Data Visualization of outliers in univariate and multivariate signal

Abstract

This project focuses on the exploration and detection of outliers in univariate and multivariate signals across three diverse datasets: catfish sales data, the Expedia dataset, and baby cry audio signals. The objective is to leverage various data visualization techniques and outlier detection algorithms to identify anomalies within the datasets, providing valuable insights into irregularities or unusual patterns. The project involves data preprocessing, exploratory data analysis, and the application of outlier detection techniques such as Isolation Forest, One-Class SVM, and KNN Outlier Detection. Section-wise outlier detection is conducted to analyze localized anomalies within the datasets. The findings from this project contribute to the understanding of outlier detection in different domains, ranging from market analysis to online travel booking and infant healthcare. Future research may focus on exploring advanced outlier detection techniques and incorporating additional features to enhance anomaly detection and interpretation.

1 Data Sources

1.1 Catfish Sales Data

1.1.1 Description

The catfish sales dataset provides information about the sales of catfish over a certain period. It includes univariate time series data representing the sales volume over time. The dataset consists of monthly sales data spanning several years, making it suitable for time series analysis.

1.1.2 Data Preprocessing

The data was loaded from a CSV file, and dates were parsed and set as the index for time series analysis. Preprocessing steps included handling missing values, which were imputed using interpolation techniques, and removing outliers to ensure data integrity.

1.2 Expedia Dataset

1.2.1 Description

The Expedia dataset comprises multivariate time series data related to online travel booking behaviors. It includes various features such as price, booking window, search criteria, and more. The dataset provides insights into booking trends, seasonal variations, and user behaviors in the online travel domain.

2 Exploratory Analysis and Outlier Detection in Catfish Sales

2.1 Introduction

The objective of this project is to analyze catfish sales data, identify patterns, and detect outliers using various statistical and machine learning techniques. Catfish sales data is crucial for understanding market trends, seasonal variations, and detecting unusual behavior that may require further investigation.

2.2 Data Loading and Preprocessing

Loaded the catfish sales dataset from a CSV file. Parsed dates and set them as the index for time series analysis. Manipulated some data points for illustration purposes.

2.3 Time Series Visualization

Plotted the original time series data of catfish sales. Identified trends and seasonal patterns in the data. Utilized vertical lines to mark the start of each year for better visualization.

2.4 Seasonal Decomposition

Performed seasonal decomposition using the additive model. Decomposed the time series into trend, seasonal, and residual components. Visualized the decomposed components to analyze the seasonal patterns and identify irregularities.

2.5 Outlier Detection Techniques

- **Isolation Forest:**
  - Used the Isolation Forest algorithm for outlier detection.
  - Scaled the data and trained the Isolation Forest model.
  - Detected anomalies in the catfish sales data and visualized them alongside normal data points.

- **One-Class SVM:**
  - Applied the One-Class SVM algorithm for outlier detection.
  - Detected anomalies and compared them with Isolation Forest results.
- **KNN Outlier Detection:**
  - Integrated KNN outlier detection using the Local Outlier Factor algorithm.
  - Detected anomalies based on local data point density and visualized them.
- **Section-wise Outlier Detection:**
  - Divided the time series data into sections for localized analysis.
  - Applied Isolation Forest for outlier detection in each section separately.
  - Visualized outliers detected in each section individually and collectively.

2.6 Conclusion

The project successfully analyzed catfish sales data, identified seasonal patterns, and detected outliers using various techniques. The insights gained from outlier detection can be valuable for understanding unusual behavior in catfish sales, which may require further investigation or action. Further analysis and interpretation of the detected anomalies could provide valuable insights for decision-making and business strategy in the catfish industry.

2.7 Future Work

Explore advanced anomaly detection techniques and ensemble methods for improved outlier detection. Incorporate additional features or external factors that may influence catfish sales, such as weather conditions or economic indicators. Conduct predictive modeling to forecast future catfish sales based on historical data and identified patterns.

3 Exploratory Analysis and Outlier Detection in Time Series Expedia Data

3.1 Introduction

The objective of this project is to analyze time series data from the Expedia dataset and detect outliers using various statistical and machine learning techniques. The project involves exploratory analysis, feature engineering, clustering, and outlier detection to gain insights into the underlying patterns and anomalies in the data.

3.2 Data Loading and Preprocessing

Loaded the time series data from the Expedia dataset. Selected relevant features for analysis, including 'price_usd', 'srch_booking_window', and 'srch_saturday_night_bool'. Imputed missing values using the mean strategy to ensure completeness in the dataset.

3.3 Isolation Forest for Outlier Detection

Applied the Isolation Forest algorithm to detect outliers in the dataset. Identified outliers based on their isolation scores and categorized them as anomalies. Visualized outliers against normal data points to understand the distribution and characteristics of anomalies.

3.4 Section-wise Outlier Detection

Divided the dataset into sections for localized outlier detection using Isolation Forest. Detected outliers in each section separately and aggregated the results for comprehensive outlier detection. Visualized outliers detected by section-wise Isolation Forest to analyze their distribution across different segments of the data.

3.5 KMeans Clustering for Feature Analysis

Utilized KMeans clustering to analyze the relationship between features and identify clusters within the data. Evaluated the optimal number of clusters using the elbow method to determine the appropriate granularity for clustering. Visualized clusters in 3D space to gain insights into the underlying patterns and structure of the data.

3.6 Principal Component Analysis (PCA)

Conducted PCA to reduce the dimensionality of the data and extract important features. Standardized and transformed the data into principal components for further analysis. Visualized the explained variance ratio to understand the contribution of each principal component to the overall variance.

3.7 Anomaly Detection with KMeans and Distance Metrics

Implemented anomaly detection using KMeans clustering and distance metrics. Calculated the distance between each data point and its nearest centroid to identify outliers. Categorized anomalies based on a predefined threshold and visualized them in the time series data.

3.8 Conclusion

The project successfully explored time series data from the Expedia dataset and detected outliers using various techniques including Isolation Forest, section-wise outlier detection, KMeans clustering, PCA, and distance-based anomaly detection. Insights gained from outlier detection can be valuable for anomaly detection, fraud detection, and anomaly-driven decision-making in the domain of online travel booking. Further analysis and interpretation of outliers could provide actionable insights for improving pricing strategies, detecting irregular booking behaviors, and enhancing user experience on the Expedia platform.

3.9 Future Work

Explore advanced outlier detection techniques such as Local Outlier Factor (LOF), One-Class SVM, or deep learning-based approaches for improved anomaly detection. Incorporate additional features and external factors such as seasonal trends, user demographics, and geographic locations for more robust analysis and anomaly detection. Conduct predictive modeling to forecast booking trends and identify potential anomalies in advance for proactive management and intervention.

4 Outlier Detection in Audio Analysis

4.1 Introduction

This report presents a computational framework for the detection of outliers in audio signals, with a particular focus on the analysis of baby cries. The motivation behind this study is to leverage audio signal processing techniques to detect anomalies which could indicate distress or health issues in infants. The objective is to employ advanced algorithms such as Fast Fourier Transform (FFT), Principal Component Analysis (PCA), and Isolation Forest to systematically identify these outliers. This study provides towards outlier detection aids in the timely recognition of atypical patterns that could signify important cues in medical and caregiving settings.

4.2 Motivation

The analysis of audio signals, specifically baby cries, holds profound importance in healthcare and parenting. Deciphering the nuances of infant cries can lead to early detection of health issues, providing caregivers and healthcare professionals invaluable insights into an infant's well-being. The motivation of this work lies in developing a computational tool that aids in this interpretation, potentially bridging the gap between quantitative data and qualitative assessment.

4.3 Objective

The primary goal of this work is to apply advanced data analysis techniques to identify outliers within audio signals. The analysis focuses on determining whether the computational models can effectively flag anomalies in the spectral features of baby cries.

4.4 Importance of Outlier Detection

Outlier detection is a critical aspect of data analysis, particularly in audio signal processing. It involves identifying data points that deviate significantly from the majority of the data. In the context of baby cries, outliers may represent cries that are unusual or indicative of distress and potential health issues. Detecting these outliers promptly can facilitate early intervention, ensuring that infants receive the care they need without delay.

4.5 Algorithmic Approach

Our approach employs a sequence of algorithms to analyze the audio signal. FFT is utilized to transform the time-domain signal into the frequency domain, providing insight into the frequency components of the audio signal. PCA reduces the dimensionality of the data, aiding in visualizing and detecting outliers. Lastly, the Isolation Forest algorithm, an unsupervised learning method, is applied to identify anomalies within the data effectively.

4.6 Observations

- **Time-Domain Visualization:** The time-domain visualization provides an overview of the amplitude fluctuations over time, offering initial insights into the cry's intensity and duration.

- **Frequency-Domain Visualization:** Analysis in the frequency domain reveals the distribution of frequencies present in the cry, indicating the pitch and spectral characteristics of the audio signal.

- **Spectrogram:** The spectrogram offers a detailed representation of frequency intensity over time, enabling the identification of patterns and anomalies with higher granularity.

- **PCA-Reduced Feature Space Visualization:** The PCA-reduced feature space visualization condenses the multidimensional feature set into a two-dimensional context, facilitating the identification of outliers by the Isolation Forest algorithm.

- **Scaled Features Analysis:** Analyzing scaled features over time highlights deviations from the typical pattern. Outliers, marked in red, may correspond to significant moments in the baby cry that warrant further investigation.

4.7 Analysis and Discussion

The analysis began with feature extraction, capturing various characteristics such as spectral centroids, bandwidth, roll-off, zero-crossing rate, and Mel-frequency cepstral coefficients (MFCCs) from the audio frames. These features collectively provide a comprehensive understanding of the audio signal's temporal and spectral properties.

Scaling these features ensures uniform consideration in subsequent analysis steps by standardizing them to have a mean of zero and a standard deviation of one, thereby mitigating the influence of features with larger numerical ranges.

Principal Component Analysis (PCA) effectively reduces the dimensionality of the feature space, transforming correlated features into linearly uncorrelated principal components. Visualization of these components aids in identifying outliers within the new feature space.

The application of the Isolation Forest algorithm on both the PCA-reduced feature space and directly scaled features reveals outliers with respect to the spectral characteristics of the audio signal. Isolation Forest's random feature selection and split value assignment isolate anomalies effectively.

Visualizations created from these analyses provide an intuitive understanding of outlier locations within the data. A color gradient assists in distinguishing between normal data points and potential outliers, enhancing clarity in data interpretation.

In conclusion, the methods applied effectively identify outliers within baby cry audio signals. This computational approach holds promise for developing automated systems to monitor infant cries for health and distress signals. Further research may explore correlations between specific outlier patterns and medical conditions, enhancing the method's applicability in pediatric healthcare.

4.8 Future Work

Future research could expand on this framework by:

- Integrating additional audio features to enrich analysis.
- Exploring alternative dimensionality reduction techniques.
- Testing the model's performance on larger and more diverse datasets of baby cries.
- Collaborating with pediatric specialists to refine outlier interpretation and relevance to health conditions.