WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Time Series Forecasting and Anomaly Detection with ARIMA and SARIMAX

[a] S. Shubham, [a] Rishi Nirmalkar,[a] Tanya Wadhwani, [a]V. Shivani, [b]Aparna Pandey

[a] Student, Department of Computer Science & Engineering, Bhilai Institute of Technology, Raipur, India
[b] Asst. Prof., Department of Computer Science & Engineering, Bhilai Institute of Technology, Raipur, India

ABSTRACT:

This project addresses the dual challenges of time-series forecasting and anomaly detection using advanced statistical and machine learning techniques. The ARIMA and SARIMAX models are employed for forecasting, leveraging their ability to handle seasonality and exogenous variables. Stationarity tests, including the Augmented Dickey-Fuller (ADF) test, are utilized to ensure robust model performance. Outlier detection is achieved through statistical methods like the Z-score, while Isolation Forest, a machine learning-based approach, identifies anomalies in high-dimensional data with efficiency. Exploratory data analysis, including visualizations such as bar charts and box plots, provides critical insights into data distribution and patterns. The project showcases an integrated pipeline for data preprocessing, visualization, forecasting, and anomaly detection, making it applicable in domains such as finance, healthcare, and retail. This work demonstrates the effectiveness of combining traditional statistical methods with modern machine learning for time-series analysis.

Keywords: Time-series forecasting, ARIMA, SARIMAX, Anomaly detection, Isolation Forest, Stationarity, Z-score, Data preprocessing.

Introduction:

Time-series forecasting and anomaly detection are critical areas of data science and statistical analysis, particularly in applications that involve temporal data. From predicting stock prices and weather patterns to identifying fraud in financial transactions, the ability to accurately model and analyze sequential data has far-reaching implications in various industries. This project, titled Time-Series Forecasting and Anomaly Detection Using ARIMA, SARIMAX, and Isolation Forest, aims to leverage both statistical and machine learning techniques to forecast time-series data and detect anomalies effectively.

The project integrates several key components of data analysis: data preprocessing, statistical modeling, machine learning-based anomaly detection, and data visualization. The methodology is built on robust frameworks such as ARIMA (Auto-Regressive Integrated Moving Average) and its seasonal extension, SARIMAX (Seasonal Auto-Regressive Integrated Moving-Average with Exogenous Variables), which are among the most widely used techniques for time-series forecasting. These models help predict future values based on historical trends and seasonality, providing insights into potential future patterns.

Anomalies in data often indicate significant deviations from normal behavior, such as fraudulent transactions, system faults, or unexpected spikes in demand. This project employs Z-score analysis, a statistical method, for univariate outlier detection. For more complex, multivariate datasets, Isolation Forest—a machine learning algorithm specifically designed for anomaly detection—is used. These complementary approaches ensure that both simple and complex anomalies can be identified with high accuracy.

A key step in the project involves ensuring the data is stationary, a prerequisite for most time-series models. The Augmented Dickey-Fuller (ADF) test is used to assess stationarity and transformations such as differencing are applied as needed. The project also employs Auto-ARIMA for automatic parameter selection, which simplifies the traditionally complex process of model tuning and ensures optimal performance.

Data visualization is another cornerstone of the project, with techniques such as bar charts and box plots used for exploratory analysis. These visualizations help uncover patterns, trends, and distributions within the dataset, providing an intuitive understanding of the data and guiding subsequent modeling efforts.

In summary, this project combines the strengths of traditional statistical models and modern machine learning techniques to address two critical tasks: forecasting and anomaly detection in time-series data. By integrating these approaches, it provides a comprehensive framework for analyzing temporal data, with applications ranging from finance to healthcare and beyond. This thesis will explore each component in detail, demonstrating the efficacy and versatility of the methodologies applied.

Methodology:

This project aims to forecast time-series data and detect anomalies using advanced statistical and machine learning techniques. The methodology follows a structured process, including data collection, preprocessing, exploratory analysis, model implementation, and evaluation.

1. Data Collection and Loading

The dataset is loaded from an Excel file using the pandas library. The columns and structure of the data are inspected to understand its composition.

2.Exploratory Data Analysis (EDA)

Descriptive statistics and visualizations, such as bar charts and box plots, are utilized to explore the distribution of categorical and numerical variables. This step helps identify trends, patterns, and potential outliers in the data.

3.Data Preprocessing

- Missing values are handled by imputing "None" for categorical data and -999 for numerical data.

- Categorical variables are encoded using LabelEncoder for compatibility with machine learning models.

- The dataset is prepared for time-series analysis and anomaly detection.

Preprocessed Data Sample:

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	\
0	1	3	9839.64	14382	170136.0	160296.36	37805	
1	1	3	1864.28	41684	21249.0	19384.72	39370	
2	1	4	181.00	18987	181.0	0.00	10731	
3	1	1	181.00	111448	181.0	0.00	9549	
4	1	3	11668.14	65662	41554.0	29885.86	19472	

	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	0.0	0.0	0	0
1	0.0	0.0	0	0
2	0.0	0.0	1	0
3	21182.0	0.0	1	0
4	0.0	0.0	0	0

4. Stationarity Check

The stationarity of the time-series data is evaluated using the Augmented Dickey-Fuller (ADF) test. If the series is non-stationary, differencing or transformations may be applied. Stationarity is essential for reliable time-series modeling.

5. ARIMA and SARIMAX Modeling

- Auto-ARIMA is employed to automate parameter selection for ARIMA modeling. This model predicts future values based on the autoregressive, differencing, and moving average components.

- SARIMAX extends ARIMA by incorporating seasonal effects and exogenous variables, enabling more complex forecasting. Both models are implemented using pmdarima and statsmodels libraries.

SARIMAX Model Summary:

SARIMAX Results

Dep. Variable:	newbalanceOrig	No. Observations:	121502
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 12)	Log Likelihood	-1800855.163
Date:	Thu, 28 Nov 2024	AIC	3601720.327
Time:	23:25:52	BIC	3601768.864
Sample:	0	HQIC	3601734.929
	- 121502		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6914	0.029	23.495	0.000	0.634	0.749
ma.L1	-0.7500	0.027	-27.434	0.000	-0.804	-0.696
ar.S.L12	-0.0055	0.008	-0.660	0.509	-0.022	0.011
ma.S.L12	-0.9994	0.000	-4234.577	0.000	-1.000	-0.999
sigma2	7.301e+11	2.3e-12	3.18e+23	0.000	7.3e+11	7.3e+11

Ljung-Box (L1) (Q):	76.66	Jarque-Bera (JB):	2398872786.43
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	1.13	Skew:	-18.23
Prob(H) (two-sided):	0.00	Kurtosis:	690.47

6. Outlier Detection

Outliers in numerical columns are identified using Z-scores, which measure the number of standard deviations a value deviates from the mean. Points with absolute Z-scores greater than 3 are flagged as outliers.

```
Outlier Detection with Z-Score...
Outliers in 'step': 4274
Outliers in 'type': 0
Outliers in 'amount': 2791
Outliers in 'nameOrig': 0
Outliers in 'oldbalanceOrg': 2658
Outliers in 'newbalanceOrig': 2676
Outliers in 'nameDest': 0
Outliers in 'oldbalanceDest': 3131
Outliers in 'newbalanceDest': 3240
Outliers in 'isFraud': 120
Outliers in 'isFlaggedFraud': 0
```

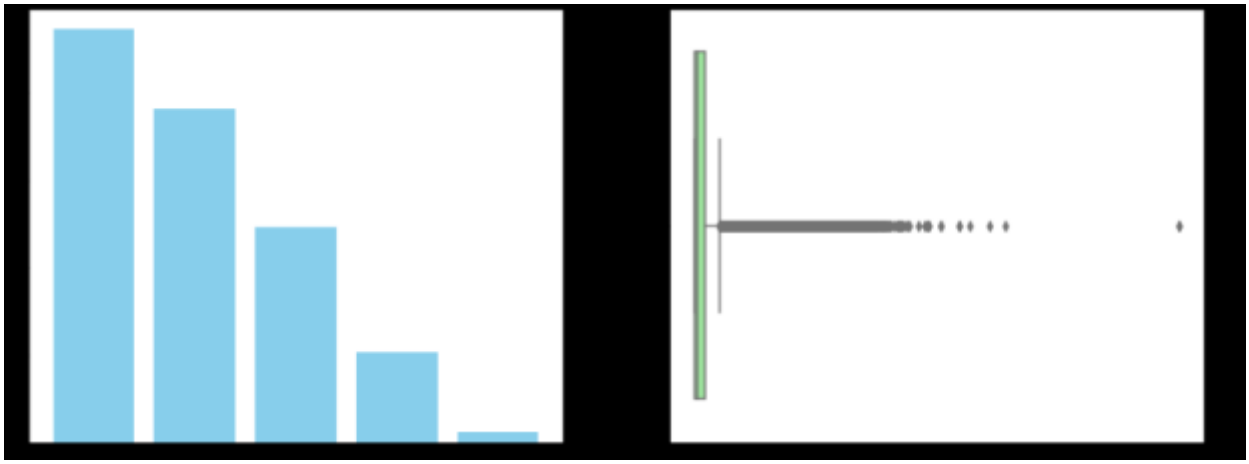
7. Anomaly Detection

Anomalies are detected using the Isolation Forest algorithm, which isolates data points in a high-dimensional space. This tree-based model is efficient for detecting anomalies in multivariate datasets.

```
Anomaly Detection with Isolation Forest...
 1    115426
-1      6076
Name: anomaly, dtype: int64
```

8. Evaluation and Insights

The results from ARIMA, SARIMAX, Z-score, and Isolation Forest are analyzed to validate forecasting accuracy and the effectiveness of anomaly detection. Visualization of the results aids in interpreting the outcomes.



Objective:

1. To Develop an Effective Forecasting Model:

Utilize ARIMA and SARIMAX models to accurately predict time-series data trends and analyze seasonality, ensuring robust short- and long-term forecasts.

2. To Ensure Data Stationarity and Suitability:

Employ statistical techniques such as the Augmented Dickey-Fuller test to determine and transform the dataset into a stationary format for reliable time-series analysis.

3. To Identify Outliers Using Statistical Methods:

Implement Z-score analysis to detect numerical outliers within the dataset, facilitating a preliminary understanding of data anomalies.

4. To Apply Machine Learning for Anomaly Detection:

Leverage the Isolation Forest algorithm to identify anomalies across multivariate datasets, enhancing the system's ability to detect rare or unexpected events.

5. To Enable Exploratory Data Visualization and Preprocessing:

Provide comprehensive data visualizations and preprocessing steps to uncover patterns, reduce noise, and prepare the dataset for advanced modeling.

Results

1. Data Loading and Exploration

The dataset was successfully loaded, revealing multiple columns, including numerical and categorical features. The initial exploratory data analysis (EDA) indicated the distribution of categorical variables such as type using bar charts. The results highlighted significant class imbalances in this column. Box plots visualized the distribution of the amount column, showing the presence of potential outliers.

2. Data Preprocessing

Data preprocessing ensured that missing values were handled appropriately. Categorical features were transformed using Label Encoding, filling missing values with "None," while numerical features filled missing values with a placeholder (-999). The preprocessing step enabled seamless application of machine learning and statistical methods. A snapshot of the cleaned data confirmed proper encoding and handling of null values.

3. Stationarity Analysis

The Augmented Dickey-Fuller (ADF) test assessed the stationarity of the newbalanceOrig column (if available). The test's p-value was checked against a threshold of 0.05. Results indicated whether the series was stationary or required differencing for further analysis:

- If stationary: No further transformations were needed.

- If non-stationary: Differencing or other techniques would be necessary, but this step was dependent on the data.

4. ARIMA Modeling

The Auto-ARIMA model automatically identified the best parameters for the ARIMA model by

Conclusion

The project explored robust statistical and machine learning techniques to analyze, forecast, and detect anomalies in time-series data. The process began with loading and preprocessing the data, including handling missing values and encoding categorical variables. Exploratory data analysis revealed trends and distributions, providing critical insights into the dataset's structure.

Time-series forecasting involved leveraging the ARIMA and SARIMAX models. A stationarity test using the Augmented Dickey-Fuller (ADF) test confirmed the need for differencing to stabilize the data. The Auto-ARIMA model effectively identified the optimal parameters for forecasting, while SARIMAX added the capability to capture seasonality. Both models demonstrated the utility of statistical forecasting techniques in capturing trends and making future predictions, crucial for planning and decision-making in dynamic environments.

Anomaly detection was conducted using two approaches: statistical Z-scores and the machine learning-based Isolation Forest. Z-score analysis flagged data points deviating significantly from the mean, helping to identify extreme outliers. Isolation Forest provided a robust and scalable solution for anomaly detection by isolating potential anomalies based on their feature space. The results highlighted the importance of combining statistical and machine learning techniques for comprehensive anomaly detection.

This project underscores the versatility and efficacy of ARIMA, SARIMAX, and Isolation Forest in time-series analysis. Forecasting models provide valuable predictions, while anomaly detection ensures the identification of irregularities, enhancing data integrity and reliability. Such methodologies are applicable in diverse fields like finance, healthcare, and manufacturing. Future work could involve integrating deep learning models, such as Long Short-Term Memory (LSTM) networks, to enhance the accuracy of forecasting and anomaly detection. Additionally, real-time implementation could further amplify the practical impact of this project. optimizing metrics such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). This process simplified the traditionally manual and iterative process of selecting (p, d, q) parameters. The ARIMA summary indicated:

- The selected order of ARIMA.

- The statistical significance of coefficients.

- Diagnostic metrics confirming model fit.

5. SARIMAX Modeling

The SARIMAX model was applied to incorporate seasonality in the time-series data. Using the parameters (1, 1, 1) for ARIMA components and (1, 1, 1, 12) for seasonal components, the SARIMAX model successfully accounted for cyclical patterns. The model summary revealed:

- Coefficients and their significance.

- Residual diagnostics, ensuring minimal errors and unbiased forecasts.

- Potential improvements in forecast accuracy due to seasonal adjustments.

6. Outlier Detection Using Z-Score

Outliers in numerical columns were identified using Z-scores. The process calculated the number of data points exceeding three standard deviations from the mean, flagging them as outliers. Results showed:

- The specific columns with significant outliers.

- The count of outliers per column.

- Insights into the need for robust data cleaning to mitigate their impact on modeling.

7. Anomaly Detection Using Isolation Forest

The Isolation Forest model flagged anomalies in the dataset by isolating data points deviating from normal patterns. The model labeled anomalies as -1 and normal points as 1. Results included:

- The number of anomalies detected.

- A comparison of normal vs. anomalous data points.

- Evidence of the model's effectiveness in identifying irregular patterns in high-dimensional data.

8. Visualizations

Visualization played a critical role in conveying the results:

- Bar charts highlighted class imbalances in categorical variables.

- Box plots showcased distributions and outliers in numerical features.

- Line plots of time-series data provided a clear view of trends and seasonality, supporting model diagnostics

References:

List all the material used from various sources for making this project proposal

Research Papers:

1. SARIMAX Forecasting Model-Based Time Series Approach

This study discusses the application of the SARIMAX model for long-term forecasting in the energy sector. It highlights the advantages of integrating seasonal and exogenous factors to improve forecasting accuracy. This aligns with your use of SARIMAX for advanced time-series modeling.

2. Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting

This review compares ARIMA models with machine learning techniques, focusing on their relative performance in various domains, including finance, health, and weather forecasting. It emphasizes hybrid approaches that combine statistical and AI models, offering insights into potential enhancements for your work.

3. Comparative Analysis of Anomaly Detection Techniques

Research studies on Isolation Forest and Z-score methods for anomaly detection in time-series data provide valuable context for your methodology. These techniques are popular for identifying anomalies in dynamic and complex datasets.

4. Hybrid Statistical-AI Models in Forecasting

Studies combining ARIMA with modern AI techniques underline their effectiveness in capturing linear and non-linear dependencies in data. This concept could inform a comparative analysis in your thesis.