COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Anomaly Detection for Time Series Data

Yuhui Li, Kittiya Thongpitak, Jingwen Zhang
Prof. Ali Hirsa; Advisor: Yossi Cohen

WELLINGTON
MANAGEMENT

Overview

Our objective is to detect anomalies within financial data using daily transaction counts and amounts from a dataset of 51 companies spanning 2016–2022. By employing regime detection, a voting-based anomaly detection system, and refining a deep learning model, we aim to develop a robust anomaly detection framework that adapts to different economic regimes and enhances anomaly detection accuracy.

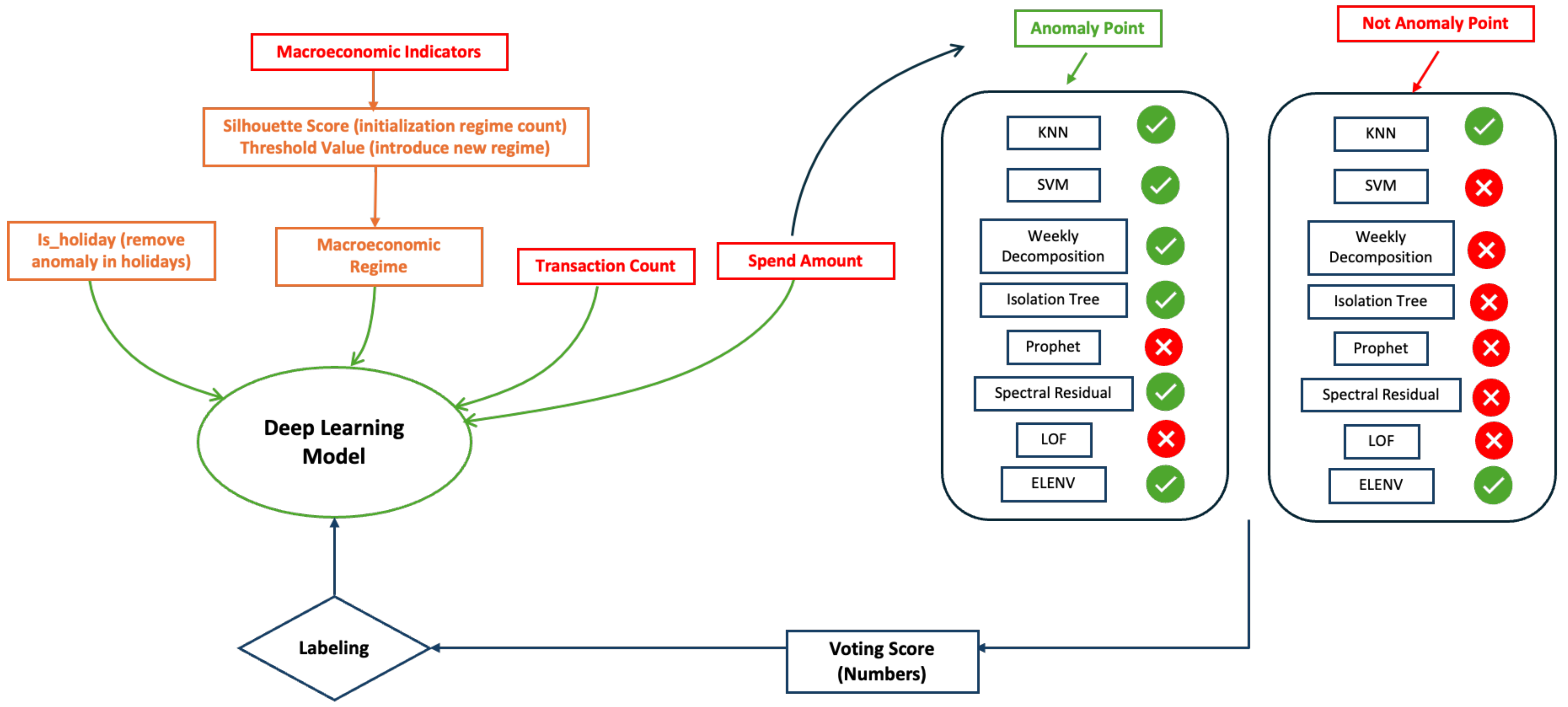
- **Regime Detection:** Building on a Robust Rolling Regime Detection(R2-RD) framework, we experimented with different threshold values for the regime detection, concluded that the current regime number is optimal. We also implemented a smoothing function to R2RD framework that merges regimes that have a duration smaller than a certain threshold.
- **Anomaly Detection Algorithms and Voting System:** We introduced SARIMA model in addition to the formerly selected eight algorithms, into a voting system to enhance the detection accuracy. We analyzed the distribution of anomalies across different regimes, plotting their percentages and amounts for each algorithm, providing a deeper understanding of the anomalies. In further enhancement, SARIMA model was also applied to each regime separately, combining results to improve anomaly detection by considering seasonality and regime shifts simultaneously. An evaluation system was also developed, calculating precision, recall, AUC-ROC and F1 scores to assess and refine the algorithms, ensuring the replacement of unfit models.
- **Deep Learning Model:** The results of the nine models were used as labels to train a deep learning model which was further refined through parameter tuning, experimenting with different epochs, batch sizes, learning rates, and optimizers.

By integrating these three components, our approach successfully highlights anomalies across various time periods and also offers a refined method for detecting anomalies in different economic scenarios. Future work could focus on these areas to further enhance the model's performance.

Data Description

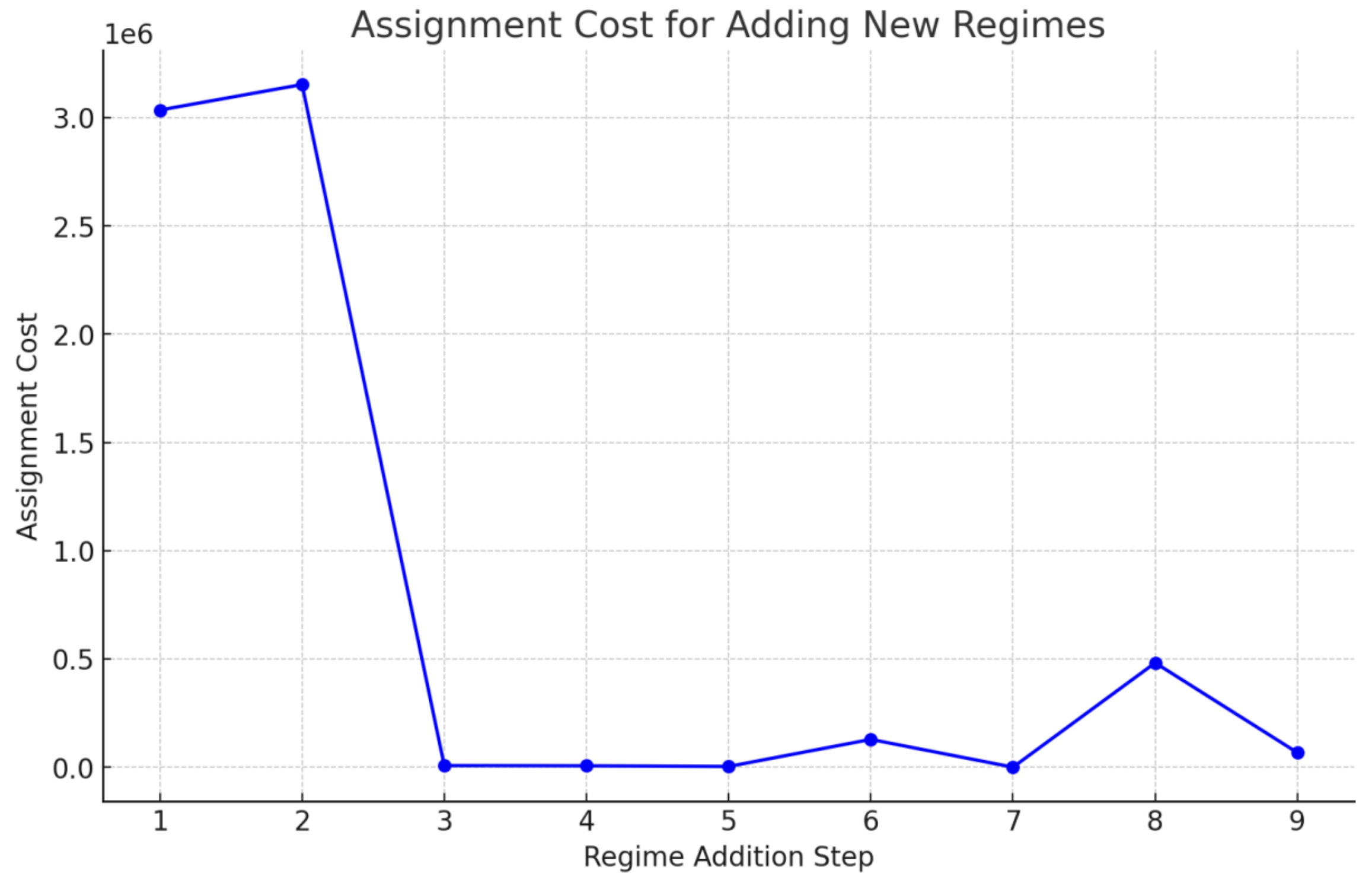
We utilized two distinct datasets:

1. **Financial Institution Dataset:** This dataset provided by Wellington Management Company comprises daily transaction data from 51 companies, covering the period from January 1, 2016, to December 31, 2022. For each company, the dataset includes two transaction count and transaction amount on daily basis.
2. **Macroeconomic Indicator Dataset:** Sourced from the Federal Reserve Economic Data (FRED), this dataset includes a selection of macroeconomic indicators relevant to the period from 2017 to 2022 on monthly basis. The indicators include: GDP, CPI, Unemployment, Interest Rate, Retail Sale and Trade Balance. It was used to supplement the financial institution dataset for regime detection.



Phase 1: Regime Detection

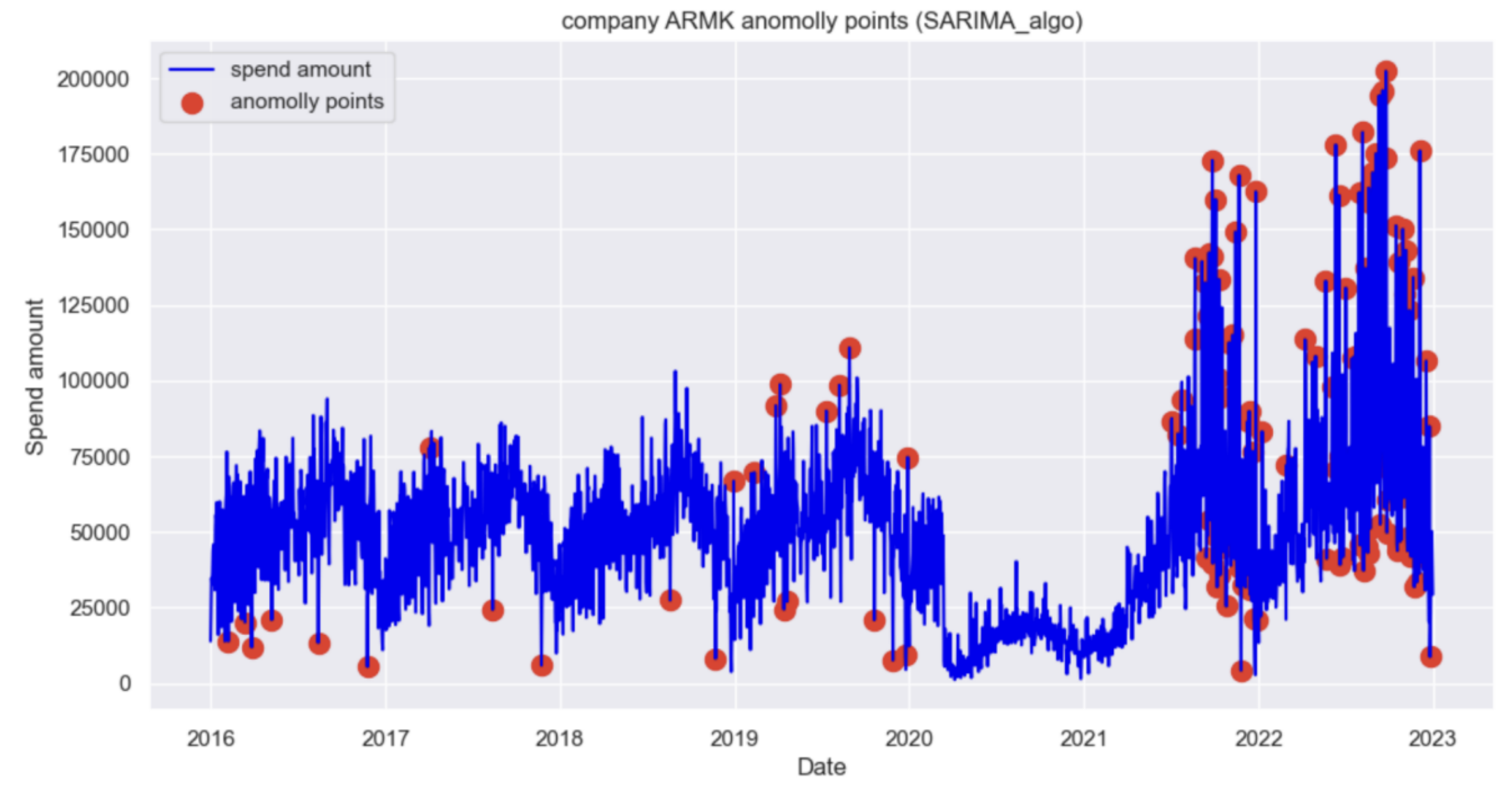
1. Threshold Value: The initial assignment cost for adding an additional regime is 3034895.767444778, with a previously calculated threshold of 3065244.7251192257. A new regime will be added if the assignment cost is less than the threshold. As a result, the three-regime (adding two regimes) model represents the optimal balance, capturing the major shifts, particularly around the COVID period, without introducing excessive and trivial regimes.



2. Smoothing Function: To smooth the regime transitions and reduce the number of fragmented regimes, we implemented a function that merges regimes with a duration smaller than a certain threshold. The function iterates through the regime sequence and merges adjacent regimes if the duration of the first regime is less than the threshold.
The updated $plot\_regime\_detection\_with\_new\_regimes$ function introduces logic to merge regimes that fall below a specified and adjustable minimum length $(min\_regime\_length = 150)$. This function aims to improve the readability and interpretability of regime segmentation by consolidating shorter, potentially noise-like regimes into adjacent larger ones.

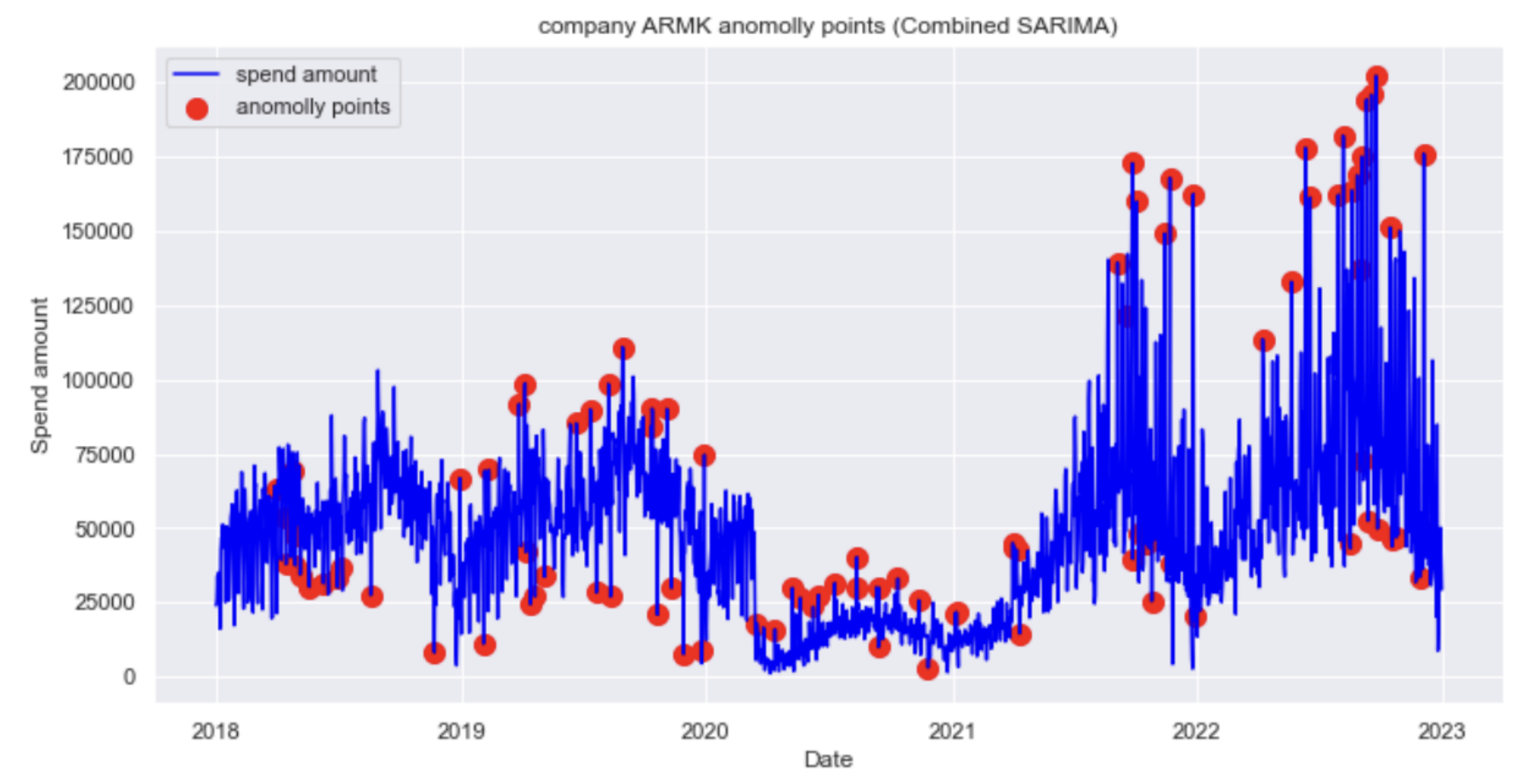
Phase 2: Voting-based Anomaly Detection

1. SARIMA model: Based on the 8 algorithms from the last group, we added the SARIMA (Seasonal AutoRegressive Integrated Moving Average) model, which is a time series forecasting model that extends ARIMA by adding seasonal components to account for seasonal patterns in the data. In the experiment conducted for the ARMK company dataset, 128 points out of 2557 data points were flagged as anomalies, representing approximately 5% of the total data.



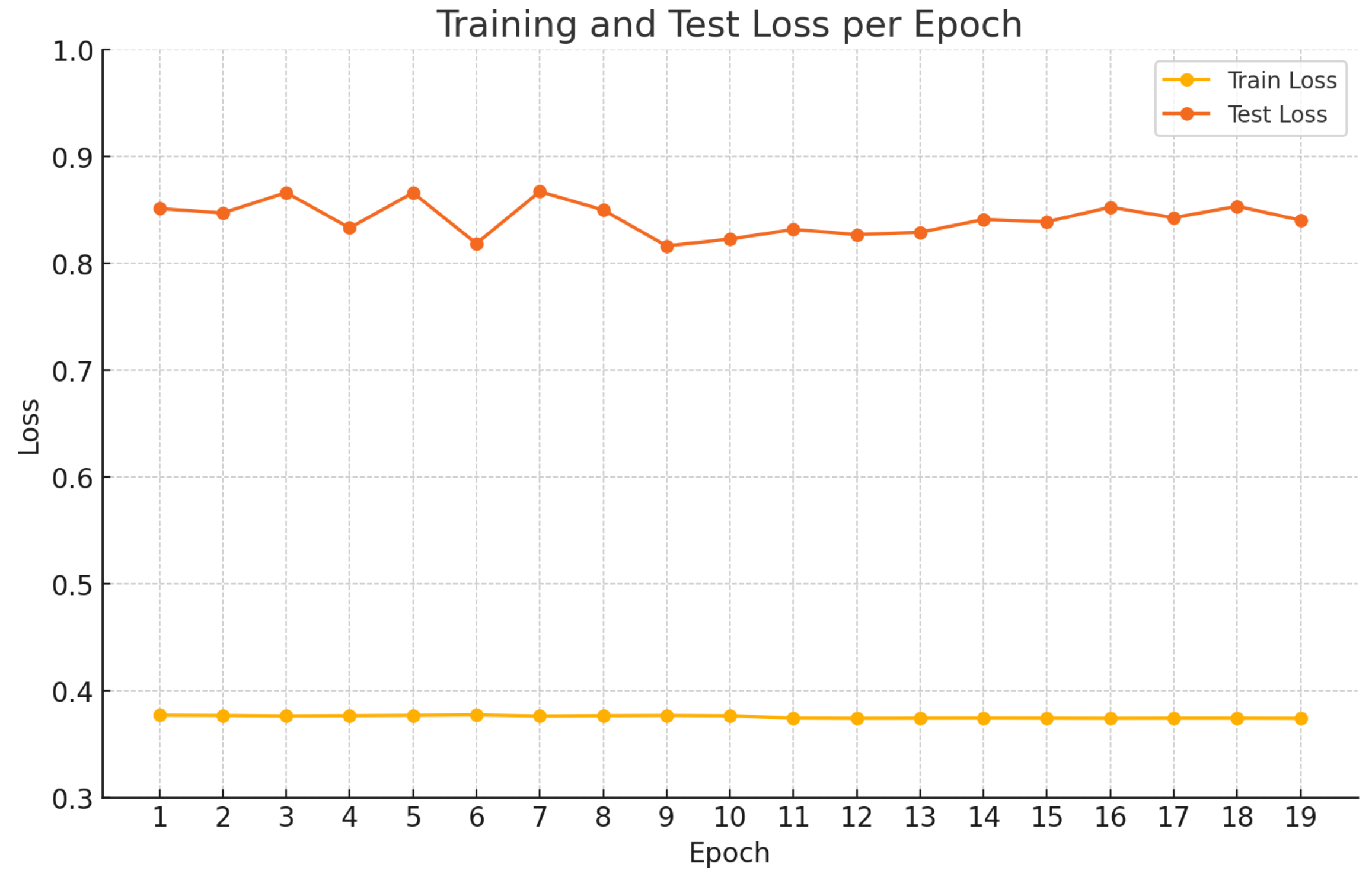
We incorporated all the algorithms, including LOF, ELEVN, Isolation Forest, KNN, and more, into a voting system. The system determined anomalies by consensus—if a data point was flagged by the majority of algorithms, it was considered an anomaly.

2. Enhanced SARIMA model: We applied SARIMA models to each regime separately, combining results to improve anomaly detection by considering seasonality and regime shifts simultaneously.
The initial SARIMA-based approach struggled to adapt to abrupt shifts in the data, leading to false positives or missed anomalies. The enhanced approach, by leveraging regime-specific models, provided a more robust framework for anomaly detection, particularly in the presence of regime shifts.

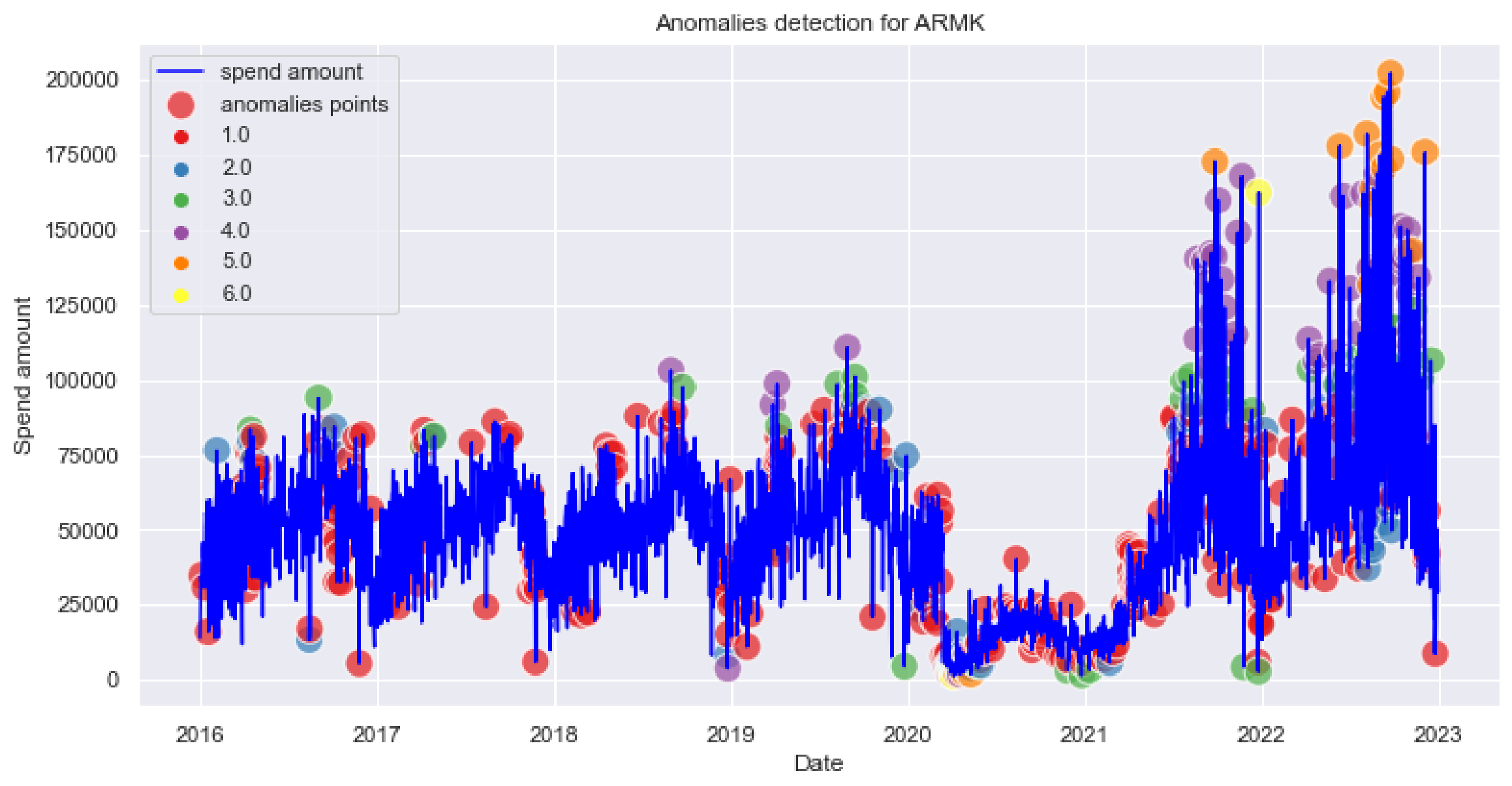


Phase 3: Refined Deep Learning Model

The current framework was built by the previous group, including the use of Knowledge Distillation model. The deep learning model was refined through extensive parameter tuning to improve model stability and performance. After experimenting with various hyperparameters, such as learning rates, batch sizes, and optimizers, the optimal setup includes a learning rate of 0.01 with a scheduler, 200 epochs with early stopping, a batch size of 64, and the Adam optimizer.



Conclusion



The sample results demonstrate that our methodology is capable of adapting to different economic scenarios, providing valuable insights into financial patterns and enabling more accurate anomaly detection. Future work could explore further model enhancements, such as including additional data sources and experimenting with other machine learning algorithms to improve performance.