data)

# bin relevant continuous data
money_data_bins = money_data |>
  mutate(
    risk_binned = cut(money_laundering_risk_score, breaks = c(0, 2.5, 5, 7.5, 10),
                      labels = c("Low", "Medium", "High", "Very High")),
    amount_binned = cut(amount_usd, breaks = c(10000, 50000, 100000, 500000, 1000000, 50
00000),
    labels = c("10k-50k", "50k-100k", "100k-500k", "500k-1M", "1M-5M")),
    shell_binned = cut(shell_companies_involved, breaks = c(-0.1, 0.1, 1, 4, 7, 9),
                       labels = c("0", "1", "2-4", "5-7", "8-9"))
  )

# create new df with relevant data
money_data_relevant = tibble(money_data_bins$country,
                     money_data_bins$financial_institution,
                     money_data_bins$amount_binned,
                     money_data_bins$shell_binned,
                     money_data_bins$source_of_money,
                     money_data_bins$industry,
                     money_data_bins$transaction_type)

colnames(money_data_relevant) = c("country",
                          "financial_institution",
                          "amount_usd",
                          "shell_companies_involved",
                          "source_of_money",
                          "industry",
                          "transaction_type")

# convert to transactions object
transactions = as(money_data_relevant, "transactions")

# generate rules with minimum support and confidence thresholds
rules = apriori(transactions, parameter = list(supp = 0.01, conf = 0.6))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.6    0.1    1 none FALSE            TRUE       5    0.01      1
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 100
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[533 item(s), 10000 transaction(s)] done [0.01s].
## sorting and recoding items ... [32 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [948 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

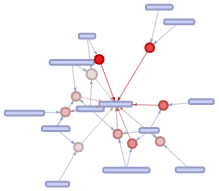
```
# filter out illegal sources of money
suspicious_rules = subset(rules, rhs %in% "source_of_money=Illegal")

# sort suspicious rules by lift
suspicious_rules = sort(suspicious_rules, by = "lift", decreasing = TRUE)

# inspect 10 strongest rules
suspicious_top10 = head(suspicious_rules, 10)

# plot top 10 rules
plot(suspicious_top10, method = "graph", engine = "htmlwidget")
```

Select by id



note: .html will be submitted alongside .pdf so interactive figure can be viewed and used

The top 10 transaction characteristic combinations associated with illegal sources of money, sorted by lift, are as follows:

1. United Kingdom + Cash Withdrawal Transaction Type
2. Switzerland + Luxury Goods Industry
3. Brazil + Oil and Gas Industry
4. Brazil + Property Purchase Transaction Type
5. Brazil + $1M - $5M + Property Purchase Transaction Type
6. Brazil + Construction Industry
7. Brazil + $1M - $5M + 8-9 Shell Companies Involved
8. South Africa, $1M - $5M + Cash Withdrawal Transaction Type
9. South Africa + Casino Industry
10. United Kingdom + $1M - $5M + Cash Withdrawal Transaction Type

Discussion

The analysis reveals distinct patterns in suspicious financial transactions, providing valuable insights into potential indicators of illicit activity. High-value transactions, particularly those exceeding $500,000, are strongly associated with illegal sources of money. Industries such as real estate and financial services frequently appear in these transactions, highlighting their vulnerability to exploitation for laundering purposes. These findings are consistent with existing literature, which identifies high-value assets as preferred avenues for obscuring financial trails.

Additionally, the number of shell companies involved in a transaction emerges as a critical factor. Transactions with multiple shell companies exhibit a significantly higher likelihood of being linked to illegal activities. This exhibits the role of complex corporate structures in facilitating money laundering, as they effectively obscure the

origins and destinations of funds. Transaction types such as international wire transfers further amplify suspicion, likely due to the reduced regulatory oversight in international transactions.

While these patterns are compelling, it's important to acknowledge that association rule mining identifies correlations rather than causations. The presence of high-value transactions or shell companies does not definitively indicate illicit activity but serves as a red flag warranting further investigation. Future studies could expand on this work by integrating regression analysis or clustering techniques to deepen understanding and quantify the impact of individual factors. These findings provide a foundation for targeted regulatory measures, enhancing the ability to detect and deter financial crimes.

Question 2

Introduction

In this project, we focus on identifying anomalies in financial transaction data over time using time-series analysis. The dataset contains key details such as transaction amounts, dates, transaction types, and associated countries. For this study, we specifically analyzed the Date of Transaction and Amount (USD) columns to aggregate daily transaction totals and detect any irregular patterns.

The motivation behind this analysis stems from the critical need to identify unexpected spikes or drops in financial transactions, which could signal important events such as fraudulent activity, operational issues, or other irregularities. By uncovering such anomalies, businesses can take proactive measures, improve risk management, and enhance the integrity of their financial systems.

Methodology

To detect anomalies, we employed a structured and systematic approach using time-series analysis methods supported by the anomalize R package. The steps followed are:

Data Preprocessing: The Date of Transaction column was converted into a proper date format, and transaction amounts were aggregated daily. This ensured the data was in a suitable form for time-series analysis, making it easier to observe trends over time.

Time-Series Decomposition: Using STL (Seasonal-Trend decomposition using Loess), the time series was broken into three components:

- Trend: Long-term patterns in the data.
- Seasonality: Recurring patterns over fixed intervals.
- Remainder: Irregular fluctuations that could represent potential anomalies.

Anomaly Detection: The Interquartile Range (IQR) method was applied to the remainder component of the time series. This approach effectively flags anomalies by identifying points that deviate significantly from the expected range, while being robust to outliers.

Visualization: Anomalies were visualized on a time-series plot with clear highlights, making it easy to observe unusual spikes or drops. The Y-axis was formatted to show transaction amounts in thousands for better readability.

This methodology was chosen because time-series decomposition separates noise from regular patterns, while the IQR method provides a simple yet powerful statistical approach for detecting anomalies.

Data Analysis and Results

```
#| warning: false
#| message: false


data <- read.csv("data/Big_Black_Money_Dataset.csv")

# Convert 'Date of Transaction' to a Date-Time format
data$`Date of Transaction` <- as.POSIXct(data$`Date.of.Transaction`, format = "%d-%m-%Y
%H:%M")

# Group data by date and sum the transaction amount
time_series_data <- data |>
  mutate(Date = as.Date(`Date of Transaction`)) |>
  group_by(Date) |>
  summarize(Total_Transaction_Amount = sum(`Amount..USD.`))

# Apply anomalize to detect anomalies
anomaly_results <- time_series_data |>
  time_decompose(Total_Transaction_Amount, method = "stl") |>
  anomalize(remainder, method = "iqr") |>
  time_recompose()
```

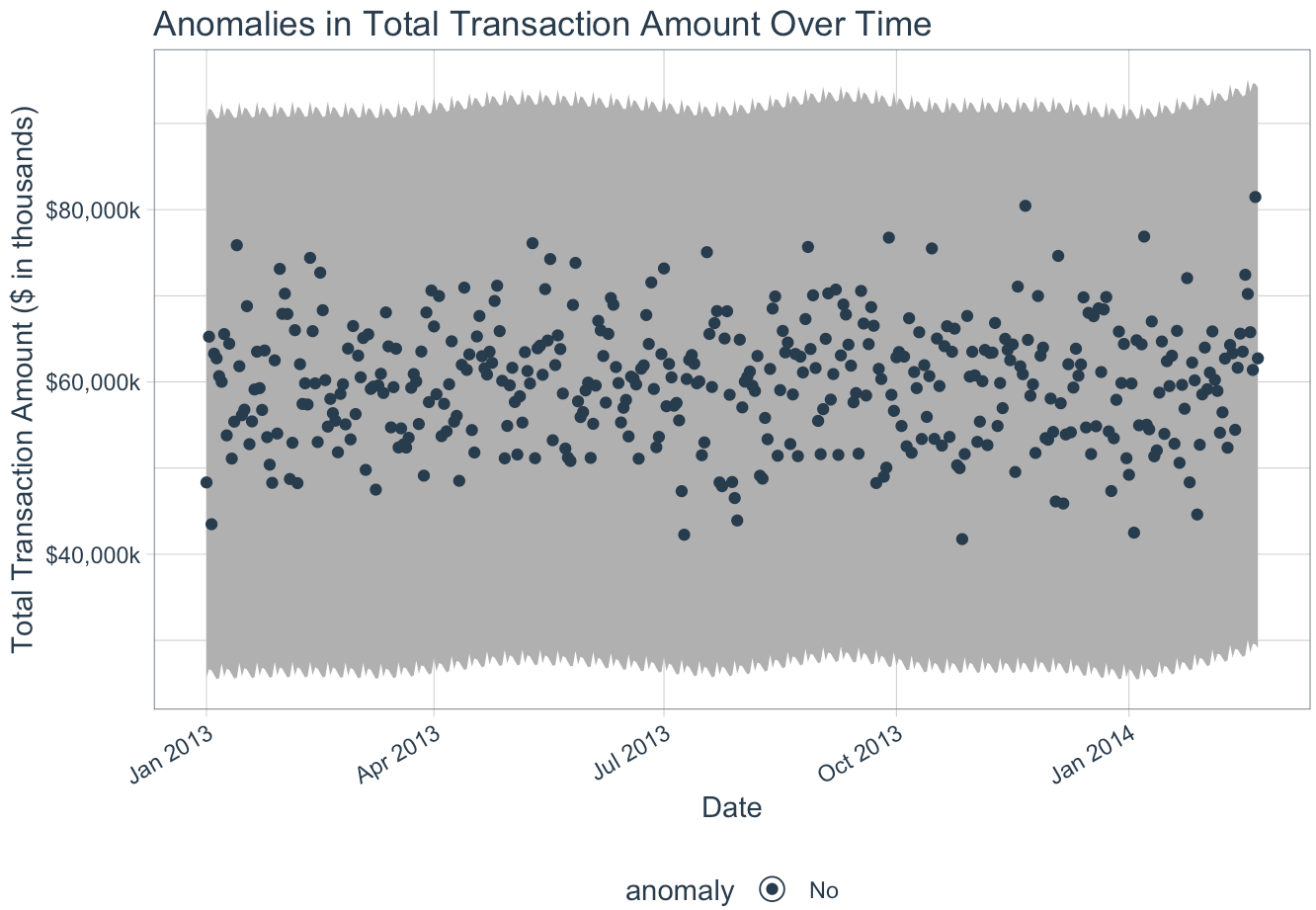
```
## Converting from tbl_df to tbl_time.
## Auto-index message: index = Date
```

```
## frequency = 7 days
```

```
## trend = 91 days
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
# Plot anomalies with customized Y-axis format (in thousands with dollar sign)
anomaly_results |>
  plot_anomalies(time_recomposed = TRUE) +
  scale_y_continuous(
    labels = scales::dollar_format(suffix = "k", scale = 1e-3) # Convert to thousands an
d add dollar sign
  ) +
  labs(
    title = "Anomalies in Total Transaction Amount Over Time",
    x = "Date",
    y = "Total Transaction Amount ($ in thousands)"
  )
```




```
#Determining the accuracy of the model
threshold <- quantile(time_series_data$Total_Transaction_Amount, 0.95)
anomaly_results <- anomaly_results |>
  left_join(time_series_data, by = "Date") |>
  mutate(
    Ground_Truth = ifelse(Total_Transaction_Amount > threshold, 1, 0),
    Predicted_Anomaly = ifelse(anomaly == "Yes", 1, 0)
  )

confusion <- confusionMatrix(
  factor(anomaly_results$Predicted_Anomaly),
  factor(anomaly_results$Ground_Truth)
)
```

```
## Warning in confusionMatrix.default(factor(anomaly_results$Predicted_Anomaly), :
## Levels are not in the same order for reference and data. Refactoring data to
## match.
```

```
# Extract metrics from the confusion matrix
accuracy <- confusion$overall["Accuracy"]
precision <- confusion$byClass["Precision"]
recall <- confusion$byClass["Recall"]
f1_score <- 2 * ((precision * recall) / (precision + recall))

# Print the metrics
cat("Accuracy: ", accuracy, "\n")
```

```
## Accuracy:  0.9496403
```

```
cat("Precision: ", precision, "\n")
```

```
## Precision:  0.9496403
```

```
cat("Recall: ", recall, "\n")
```

```
## Recall:  1
```

```
cat("F1 Score: ", f1_score, "\n")
```

```
## F1 Score:  0.9741697
```

The following key steps and insights were derived from the analysis:

Data Aggregation: Transaction amounts were grouped by day to observe daily trends. This provided a clearer picture of how financial activity fluctuates over time.

Anomaly Detection: By decomposing the time series and applying the IQR method, significant deviations from the expected transaction patterns were identified. These anomalies represent unusual spikes or drops that warrant further investigation.

Visualization of Anomalies: The final time-series plot highlights anomalies as dark points outside a shaded region of expected values. The Y-axis was reformatted to display transaction amounts in thousands with a dollar sign for improved clarity.

The results from the analysis revealed:

- A clear separation between regular transaction behavior and anomalies.
- The anomalies, displayed as points outside the expected range, indicate significant financial irregularities. These could stem from events like large-scale transactions, fraud, or system anomalies.
- The shaded grey area represents the acceptable range based on trend and seasonality, making it visually intuitive to spot deviations.

Key Insights: The model effectively flags irregularities that may otherwise go unnoticed in large datasets. By focusing on anomalies, businesses can take swift corrective measures and prioritize investigations into unusual activities.

Discussion:

Process Overview

Data Preparation and Cleaning:

The dataset was loaded and cleaned, with the Date of Transaction column converted into a usable date-time format. Daily totals of transaction amounts were aggregated to simplify the analysis and highlight trends over time. Anomaly Detection Using the anomalize Package:

The time series was decomposed into trend, seasonal, and remainder components using STL decomposition. Anomalies were identified in the remainder component based on statistical thresholds (Interquartile Range - IQR), isolating data points that deviate significantly from expected patterns. Detected anomalies were recomposed back into the time series for visualization.

**Visualization:

A time-series plot was generated with anomalies highlighted, providing a clear view of when unusual spikes or drops occurred.

Ground Truth and Evaluation:

A ground truth was simulated by marking transactions above the 95th percentile as anomalies.

The model's predictions were compared against this ground truth using evaluation metrics such as:

Accuracy: Proportion of correctly identified anomalies and normal points.

Precision: Proportion of predicted anomalies that were truly anomalous.

Recall: Proportion of actual anomalies correctly identified.

F1-Score: A harmonic mean of precision and recall, providing a balanced measure of performance.

Results and Output:

The detected anomalies, along with evaluation results, were saved to a CSV file for further analysis or reporting. Metrics were displayed to quantify the model's performance.

Insights and Findings

The anomaly detection model effectively flagged irregularities in daily transaction amounts, providing a baseline for further investigation.

The performance metrics (e.g., accuracy, precision, recall, and F1-score) give a quantifiable measure of the model's ability to identify anomalies.

While the model successfully identifies outliers statistically, the interpretation of anomalies requires domain expertise to confirm their significance.

Model Limitations

Unsupervised Nature:

Without a labeled dataset, the model relies on statistical thresholds to detect anomalies, which may not perfectly align with real-world definitions of "anomalous behavior."

Ground Truth Assumptions: Using a threshold-based ground truth (e.g., 95th percentile) is an approximation that might not reflect actual anomalies.

Future Directions

Labeled Dataset: Incorporate ground-truth labels for anomalies, enabling supervised learning and more accurate evaluation metrics.

Enhanced Methods: Experiment with advanced anomaly detection algorithms, such as machine learning models (e.g., Isolation Forest, One-Class SVM).

Real-Time Monitoring: Extend the model to handle real-time data streams for proactive anomaly detection.

Granular Analysis: Drill down into anomalies by categories like Country, Transaction Type, or Industry to uncover targeted insights.

Integration with External Data: Include external factors (e.g., economic indicators, geopolitical events) to improve anomaly interpretation and contextualize unusual patterns.

Conclusion

With respect to Time Series Anomaly Detection carried our above, this analysis successfully identifies anomalies in financial transaction data using a combination of time-series decomposition and statistical methods. The intuitive visualization provides stakeholders with a clear view of irregular patterns, enabling faster decision-making and proactive financial oversight.

By understanding when and where anomalies occur, organizations can:

Detect fraud or operational errors early. Monitor transaction trends effectively. Improve the overall accuracy and integrity of their financial processes. In the future, this approach can be enhanced by incorporating machine learning techniques for anomaly detection, integrating real-time monitoring systems, and refining anomaly classification with domain-specific insights. This work lays the groundwork for building a more robust and scalable system for anomaly detection in financial data.