2019

Forecasting Anomalous Events And Performance Correlation Analysis In Event Data

Sonya Leech [Thesis]
Technological University Dublin

Forecasting Anomalous Events And Performance Correlation Analysis In Event Data

Sonya Leech

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

Date: 01-06-2019

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Date: 01-06-2019

Abstract

Classical and Deep Learning methods are quite common approaches for anomaly detection. Extensive research has been conducted on single point anomalies. Collective anomalies that occur over a set of two or more durations are less likely to happen by chance than that of a single point anomaly. Being able to observe and predict these anomalous events may reduce the risk of a server's performance. This paper presents a comparative analysis into time-series forecasting of collective anomalous events using two procedures. One is a classical SARIMA model and the other is a deep learning Long-Short Term Memory (LSTM) model. It then looks to identify if an influx of message events have an impact on CPU and memory performance.

The findings of the study conclude that SARIMA was suitable for time series modeling due to the elimination of heteroskedasticity once transformations were implemented, however it was not suitable for anomaly detection based on an existing level shift in the data. The deep learning LSTM model resulted in more accurate time-series predictions with a better ability to be able to handle this level shift. The findings also concluded that an influx of event messages did not have an impact on CPU and memory performance.

Signed: Sonya Leech

Keywords: ARIMA, SARIMA, LSTM, Anomaly, Collective, Forecasting, Time Series Modelling

0.1 Acknowledgements

I would like to express my sincere thanks to my college supervisor Dr. Bojan Bozic for his guidance, support and tremendous encouragement throughout this work. I would also like to thank Dr. Jonathan Dunne who is our companies data science architect for his outstanding guidance and support throughout this project. He has been an encyclopedia of knowledge. His continued guidance and support has helped me realize that anything is possible once you have patience, dedication and little time on your hands. Without his efforts, I would not have achieved so much. I would also like to wholeheartedly thank my partner Glenn for the endless nights he has been left on his own while I plugged away at getting through college. Not forgetting my father he has been my rock, always there to take the reigns when life gets in the way and his encouragement has helped me keep going and not give up.

Contents

List of Figures

List of Tables

List of Acronyms

ACF: Auto Correlation

ARCH: Autoregressive Conditional Heteroskedasticity

ADF: Augmented Dickey Fueller

AD: Anderson-Darling

ARIMA: Auto-Regressive Integrated Moving Average

CVM: Cramér–von Mises

CPU: Central Processing Unit

CH: Canova and Hansen

DF-GLS: DickeyFuller-Generalized Least Squares

GARCH: Generalized Autoregressive Conditional

GMM: Gaussian Mixed Model

IT: Information Technology

KPSS: Kwiatkowski Phillips Schmidt Shin

Loess: Locally weighted regression

LM: Engle's Lagrange multiplier

LB: Ljung-Box Q-Test

LSTM: Long Short Term Memory

PACF: Partial Auto Correlation

PP: Phillips-Perron

RSVM: Robust Support Vector Machines

RNN: Recurrent Neural Network

SVM: Support Vector Machines

STL: Seasonal Trend Decomposition

SARIMA: Seasonal Auto-Regressive Integrated Moving Average

SW: Shapiro-Wilk

STD: Standard Deviation

KPSS: Kwiatkowski Phillips Schmidt Shin

Loess: Locally weighted regression

LM: Engle's Lagrange multiplier

LB: Ljung-Box Q-Test

LSTM: Long Short Term Memory

PACF: Partial Auto Correlation

PP: Phillips-Perron

RSVM: Robust Support Vector Machines

RNN: Recurrent Neural Network

SVM: Support Vector Machines

STL: Seasonal Trend Decomposition

SARIMA: Seasonal Auto-Regressive Integrated Moving Average

SW: Shapiro-Wilk

STD: Standard Deviation

Chapter 1

Introduction

When unusual patterns occur in data this is classified as an anomalous event also known as an outlier. An outlier is a single extreme event. Detecting these anomalous events can be considered a support aid for a variety of different business organizations. It can be used in cyber security to aid to detect cyber attacks (Chandola, Banerjee, & Kumar, 2009). It can also be used in the financial sector for credit card fraud or the betting domain for gambling fraud. It can also aid in intrusion detection for network security or even in census data (Lu, Chen, & Kou, 2003). Being able to predict when a system or application log message is exceeding the normal operational bounds allows the IT support people become more proactive than reactive to their business process.

A collective anomaly is when more than one irregularity occurs consistently over a set amount of observations in a dataset. These collective anomalous events will fade out single point anomalies effectively reducing noise in the anomalous forecast process. These collective anomalies have been studied in time series data and LSTM models (Bontemps, McDermott, Le-Khac, et al., 2016). Our research is based on collective anomalous events.

These irregularities in the data can be identified using common measures of location like the mean or median value of a distributed dataset while traversing over those time-

series data using a rolling window (Box & Jenkins, 1970). To identify these anomalous events we introduce SARIMA, GARCH, and LSTM models. These linear, non-linear and network models capture the stationarity and volatility of the data (Box & Jenkins, 1970). These models are widely used in time series data and forecast analysis (Lasisi & Shangodoyin, 2014),(Engle, 2001).

1.1 Problem Statement

High-end applications generate thousands of log messages per minute (Jayathilake, 2012). As more applications are added to servers the volume and velocity of the data become exponential (Huang, 1998). Manually sifting through this log data to find the root cause of errors that impact on application or server performance becomes unrealistic and extremely time-consuming (Jayathilake, 2012). Having an important application go down at any given time may cost a business thousands of dollars in failed Service Level Agreements. The data is analysed to find the fault is often noisy, heterogeneous and suffering from high dimensionality. This makes finding the fault quite complex and time-consuming.

As log data is textual, this time series multivariate data suffers from a lack of labelled data. When sifting through the data it is important to grasp what factors are important to keep and which can be discarded. The parsing of the log data should be accurate while providing usable data for analysis. This labelled data is necessary to support in the aid of quick fault diagnosis and efficient relevant data retrieval.

Initial diagnosis of an event will start with a support person trawling through log data looking for a status type of " error " just before the issue occurred. With many developers working cross site on an application the standard definition of these classifications may produce a false positive. A warning classification by one person might be an error classification to another. This may lead to lengthier delays in pinpointing the fault detection due to filtering out incorrectly labelled log data.

Anomaly detection has been researched under many different titles but all leading to-

wards the same research field, example's of those are outlier detection, noise detection, exception mining, anomaly detection to name a few (Hodge & Austin, 2004). Outliers in data have a strong impact on predictions. Some outliers are defined as noise. Singh and Upadhyaya describes the noise as *"a phenomenon in data which is not of interest to the analyst, but acts as a hindrance to data analysis"*. When looking for outliers one needs to look for unusual patterns or behaviours in data. It was often the case that outliers in data were removed from a dataset to reduce noise. As more research was conducted around outliers it then became widely accepted and used to detect when a process deviates from the norm and under what conditions the deviation occurs. It became so widely popular that some business domains apply strict confidentiality to the anomalous methods used for its analysis like crime and terrorist activities (Singh & Upadhyaya, 2012).

From this comes a need for an automated anomaly detection tool (Chandola et al., 2009) that can identify rare events or behaviours in data that differ significantly from the norm. These anomalies can come in the form of point, contextual or collective anomalies (Chandola et al., 2009). Being able to track, control and understand these anomalous events can aid a business in its ability to better handle and control these events. Some of these anomalous events may impact or bottleneck the performance of a server leading to significant cost implications. Such is the case that when Amazon has an additional 100 ms delay in their response times it impacts them by a 1% reduction in sales (Ibidunmoye, Hernandez-Rodriguez, & Elmroth, 2015).

1.2 Organisation of Dissertation

This dissertation is organised as follows:

Chapter 1 gives an introduction to the background of the research and identifies the problem statement. It then identifies the scope and limitations of the research. Chapter 2 provides relevant background literature reading in the domain of time series forecasts as well as research developments within that domain. It then goes through

the research objectives and methodologies. Chapter 3 gives a brief review of the data. Chapter 4 brings the reader through exploratory analysis. Time series modelling of the data via SARIMA, GARCH and LSTM is conducted in chapter 5. Anomaly detection is covered in chapter 6. Performance analysis is reviewed In chapter 7 for CPU and memory metrics. Chapter 8 goes through the evaluation of the results. Chapter 9 contains the conclusion and future work identified.

1.3 Scope and Limitations

1.3.1 Scope

The scope of the research is to classify log event data and conduct time series models for anomaly detection using both classical and deep learning methods with a comparative analysis done on the results. Naïve Bayes and K-Mode cluster models will be implemented for classification analysis. SARIMA, GARCH and LSTM models will be implemented for time series anomaly detection analysis. Performance analysis will then be analysed on CPU, memory and disk space metrics to see if anomalous events have an impact on the performance of a server.

1.3.2 Limitations

Due to the volume of the workload, some limitations were identified.

Classification of the data was not implemented. The existing predefined severity event types within the dataset was used for classification. Those severity types were Info, Warn and Error. This is defined as a limitation as a higher level of abstraction of log event data was used whereby it would have been more appropriate to do a deeper dive classification of the different types of messages to further identify which types of textual events are causing anomalies. For example, showing that an event of type error is anomalous would not be as beneficial as showing an anomalous event of type

"connection limit reached".

Anomaly detection was only implemented on Info type event messages. Warn and Error type events were ignored from the anomaly detection analysis. This was a limitation as these types of events are only informational. It may have been more appropriate to pick warn or error type messages as these types of messages are more of an indication of a process heading towards an out of control event than that of an informational message.

Missing data was ignored and no imputations were implemented. This missing data occurred at the start of the dataset and was then filtered out so as not to be analysed. Because there was very little missing data, the limitation is minor but their needs to be a method implemented to impute missing data in the future.

Two transformations should have been implemented on the dataset to eliminate the existing level shift identified in the data. This limitation rendered the SARIMA model not suitable for anomaly detection as the data still contained seasonality. It might have been the case that the model may have performed better than that of LSTM had the 2nd transformation been done on the dataset.

Although CPU and memory were analysed from a performance perspective disk space was excluded from the analysis. This was a limitation as it was reduced from the scope of the research and it may be the case that an influx of messages might have caused the disk space to increase.

For anomaly detection, a simple two standard deviation metric was used. Three standard deviations were initially implemented but were removed from the analysis since only extreme values outside of the 99.7% confidence interval would have been captured from the Gaussian distribution of the data. It would have been better if a level shift algorithm was implemented to better detect the anomalous events within the existing shift in the data.

Chapter 2

Literature Review

2.1 Classification

Currently reviewed research papers identify in-depth studies on how to cleanse and classify data. Some of the areas of research are related to spam mail (Delany, Cunningham, & Coyle, 2005), website classification (Delany et al., 2005) and classification of emails (Youn & McLeod, 2007) but not much research can be found around the classification of application log data. Decision Trees, Random Forests, Support Vector Machines and Naïve Bayes models are useful approaches to classification using supervised machine learning algorithms. A common approach is to use Naive Bayes classification algorithm as it does not require parameter tuning and is easy to implement although Support Vector Machines (SVM) would tend to have a higher precision value than that of Naïve Bayes. (Ting, Ip, & Tsang, 2011). As the data will be continuously streamed into the model an issue may arise with unforeseen data, therefore, a new approach needs to be applied. Implementing an unsupervised clustering k-mode machine learning algorithm would be the best approach. This algorithm clusters the categorical data into partitions based on similarity (Sharma & Gaud, 2015) and has a higher degree of accuracy over that of the Naïve Bayes model. When using a cluster approach a k number needs to be identified to support the number of clusters for the

data. As this data is streamed and has high volume - identifying the proper k number would be flawed (Sharma & Gaud, 2015). We attempt to address this problem using a Gaussian mixed model (GMM) method to automatically define the number of clusters required based on the incoming data. (Celeux & Soromenho, 1996).

2.2 Time Series Models

Time series forecasting for anomaly detection needs past and present observations as an aid to help determine future values. Box and Jenkins briefly describe a stochastic process and how to make a forecast. *"A model which describes the probability structure of a sequence of observations is called a stochastic process....To make a forecast is to infer the probability distribution of future observation from the population, given a sample z of past values.* An important factor of the stochastic process is the test for stationarity. These tests are necessary because most data is not stationary by default like for example volatile stock prices. The most common models to handle non-stationary time series data is ARIMA, SARIMA and GARCH. ARIMA and SARIMA are implemented when the data is stationary or when seasonality and trend exist. The model needs to present conditional mean and constant variance (Box & Jenkins, 1970). GARCH is implemented when the data is volatile and contains heteroskedasticity by having conditional variance and zero mean (Bollerslev, 1986). The main difference with the GARCH model is that while ARIMA and SARIMA bring back actual predicted values the GARCH model uses the difference of the data to determine a prediction variability value. [1] The GARCH model in its own right is more suited for economic type data (Engle, 2001) and is only suited if you are looking for a variance prediction value. To determine which model to use the data would need to be analysed for trend, seasonality and volatility before an assumption can be made. These classical models are least square models with its performance analysed using the residual errors of the model.

[1] With time series data the percentage difference or the variance of the data points are used to draw a variance prediction value from GARCH.

LSTM networks are a type of Recurrent Neural Network (RNN) that can be used as an approach for anomaly detection. Using unsupervised clustering of the data LSTM learns the relationship between past and current data, using learned weights [(Hochreiter & Schmidhuber, 1997), (Bontemps et al., 2016)] which can then model and capture normal behaviour. Gaussian assumptions can then be used to determine if the predicted values are anomalous by smoothing past errors and comparing them to new errors.

2.3 Stationarity

A stationarity process is also known as a stochastic process. It is when the properties of the time series do not change over time. To be strictly stationary the distribution of the time series data needs to be unaffected by any shift of (n=?) times plotted along the axis of the time series data. Data can be tested for stationarity by looking at its variance, mean and autocorrelation function. Another term for the autocorrelation function is the Spectral Density function (Box & Jenkins, 1970). A constant mean (m=1) is an indication that the time series is stationary if this value holds throughout all times within the time series data. A constant variance which measures the spread of the time series is another indication of stationarity. A histogram can be plotted to determine the shape of the data for its variance. The shape of the time series if stationary would contain a probability distribution of the time series as a Multivariate Normal Distribution. This would be represented as a Gaussian process which when plotted in a histogram contains a Gaussian bell-shaped curve. A weak stationarity process is also known as covariance stationarity is where the variance in the time series does change with time. Another test can be implemented using the periodogram. The periodogram uses the Spectral Density Function. It uses sine and cosine waves with different frequencies of the time series data. It is used to check the randomness of the residuals of a time series after fitting a model to the data (Ashot Vagharshakyan, 1999). It was first used by Schuster in 1898 (Schuster, 1898).

A unit root is a collection of random variables indexed by points in time series data. This is also known as a stochastic process (non-deterministic random process). A Unit root tests to see if a shock in the data has a permanent effect. If unit root exists then the time series is deemed not stationary.

There are different types of unit root tests:

1. Augmented Dickey-Fuller
2. Phillips-Perron
3. Kwiatkowski Phillips Schmidt Shin
4. Elliott Stock Rothenberg ADF-GLS

ADF

With ADF its null hypothesis is that there is unit root which implies non-stationarity. It was developed by David Dickey and Wayne Fuller in 1979 (Dickey & A. Fuller, 1979). As part of its computation, it fits the regression model by Ordinary Least Squares (OLS) starting at the lag of the first difference. It tests for an independent normal random variable with a mean and variance of zero. If $P < 1$ it implies stationarity with a limiting distribution of normality. If $P > 1$ it implies non-stationarity with a limiting distribution called Cauchy. If $P = 1$ it assumes a random walk and a transformation would need to be done on the time series data (Dickey & A. Fuller, 1979). If the p-value is significant it recommends using the ADF Test Statistic. This test should be used with caution as it has a high Type I error rate. It also suffers from a "near observation equivalence" problem as it cannot distinguish between true unit-root processes of 0 and near unit-root processes that are close to zero.

PP

PP is a non-parametric unit root test. Its null hypothesis is that a time series is integrated to the order of 1 which is non-stationary as it has been first differenced. It supports weakly dependent and widely dissimilar distributed data and its time series models do not need to be stationary. It looks at drift or drift and a linear trend. One of its assumptions is that its sequence of innovation is 0 for all time series (Peter

C. B. Phillips, 1988). The PP test is a variation of the DF test. To allow for serial correlation its test statistic is based on a regression line without any modification. When computing the test statistics to ensure that serial correlation has no impact on the widely dissimilar distributions a heteroskedasticity and autocorrelation consistent estimator (HAC) is used.

KPSS

KPSS test was developed to give more grounding in unit root tests. Shin, Kwiatkowski, Schmidt, and Phillips defined the test as *"We propose a test of the null hypothesis that an observable series is stationary around a deterministic trend. The series is expressed as the sum of the deterministic trend, random walk, and stationary error, and the test is the LM test of the hypothesis that the random walk has zero variance."*

ADF-GLS

ADF-GLS is a modification to the ADF test. It aims to have more power than that of the ADF test when an unknown mean or trend is present. ADF-GLS is first estimated by a generalized least squares (GLS) model followed by a DF test to test for unit root. Elliott, Stock, and J. Rothenberg cited that *"Employing a model common in the previous literature, we assume that the time series data were generated where dt is a deterministic component and vt is an unobserved stationary zero-mean error process whose spectral density function is positive at zero frequency"*. Its initial experiments confirm that it works well when the sample size is small (Elliott et al., 1996).

To identify patterns in data, techniques that can be used are smoothing, fitting a curve or running an Auto Correlation plot on the time-series data. Pattern identification can be accomplished by looking at a sequence of values that follow an order that is not random ie it does not happen by chance. Being able to extrapolate these patterns allow us to better predict for the future. A trend pattern would have a linear positive or negative gradient. The smoothing operation functions are the moving average function, the median function or the exponential weight function. The moving average function will have a set window size example (s=7) that will return the moving average of the time-series. This would only repeat itself every 7 points in the time-series. Statsoft

recommends *"Medians can be used instead of means. The main advantage of median as compared to moving average smoothing is that its results are less biased by outliers (within the smoothing window). Thus, if there are outliers in the data (e.g., due to measurement errors), median smoothing typically produces smoother or at least more "reliable" curves than moving average based on the same window width. The main disadvantage of median smoothing is that in the absence of clear outliers it may produce more "jagged" curves than moving average and it does not allow for weighting."* Smoothing out the data removes noise and cancels the outliers in the data. Robert J Hyndman recommends that for non-seasonal data the model parameter should be 10 and for seasonal data the parameter should be 20.

2.4 Seasonality & Trend

Seasonality in economics has been defined by Hylleberg as *"Seasonality is the systematic, although not necessarily regular, intra-year movement caused by the changes of the weather, the calendar, and timing of decisions, directly or indirectly through the production and consumption decisions made by the agents of the economy"* . Seasonality is an important factor in time series modelling as to exclude this factor leads to building an inaccurate forecasting model. Peart also describes its importance as *"Every kind of periodic fluctuations, whether daily, weekly, quarterly. or yearly must be detected and exhibited not only as a subject of study in itself but because we must ascertain and eliminate such periodic variations before we can correctly exhibit those which are irregular or non-periodic and probably of more interest and importance"* Peart.

Seasonal Decomposition is where the data has been decomposed into seasonality, trend and remainder components. This process is called Seasonal Adjustment or Deseasonalizing. STL is a Seasonal Decomposition that uses a set of sequential smoothing operations based on Loess (Locally Weighted Regression) smoother. The eigenvalue and frequency analysis results determine which part of the data is trend and

seasonality. To identify trends and patterns within the dataset it removes the seasonal patterns. STL supports missing values by a using a dependant and independent variable (x, y). It uses a regression of (x) which is a smoothing of (y) therefore allowing x to be computed for any value (x) along the independent variable scale (Cleveland & Cleveland, 1990). It also handles data diverging from normality also known as aberrant data through its computation of using an inner loop that is nested in an outer loop (Cleveland & Cleveland, 1990). As each iteration passes through the inner loops it updates the seasonal and trend components. The outer loop calculates a robustness weight. If the time series point diverges from normality and results in a high residual value it then passes a zero weight back to the inner loop which will then be used as part of the inner loop computation. The residuals will show if variability exists within the time-series data. Parameters defined in seasonality are frequency and periodicity. The seasonal dummy variable for periodicity for month is (n=12), for day (n=365) and (n=24*365) for hourly. The frequency is defined by the aggregation of the data.

Seasonal Decomposition is where the data has been decomposed into seasonality, trend and remainder components. This process is called Seasonal Adjustment or Deseasonalizing. STL is a Seasonal Decomposition that uses a set of sequential smoothing operations based on Loess (Locally Weighted Regression) smoother. The eigenvalue and frequency analysis results determine which part of the data is trend and seasonality. To identify trends and patterns within the dataset it removes the seasonal patterns. STL supports missing values by using a dependent and independent variable (x, y). It uses a regression of (x) which is a smoothing of (y) therefore allowing x to be computed for any value (x) along the independent variable scale (Cleveland & Cleveland, 1990). It also handles data diverging from normality also known as aberrant data through its computation of using an inner loop that is nested in an outer loop (Cleveland & Cleveland, 1990). As each iteration passes through the inner loops it updates the seasonal and trend components. The outer loop calculates a robustness weight. If the time series point diverges from normality and results in a high residual value it then passes a zero weight back to the inner loop which will then be used as part of the inner loop computation. The residuals will show if variability exists within

the time-series data. Parameters defined in seasonality are frequency and periodicity. The seasonal dummy variable for periodicity for month is (n=12), for day (n=365) and (n=24*365) for hourly. The frequency is defined by the aggregation of the data.

Hylleberg defined different types of seasonal adjustment methods:

1. X11 Method
2. Unobserved Component Models For Seasonal Adjustment Filters
3. Model-Based Estimating Structural Models Of Seasonality
4. Model-Based ARIMA Models
5. Model-Based Periodic Variance
6. Model-Based Box-Jenkins

All of the above models have not been defined in the literature but have been documented as evidence that different tools exist.

X-11 procedure was developed by Julius Shiskin in the 1950's (B. Q. Dominique Ladiray, 2001). It is based on a single time series using a sequential moving average filter. It was developed to support seasonal adjustment and decomposition of monthly and quarterly series. Its components consisted of a seasonal component, a combined trend and cycle component, a trading day component which looks at the composition of the day-of-the-week at a month and quarter time series. Another one of its components measures the effect of the Easter holidays and finally an irregular component that covers all the other fluctuations not picked up by the other components. (Ladiray & Quenneville, 2001). X-11 models are "Additive" and "Multiplicative". The difference between both models is that the Additive Model adds the components together and the Multiplicative Model multiplies the components together.

Additive Model = Ct + St + Dt + Et + It
Multiplicative Model = Ct * St * Dt * Et * It

It components are represented as:

1. C: Cycle
2. S: Seasonal,
3. D: Day of Week
4. E: Easter Holiday
5. I: Irregular

Box-Jenkins is a combination of an AR and MA model. It was developed by Box and Jenkins (George E. P Box, 1976). AR is the Autoregressive model and MA is the Moving Average model. Combined it is known as the ARMA Model. An assumption with using Box-Jenkins is that the data is stationary, ie constant in mean and variance for all values in the time-series. If there is seasonality in the data then Box-Jenkins can support seasonality by using the SARIMA Model. Box-Jenkins identifies seasonality using autocorrelation and partial autocorrelation correlogram. The correlogram looks at the correlation between different lags on the time-series. It looks at the current period against past periods to determine if seasonality exists (t-1). The first lag on the plot is an autocorrelation onto itself and as such should be ignored. The partial autocorrelation looks at the moving average value from the time-series data.

Portmanteau Test can be used to statistically determine if there is a correlation in the time series data. It looks at the residuals of the model to test for correlation (Jennifer Castle, 2010). The Ljung–Box and Box–Pierce are different types of Portmanteau tests. G. M. Dominique Ladiray Jean Palate and Proietti commented that *"The detection of the various periodicities must be done before any modelling of the time series. Among the statistical tools that can be used in this respect, the most efficient are certainly: the spectrum of the series, the Ljung-Box test and the Canova-Hansen test"*. When using a Box-Jenkins (George E. P Box, 1976) approach a minimum dataset would be of no less than 50 observations but a recommendation would be 100.

Canova-Hansen

Canova and Hansen is a statistical test to see if there is seasonality in the data against that of the null hypothesis which implies unit root exists at all of the seasonal frequencies except zero. (Taylor, 2003)

2.5 Goodness Of Fit Tests

Testing for the goodness of fit determines whether the model selected is the best fit model for the data. Data can either be parametric or non-parametric. With parametric data, we make an assumption or an inference about the parameters of the data based on a sample of the population which is then used as estimated model parameters if all assumptions hold. Non-parametric data assumes that the data does not have a normal distribution, has no characteristical structure and that all assumptions don't hold. In statistical models to understand which values to assign to a parameter either an Ordinary Least Square (OLS), a Methods Of Moments (GMM) or a Maximum Likelihood method can be used.

GMM

GMM was developed by Karl Pearson in 1894 (Encyclopedia.com, n.d.) and published in his journal "Biometrikia" in 1936 (Fisher, 1937). GMM can be used in both parametric and non-parametric data. Using population moment conditions we can understand the variance, mean, skewness and kurtosis of the population. This allows us to understand the shape of the distribution of the data and from this, we can estimate the parameters of the data for the model under certain moment conditions (Wooldridge, 2001).

MLE

MLE became very popular in 1992 by Ronald Fisher (Aldrich et al., 1997). With MLE the goal is to find the best way to fit a distribution of any type to the data. For

example, any distribution that is normal once fitted to an experiment with that same distribution should have the same symmetrical shape with no skewness. All the data should lie around the mean value. Any data that does not fit the shape is then considered to be part of different distribution and the likelihood of predicting the values is low.

Normality Tests

Normality tests are important when you want to ascertain the confidence interval. For time series data to be of normal distribution the data from the quantile plot should fit along the regression line. Any data that drifts further from the regression line indicates that there is uncertainty that the data is normally distributed. There are different ways to detect that time series data is of a normal distribution. Quantile Plots, Box-Plots and histograms are some good visual diagnostic aids. Although these graphical aids are good indicators of normality to be truly sure of your findings - strong statistical tests should be conducted that tests whether the data is of a normal Gaussian distribution. These statistical tests are Skewness, Kurtosis, Shapiro-Wilk, Anderson-Darling and Cramér–von Mises.

For skewness, the value should be zero and for kurtosis, the distribution of the data should be equal to 0 or for excess kurtosis should be equal to 3.0. A greater than 3 kurtosis indicates a heavy-tailed distribution and a kurtosis less than 3 has a light tailed distribution (Mohd Razali & Yap, 2011).

The Shapiro-Wilk test is a left tailed test. On initial development, it only showed good results when the data size was n $<$50. This was enhanced further with the implementation of the AS R94 algorithm which effectively allowed it to support a larger sample size (Mohd Razali & Yap, 2011).

Anderson-Darling is an enhancement of the Cramér–Von Mises test, it focuses more

on the tails of the distribution. During the running of the test, it calculates the critical values for each distribution(Mohd Razali & Yap, 2011).

Time Series Dependence Test

When modelling data the time series needs to be independent. This is important as if the time series data is dependent on other time lags then this means that trend or seasonality still exist in the model and as such may produce inaccurate predictions. To test for independence a Ljung–Box can be implemented.

The Ljung–Box test is a portmanteau test which checks the residuals of the ARIMA model for white noise. Its null hypothesis is that data is independently distributed. The alternative hypothesis is that the data is not independently distributed and exhibits serial correlation. It tests the overall randomness based on the number of lags defined which is different from testing for randomness at each distinct lag (Ljung & Box, 1978). A residual ACF test is deemed more powerful than that of a Pearson test.

Heteroskedasticity - ARCH Test

Heteroskedasticity is when the errors of the model are not constant over time. Over time the errors span out in range (Bollerslev, 1986). This makes the prediction volatile. GARCH models treat heteroskedasticity as a variance to be modelled (Engle, 2001). Engle's LaGrange Multiplier test can be used to test for heteroskedasticity.

2.6 Anomaly Detection

Anomalous detection can be implemented by looking at points in time. A single point that is distant from the majority of observations can be considered an anomaly. Considerations need to be taken to decide under what conditions a deviation is classified as an anomaly. Different classifications can be implemented, those are point, collective and contextual (Singh & Upadhyaya, 2012).

Two types of outliers are discussed. Those are the Additive Outlier (AO) (Fox, 1972) and the Innovational Outlier (IO) (Balke, 1993). The AO outlier occurs over a single observation like a point in time. It is something that may occur due to random chance. The IO outlier is identified when it remains an outlier over several observations. It does not drop back to a normal value until some time has passed (Tsay, 1988).

Changes in the structure of data can also be an outlier. Different types of changes exist. Three of these structures are discussed in the paper. Those are the Level Shift(LS), the Variance Change (VC) and Transient Change(TC) (Tsay, 1988).

LS in time series data is when the data abruptly changes and remains at that abrupt change until a constant amount of time has lapsed (Balke, 1993). VC is when the variance of the data changes over time. Transient changes occur over time like a stepping change or a gradual slope change. To identify these structural changes in the data one would need to read in all the data as a batch process. Another process to detect outliers is to use a sequential iteration over the time series data. Tsay describes the importance of not ignoring these changes in the data *"Outliers, level shifts, and variance changes are commonplace in applied time series analysis. However, their existence is often ignored and their impact is overlooked."*

2.7 Model Evaluation

Before data can be modelled, the model parameters (p,d,q) need to be defined. These model parameters can be used as an aid to define which model to use. Mehdiyev, Enke, Fettke, and Loos comments that *"Some methods indicate superior performance when error based metrics are used, while others perform better when precision values are adopted as accuracy measures"*.

RMSE has been criticized as being heavily misinterpreted and should be removed from the literature since it is calculated based on the variance of three measures than that of one. Those measures are the distribution of the error magnitude, the average error

magnitude and the square root of the number of errors (Willmott & Matsuura, 2005). This theory is rejected by (Chai & Draxler, 2014) who say that RMSE should be used when the errors of the model are Gaussian. MAE and MSE have been used by Babu and Reddy on their hybrid ARIMA and ANN models. MSE was used by Shipmon, Gurevitch, Piselli, and Edwards on their time series anomaly detection paper.

2.8 Summary of Literature

SARIMA models are better suited when the data contains trend or seasonality and GARCH models work better under conditions of volatility. GARCH models have shown high prediction success rates in finance but the model predicts variance for each error term (Engle, 2001) and may not be suited for anomaly detection when looking for anomalous events in counts of application log data. Siami-Namini and Namin conducted a test on LSTM and ARIMA in time series data. The results of the test confirmed that LSTM outperformed ARIMA by 85% and that setting the value epoch = 1 generates a reasonable prediction model.

Stationarity is important in time series modelling for ARIMA and GARCH. The ADF test suffers from high Type I error rates. The ADF test is the most commonly used test (Davidson & Mackinnon, 2012). The PP test was found to have poor performance on a finite sample size (Davidson & Mackinnon, 2012). ADF-GLS is a modification to the ADF test. Its initial experiments confirm that it works well when the sample size is small (Elliott et al., 1996). Davidson and Mackinnon documented that ADF-GLS has more advantage over the ADF test. The KPSS test was developed to give more grounding in unit root tests (Shin et al., 1992).

Understanding how well data is distributed can be used to determine the approximation of a value based on its normal distribution. There is much goodness of fit tests. A statistical skewness test with a value of zero ascertains a normal distribution. A kurtosis test should have a distribution value of zero or for excess kurtosis should be equal to 3. A Chi-Square test is used when the data is categorical. Mohd Razali and

Yap conducted tests on SW, AD and CVM to see which test performed the best. The outcome of his results concluded that SW was the best performing test with the AD test coming a close second. Testing for normality is important but it may be hard to pass a normality test because with large sample sizes any kind of deviation from normality will cause the data to become non-normal.

Lasisi and Shangodoyin studied outlier detection on airport data. They looked at Innovation (IO), Level Shift(LS), Additive (AO) and Transient Change (TC) Outlier algorithm's on ARIMA Models. Their findings concluded that combined usage of AO, LS and TC captured 60% of their outliers with LS producing the best results.

2.9 Gaps in Literature

It is very hard to identify gaps in the literature. This area of research is heavily researched. There are many aspects to time series modelling and anomaly detection. We can see from the literature that many avenues need to be explored to aid in determining the right techniques and tools to use. The more you dive into the research the more avenues that open up. Although no gaps in the literature have been defined, this is not to say that they do not exist. Further research should be conducted on one element of the topic to identify gaps in the literature.

2.10 Research Aim and Objective

2.10.1 Research Question

Is K-Mode clustering able to classify log event data with a better accuracy rate than that of Naïve Bayes? Can LSTM outperform SARIMA or GARCH by forecasting a better prediction accuracy measure to support anomalous events? Does an anomalous event harm a servers CPU, memory or disk space usage?

2.10.2 Research Aim

This research aims to look at how well different classification models perform against each other on the same dataset. Once classified we then look to compare and contrast on how well classical and deep learning models predict anomalous events. A question then needs to be answered to see how correlated anomalous events are with that of performance metric data and is that correlation positive or negative.

2.10.3 Objective

Labelling data correctly can lead to shorter resolution times in fault diagnosis. Implementing Naïve Bayes and K-Mode clustering will show a difference in terms of accuracy of its labelling of the data into different classified groupings. The evaluation of the test will provide evidence as to which model is better suited for the given data.

Manual streaming of large datasets can be cumbersome when looking for fault detections when an anomalous event occurs. Can this be better addressed using a deep learning network model like LSTM and how does it compare against a more classical model like SARIMA or GARCH?

Understanding if an anomalous event of type a has, for example, a 10% higher impact on CPU, memory or disk space performance than that of an anomalous event of type b helps weight and prioritise anomalous events mean time to resolution. This research should provide evidence as to the strength of the relation between the event type messages and the CPU, memory and disk space usage. This allows us to provide evidence as to whether these anomalous events have an impact on server performance.

The objective of the research is to provide evidence as to the strength of the correlation on a server's performance under the condition of an anomalous event. An anomalous event is composed of some pre-defined rules and modelled using both deep learning and classical models.

2.10.4 Research Methodologies

With 4 months of data time series SARIMA, GARCH and LSTM models will be implemented to identify the best-suited model for anomalous events in log event type data. CPU, disk space and memory metrics will then be used to determine if these anomalous events have either a positive or negative impact on the performance of a server. The goodness of fit tests on these time series models will be tested on residual errors. RMSE will be used to identify the best (p,d,q) parameters for model prediction. The detection of anomalous events will be based on two standard deviations. Exploratory analysis will be conducted on daily and hourly data.
This study can be summarized as follows:

- Exploratory analysis will be conducted on event type messages of Info, Warn and Error for daily and hourly data using descriptive statistics.
- Pearson's correlation analysis will be conducted on the three event type messages to see if they have a linear correlated relationship with each other.
- Distribution analysis will be conducted on the data to determine its shape using statistical tests of skewness and kurtosis. Graphical histograms and quantile plots will also be used to conclude as to whether the data is of a Gaussian distribution.
- Unit Root tests will be conducted on the data using ADF and KPSS tests.
- Seasonal and trend analysis will be conducted using auto correlation and partial autocorrelation functions. STL and CH tests will also be used.
- Transformations will be done on the data to remove trend and seasonality if identified in the analysis. Those transformations are a natural log, 1st difference and square root.
- An LM test will be conducted on the residuals of a GARCH (1,1) model on all transformed and non transformed data to confirm if a heteroskedastic ARCH effect exists in the data.

- A SARIMA model will be implemented for time series prediction. This model will support seasonality and trend. The auto ARIMA function will iterate through each of the (p,d,q) parameters to identify the best fit model with the lowest AIC and RMSE value. The lowest RMSE of the iterated model will be used to identify model parameters.
- The residual errors of the SARIMA model will be analysed.

  - Auto correlation will be plotted on the residual errors to confirm if there are any lags outside of the confidence interval limits.

  - A Ljung Box test will be conducted on the residual errors to determine time dependency.

  - Normalcy tests will be conducted using AD, Shapiro-Wilks and CMV tests.
- An LSTM model with parameters of a repeat iteration of 10 using an epoch size of 1, 10, 50 and 100 will be tested for time series predictions. The test with the lowest RMSE will be used for time series predictions.
- Two STD values will be used for both SARIMA and LSTM prediction models for anomaly detection.
- Pearson's correlation test will be conducted on event type messages against that of the percentage of CPU used, free memory and disk usage.

Chapter 3

Data Understanding

For this research log event, data will be captured. This data will come from a Kafka application log server. The data will be pulled every hour, parsed, cleansed, aggregated and fed into a Mongo database. Disk usage, CPU and memory information will be pulled hourly from a graphite server for the performance metrics. This data is already summarized so no aggregation will be necessary. For the initial exploratory analysis, a dashboard will be created and hosted on a local server. The dashboard will be made up of D3 time series charts. These charts will be near real-time as they will be fed from the automatic pushes of the aggregated data pulls. All code will be developed in python. For LSTM modelling the model will be implemented on Keras which sits on top of Tensor flow through the python Jupyter notebook. An architectural diagram of the platform is shown in figure 3.1

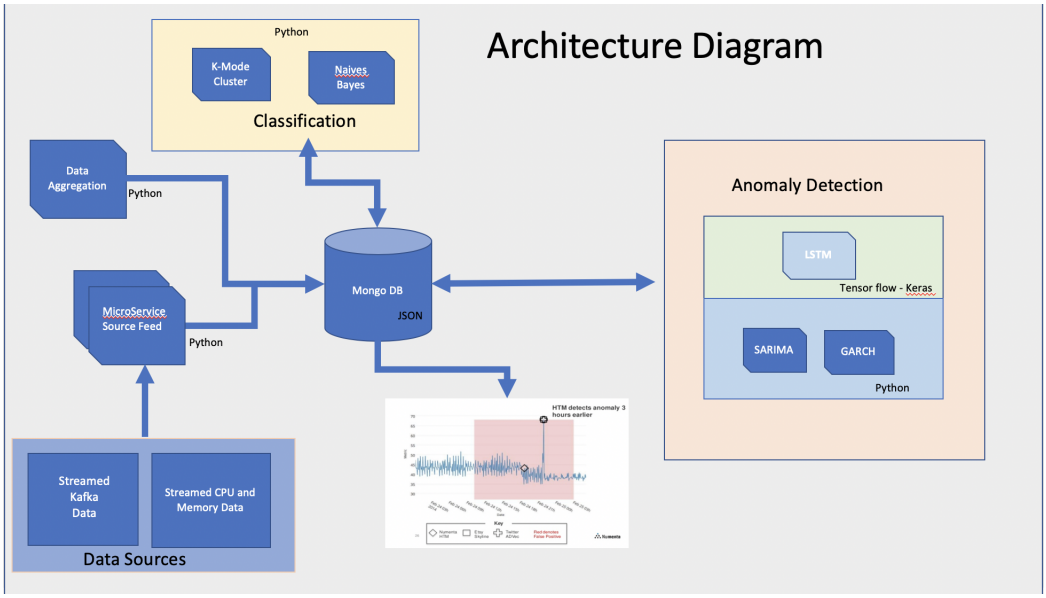

Figure 3.1: Anomaly Detection Architectural Diagram

3.1 Feature Identification

Kafka application log data was analysed. Exploration of the data concluded that there were only three types of log levels active on the server. Those were info, error and warn. No other log level severity type like debug or trace was found within the logs. For the performance metrics memory_percent_free and cpu_pct.use variables were identified.

Features identified:

- Info
- Error
- Warn
- memory_percent_free
- cpu_pct.use

- Timestamp

3.2 Data Cleansing

Data parsing will be done on the textual log messages to parse out the timestamp and severity type. Deeper parsing of the textual message itself will be conducted to bring back only the first 100 characters. For data cleansing, any row with no timestamp starting with 2018 or 2019 within the first set of characters will be removed from the dataset. All words will be converted to lower case. All stop words, punctuations, white spaces and numbers will be removed from the data set.

3.3 Missing Data

Initial observations identified that 66 individual hours of data was missing which equates to 2.75 days of data. The missing data occurred around the same time interval and was not widely dispersed throughout the data set. No imputation was implemented for this missing data. This missing data were included in the analysis for daily exploration but was excluded for hourly.

Chapter 4

Exploratory Analysis

An exploratory analysis was first conducted on daily data. The data analysed was from 21/12/2018 - 26/02/2019. Hourly data from 01/01/2019 - 13/04/2019 was then analysed and used for the remainder of the research objectives as specified in chapter 2.

Figure 4.1 shows the time series dashboard that was created. It is used to visually inspect if there is any correlation between any of the different event type messages. The dashboard has been filtered to show a subset of the data from 11th January to 5th February because the 20th of January was the highest producer of log message events during the initial exploratory analysis phase.

Figure 4.1: Exploratory Analysis Dashboard

4.1 Daily Analysis

For the initial two months of the data table, 4.1 shows that a total of 1.5 million event type messages were produced. Out of those messages the error type events produced the highest number of events equating to a total of 55.5%.

	Total	**Percent**
Total	1,574,682	100%
Info	560,828	35.62%
Error	874,336	55.52%
Warn	139,518	8.86%

Table 4.1: Count of Aggregated Log Data

For any analysis herein the alpha will be 0.05

4.1.1 Normality

Distribution analysis was done for each of the severity type messages. The first graph to the left in figure 4.2 shows info type messages not being of a normal distribution. The graph displays a platykurtic kurtosis with positive right-tailed skewness. Its quantile plot underneath it does confirm that the data is not normally distributed but does observe some fitting on the regression line. We also observe some outliers in the data. The middle and right graphs which show the warn and error distributions indicate a very volatile dataset due to the high volume of low counts of messages and a small volume of high count messages. The three quantile plots show that the data is not of a normal distribution for each of the severity type events.

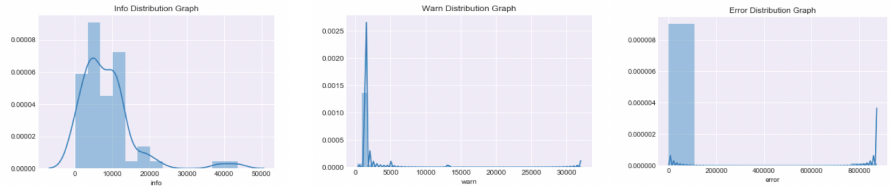

Figure 4.2: Info, Warn, Error Daily Distribution Analysis

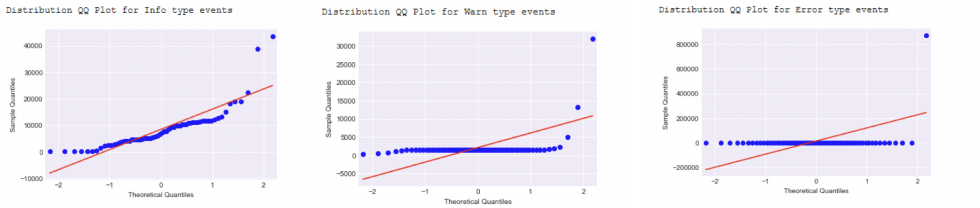

Figure 4.3: Daily Quantile Plots

For data to be of normal distribution its skewness should be zero and its kurtosis should be three. As per table 4.8 info, warn and error do not conform to the skewness and kurtosis values to be of a normal distribution.

	Skewness	**Kurtosis**
Info	2.5	9.1
Warn	6.7	48.7
Error	34.7	1261

Table 4.2: Daily Skewness-Kurtosis

SW and AD normalcy goodness of fit tests were conducted on the data.

Log Type	**Test**	**Test Statistic**	**P Value**
Info			
	SW	0.7	0.0
	AD	3.2	1.0
Warn			
	SW	0.1	3.2
	AD	21.6	1.0
Error			
	SW	0.0	0.0
	AD	585.9	1.0

Table 4.3: Daily Goodness Of Fit Tests

SW Test

Null Hypothesis: The data is normally distributed.

Alternative Hypothesis: The data is not normally distributed.

If p-value < 0.05 reject the null hypothesis. The data is not normally distributed.

AD Test

Null Hypothesis : The data is normally distributed.

Alternative Hypothesis: The data is not normally distributed.

Critical values [10%: 0.62, 5% : 0.74, 1% : 1.03]

If test statistic > critical values : Reject the null hypothesis the data is not normally distributed.

Normalcy Results

Info :

We reject the null hypothesis of the SW test p=0.0. There is statistical evidence to suggest the data is not of a normal distribution. With the AD test (test statistic =3.2 > 5% at 0.74) we reject the null hypothesis. The data is not normally distributed.

Warn :

We fail to reject the null hypothesis of the SW test p=3.2. The AD test (test statistic =21.6 > 5% at 0.74) is showing strong evidence to suggest that the data is not normally distributed.

Error :

We reject the null hypothesis of the SW test p=0.0. There is statistical evidence to suggest the data is not of a normal distribution. The AD test (test statistic =585.9 > 5% at 0.74) is showing strong evidence to suggest that the data is not normally distributed.

4.1.2 Unit Root Tests

Unit Root tests were conducted.

ADF

Null Hypothesis : Data has unit root (implies not stationary)

Critical Values : [10%: -2.59, 5%: -2.90, 1%: -3.53]

P value < 0.05 : Reject the null hypothesis. The data is stationary.

If test statistic < critical values. Fail to reject the null hypothesis the time series has unit root and is not stationary.

KPSS

Null hypothesis : The data is stationary and does not have unit root.

KPSS critical Values : [10% : 0.34, 5% : 0.46, 1% : 0.73]

If test statistic < critical value : Fail to reject the null hypothesis. The data does not contain unit root and is stationary.

Critical values for all tests:

ADF Critical Values : [10% : -2.59, 5% : -2.90, 1% : -3.53,]

KPSS Critical Values : [10% : 0.34, 5% : 0.46, 1% : 0.73]

	Test Statistic	**P Value**
ADF	-2.21	0.20
KPSS	0.38	0.08

Table 4.4: Daily Info Unit Root Values

	Test Statistic	**P Value**
ADF	-8.15	0.00
KPSS	0.12	0.10

Table 4.5: Daily Warn Unit Root Values

	Test Statistic	**P Value**
ADF	-8.06	0.00
KPSS	0.1	0.10

Table 4.6: Daily Error Unit Root Values

Results: Unit Root Tests

Info :

In the ADF test (p=0.20, test statistic (-2.21) $<$ critical value (-2.90)). We accept the null hypothesis. The time series has unit root and is not stationary. For the KPSS test (test statistic (0.38) $<$ critical value (0.46)) therefore we fail to reject the null hypothesis, the data is stationary.

Warn :

We reject the the null hypothesis for the ADF test (p=0.00, test statistic (-8.15) $>$ critical value (-2.90)). The time series has no unit root and is stationary. For the KPSS test (test statistic (0.12) $<$ critical value (0.46)). We fail to reject the null hypothesis the time series is stationary.

Error :

We reject the the null hypothesis of the ADF test (p = 0.00, test statistic (-8.06) $>$ critical value (-2.90). The time series has no unit root and is stationary. We fail to reject the null hypothesis for the KPSS test (test statistic (0.1) $<$ critical value (0.46)). Which implies the time series is stationary.

Mean and Variance Analysis

For forecast analysis, the data needs to be stationary in mean and variance for it to fit an ARIMA Model. The data was split into 2 random samples. Mean and variance tests were conducted on both samples.

	Info	Error	Warn
Mean 1	10227	2428	19007
Mean 2	4720	1439	0
Mean Diff	5507	989	19007
Variance 1	73815141	23224782	16608359
Variance 2	1605014	14	0
Variance Diff	72210127	23224768	16608359023

Table 4.7: Daily Mean-Variance Analysis

Table 4.7 shows that the data is not stationary in mean and variance as there are significant differences in mean on both samples for each severity type message. This is the same for the variance test.

4.1.3 Seasonality & Trends

Trend and seasonal graphs were created for info, warn and error type events. STL decomposition was done with the frequency set to weekly using the additive model. A monthly period was ignored due to the lack of initial data for analysis.

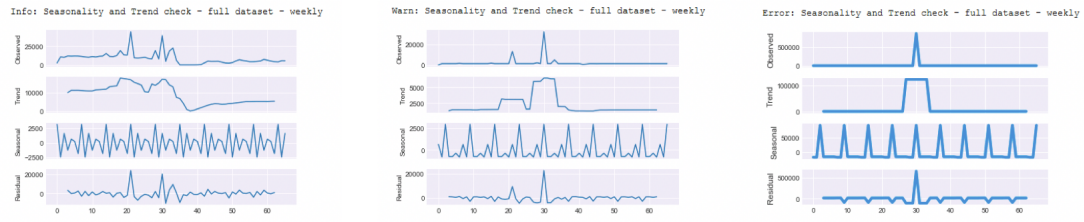

Figure 4.4: Daily Seasonal Decomposition Analysis

As per table 4.4 we visually observe that trend and seasonality do exist in the dataset. The graphs are displayed in order of observed, trend, seasonality and residuals. The trend is shown in the 2nd graph of the grouped graphs. For trend info type events

do show a variance change while warn events to show a transient type change and error events show the same as warn but not as apparent. Seasonality is shown in the third row of the grouped graphs and there does seem to be a repeat pattern over the time series. These patterns become more apparent when higher levels of frequency are used. A correlogram was also created to identify trends in the dataset.

Auto correlation - Partial Autocorrelation

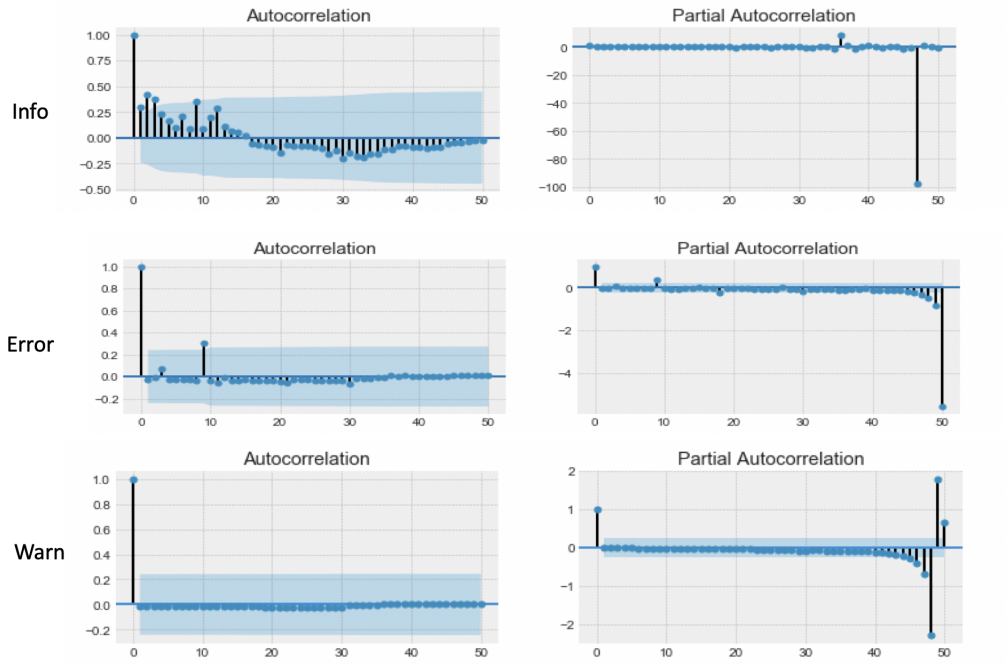

Figure 4.5: Daily ACF PACF

For the correlogram, the first fifty lags were used. This gave fifty data points within the time-series to be tested for correlation and trends. We observe from the Info correlation chart that Info type events do show trend in the dataset while warn and error do not show any trends.

A statistical CH test was conducted to see if the data contained seasonality with the results concluding that there was no evidence of seasonality or trends in the dataset. An ARIMA difference utility test using ndiff was implemented to see how many times

we difference the data to remove trend. The result of the test indicated that there was no trend in the data set for any of the different types of severity messages. As such we conclude that there is no statistical evidence to suggest seasonality or trend exist but this may be taken with caution due to the graphical evidence presented.

4.1.4 Correlation

Pearson's correlation analysis was implemented on the daily data to see if any of the event types have any type of relationship with each other. It is observed from figure 4.6 that info type events have a very strong correlation with warn events (0.7). Info events also have a strong correlation with error events (0.5). Error and warn events do show a significant correlation with each other of (0.9). The results of the Pearson's test conclude that there is strong statistical evidence of relationships between each of the different event types.

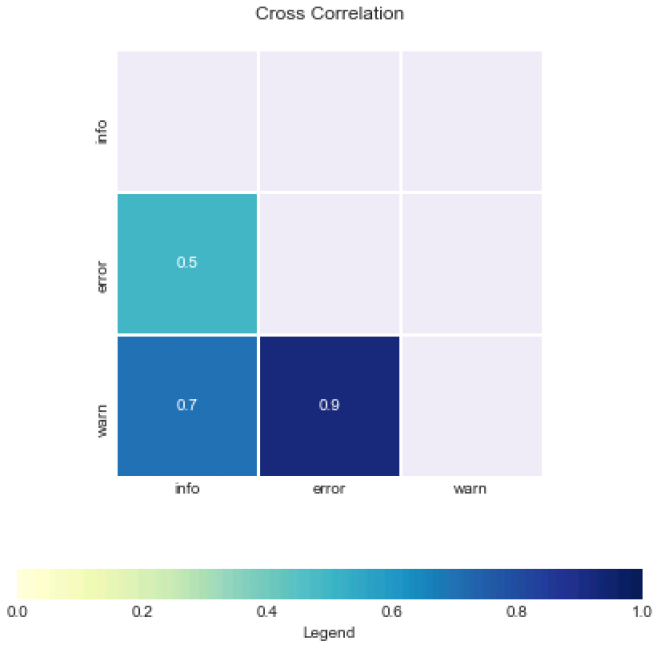

Figure 4.6: Daily Pearson Correlation Test : Info : Warn : Error

4.2 Hourly Analysis

The hourly analysis was conducted on the event type data.

4.2.1 Normality

Histograms and quantile plots were graphed on the hourly data to indicate as to the data's distribution and shape.

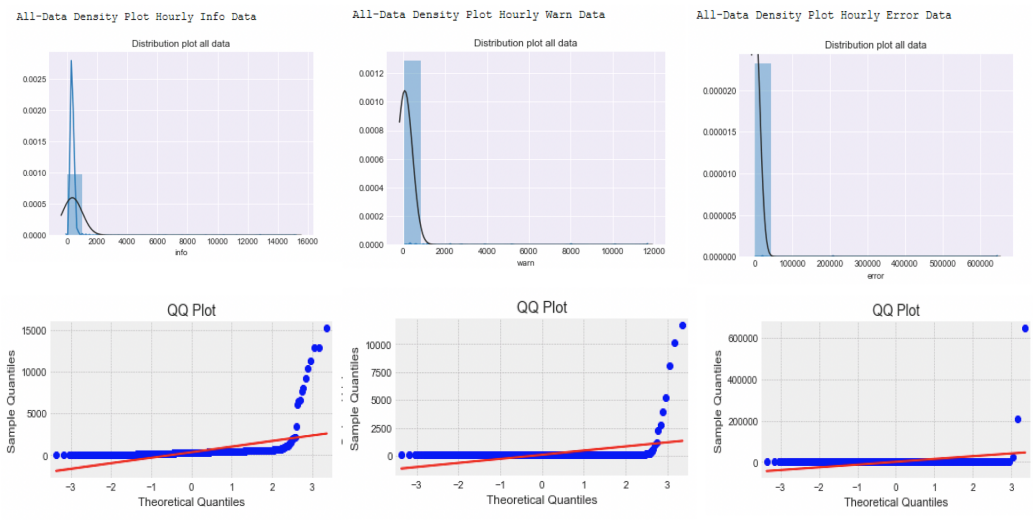

Figure 4.7: Hourly Distribution Analysis

Figure 4.7 shows that the data does not conform to a normal distribution as the data is not a symmetrical shape on the histograms and does not fit along the regression lines in the quantile plots. The histograms also show that the data is contained within a small range of values.

As per table 4.8 info, error and warn do not conform to the skewness and kurtosis tests to be of a normal distribution.

	Skewness	**Kurtosis**
Info	215.4	272.0
Warn	25.1	680.9
Error	45.2	2144.2

Table 4.8: Hourly Skewness - Kurtosis

Normalcy goodness of fit tests was conducted on the data. SW and AD tests were implemented.

	Log Type	**Test Statistic**	**P Value**
Info			
	SW	0.2	0.0
	AD	552.8	1.0
Warn			
	SW	0.1	0.0
	AD	980.4	1.0
Error			
	SW	0.0	0.0
	AD	997.5	1.0

Table 4.9: Hourly Goodness Of Fit Tests

SW Test

Null Hypothesis: The data is normally distributed.

If p-value < 0.05 reject the null hypothesis. The data is not normally distributed.

AD Test

Null Hypothesis : The data is normally distributed.

Critical values [10% : 0.65, 5% : 0.78, 1% : 1.09]

If test statistic > critical values : Reject the null hypothesis the data is not normally distributed

Info, warn and error event types reject the null hypothesis for the AD test. The data is not normally distributed. The SW test also rejects the null hypothesis, the data is not normally distributed.

4.2.2 Stationarity

The stationarity tests that were implemented on the hourly data.

ADF Test

Null Hypothesis: Data has unit root(implies not stationary).

Critical values: [10% : -2.56, 5% : -2.86, 1% : -3.43]

P value < 0.05: Reject the null hypothesis, the data is stationary.

If ADF statistic > critical values: Reject the null hypothesis of unit root. The time series is stationary.

KPSS Test

Null hypothesis for the KPSS test : The data is stationary

Critical values: [10%: 0.34, 5% 0.46, 1%: 0.73]

If test statistic < critical value : Fail to reject the null hypothesis, the data is stationary.

ADF critical values: [10% : -2.56, 5% : -2.86, 1% : -3.43]

KPSS critical values: [10%: 0.34, 5% 0.46, 1%: 0.73]

	Test Statistic	P Value
ADF	-19.27	0.00
KPSS	1.28	0.01

Table 4.10: Hourly Info Stationarity Values

	Test Statistic	P Value
ADF	-13.49	0.00
KPSS	0.27	0.1

Table 4.11: Hourly Warn Stationarity Values

	Test Statistic	P Value
ADF	-27.00	0.00
KPSS	-0.13	0.1

Table 4.12: Hourly Error Stationarity Values

Results: Unit Root Tests

The time series is stationary for info and error type events as they pass both the KPSS and ADF test. Warn type events do pass the ADF test but fail the KPSS test.

4.2.3 Seasonality & Trend

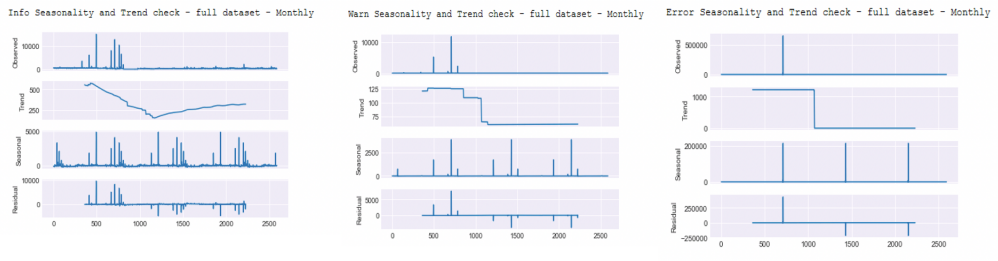

Figure 4.8: Hourly Seasonality-Trend

A trend and seasonality graph was created. The graph was based on hourly data. As per table 4.8 it is visually observed that there was a negative followed by a positive trend detected in the monthly time series data for info event types. A step downward type trend was detected for warn and error type events. Seasonality is observed for each severity event type.

A statistical CH test was conducted to see if the data contained seasonality. The test was implemented for daily, weekly and monthly frequencies. The results conclude that there is no evidence of seasonality in the dataset for daily and weekly data but there was evidence of seasonality in monthly data for info and warn but none for error.

4.2.4 Correlation

ACF-PACF

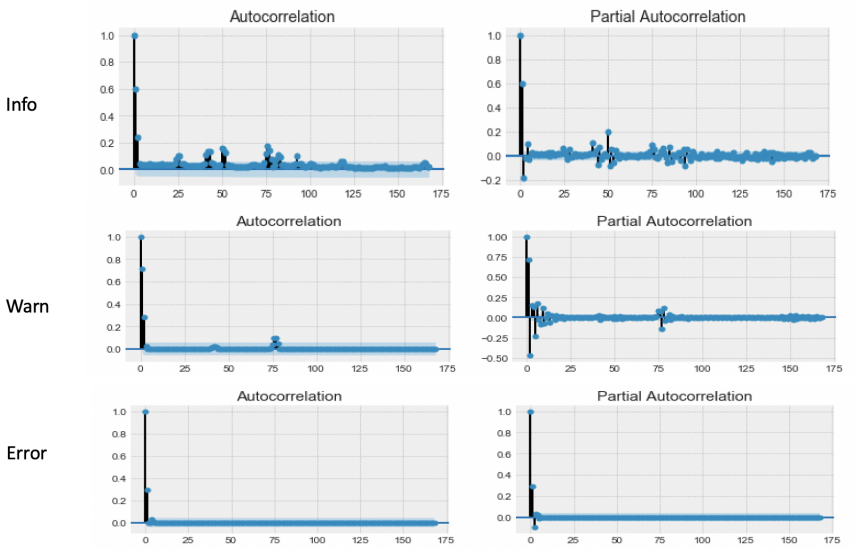

Figure 4.9: Hourly ACF-PACF

For the correlation chart in figure 4.9 , the first 168 lags were used. This is a representation of one week's worth of time series data. We observe from the info correlation chart that there is still some correlation within the time series data around lag twenty-five

onwards which is an indication that the time series data is dependant on its previous time series observations. The partial autocorrelation chart shows that there is still some residual noise which exceeds the significance threshold. For warn, there appears to be no correlation on the data except for the first three lags with the PACF still showing some residuals on the first ten lags. For error, there was no correlation at all through the time series except around lag two. Looking at the p and q values for ARIMA modelling

Info: The data would need to be differenced to become more stationary

Warn: The data would need to be differences to get rid of the residual noise on the PACF plot

Error : (1,1)

We conclude from the hourly data that transformation should be implemented on info and warn and no transformation is required for error data.

4.2.5 Cross Correlation

Hourly cross-correlation analysis was done on all the different event types. The lag value was set to twenty-four which represents a full day. Cross-Correlation was conducted on all untransformed events.

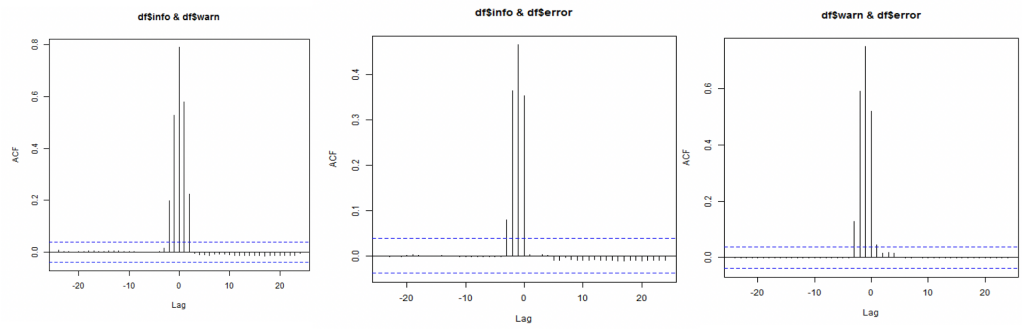

Figure 4.10: Hourly Cross Correlation

The results of the cross-correlation charts imply that info and warn have a significant correlation at lag one and lag two. This implies that an info event will become a warn

type event within the first two lags which represents the t-1 with a less significant correlation at t-2. The info and error correlation chart have significant correlation at lag one. This implies that an Info type event does have a strong correlation with an error type event at t-1. The warn and error correlation chart also show significant correlation at lag one. This implies that a warn type event will result in an error type event within the first hour as the time will be t-1.

Chapter 5

Time Series Modelling

5.1 GARCH

A GARCH Model was implemented on each of the non transformed and transformed datasets so that an LM test could be conducted. The result of the LM test was to conclude if heteroscedasticity occurred in the model and if so it implied that the data was volatile and not suitable for ARIMA modelling.

LM Test Results On Non-Transformed Data

	Test Statistic	**P Value**
Info	-4.37	0.99
Warn	0.02	1.00
Error	11.03	0.27

Table 5.1: Hourly Engle's LM Test for Autoregressive Conditional Heteroscedasticity

From table 5.1 :
Info p-value (0.99) > test statistic (-4.37). We reject the null hypothesis and conclude that heteroscedasticity does exist.

Warn p-value (1.0) > test statistic (0.02). We reject the null hypothesis an arch element does exist.
Error p-value (0.27) < test statistic (11.03). We fail to reject the null hypothesis, no arch element exists in the error dataset.
This indicates that info and warn require a transformation before ARIMA modelling can be implemented. The error dataset requires no transformation.

5.2 ARIMA

For time-series forecasting an Auto ARIMA model was tested to automatically identify the best order of the p, d, q values. As seasonality was detected via the seasonal decomposition function a SARIMA model was implemented. The parameters of the model were set to forecast four hours into the future with the seasonal parameter set to "True" and the seasonal period set to "24" which represents hourly data over one day. A set of tests were conducted on the model to aid in the acceptance of the best fit model.

Model Tests:
ADF
KPSS
CH
STL
LM
LB
SW
AD
CM

Transformations were implemented on the data due to the detection of seasonality and trends from the initial exploratory analysis. To avoid over-fitting an analysis was conducted on each model order and comparing it to that of other model p, d, q orders. The results of all tests are based on the independent statistical tests implemented.

An LB test was conducted to test for autocorrelation on the residuals. The results of the LB test indicate if the time series data is dependent on previous time series lags. For prediction modelling the time-series data should not be dependent on previous time series values as it implies seasonality or trend may exist in the data.

LB Test On Non-Transformed Data

	Test Statistic	**P value on Chi Square Distribution**
Info	0.73	0.99
Warn	0.03	1.00
Error	2.85	0.98

Table 5.2: Hourly Ljung-Box Q-Test

The LB Null Hypothesis tests that no serial correlation exists up to lag ten.
Info p value (0.99) $>$ alpha (0.05)
Warn p value (1.00) $>$ alpha (0.05)
Error p value (0.98) $>$ alpha (0.05)
The test concludes that we fail to reject the null hypothesis. There is no correlation in the time series residuals. This can be further demonstrated by figure 5.1 and table 5.3 which show all the p values are less than 0.05.

LB Graph For Each Severity Type

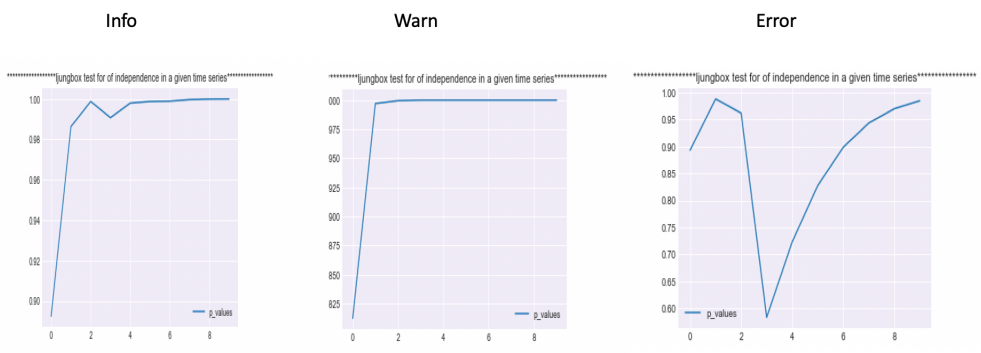

Figure 5.1: Un-Transformed Hourly LjungBox Test

	Info	**Warn**	**Error**
p-value	0.89	0.98	0.89
p-value	0.98	0.99	0.98
p-value	0.99	0.99	0.96
p-value	0.99	0.99	0.58
p-value	0.99	0.99	0.72
p-value	0.99	0.99	0.82
p-value	0.99	0.99	0.89
p-value	0.99	1.	0.94
p-value	0.99	1.	0.96
p-value	0.99	1.	0.98

Table 5.3: Hourly Ljungbox P Values

5.2.1 Info Hourly ARIMA Analysis

ARIMA analysis was conducted on Info type events. Three transformations were attempted on the dataset. These transformations include natural log, first difference and square root.

Some initial discrepancies were evident in the hourly analysis and were noted for the

rest of the ARIMA modeling analysis. It was observed that seasonal decomposition often concludes that trend and seasonality do exist but the CH test does not always pick this up. This is due to the limitations of the CH test and the fact that it is sensitive to data not being transformed and it is not able to identify higher-level trends within the dataset. With that, the seasonal decomposition test took more power than that of the CH test. It was also observed that where trend existed the ADF test was not detecting that trend existed, this may be as a result of the near observation equivalence problem that ADF suffers from. With that, a SARIMA model was tested where seasonal decomposition showed that either trend or seasonality existed. Further to this point all results in the tables are reflective of statistical tests and are not an accurate assumption but are a guide in our analysis.

Type	**Test**	**Untransformed**	**Log(x)**	**1st Diff**	**Sq Root**
No Unit Root					
	ADF	True	True	True	True
	KPSS	False	False	True	False
Trend					
	Ndiff ADF	True	True	True	True
No Seasonality					
	CH	False	False	True	False
	STL	False	False	False	False
No ARCH Effect					
	LM	True	True	True	True
TS Indep.					
	LB	True	True	True	True
Normal Dist					
	SW	False	False	True	False
	A-D	False	False	False	False
Model					
	SARIMA	(1, 0, 2)x	(2, 0, 3)x	(1, 0, 0)x	(2, 0, 3)x
		(1, 0, 1, 24)	(0, 0, 2, 24)	(2, 0, 2, 24)	(1, 0, 0, 24)
Model Score					
	RMSE	122.85	0.45	2.01	2.93
	AIC	39756.63	1955.72	10547.20	15564.92
	Accepted or Rejected	Reject	Accept	Reject	Reject

Table 5.4: Hourly Info Model Transformation Analysis

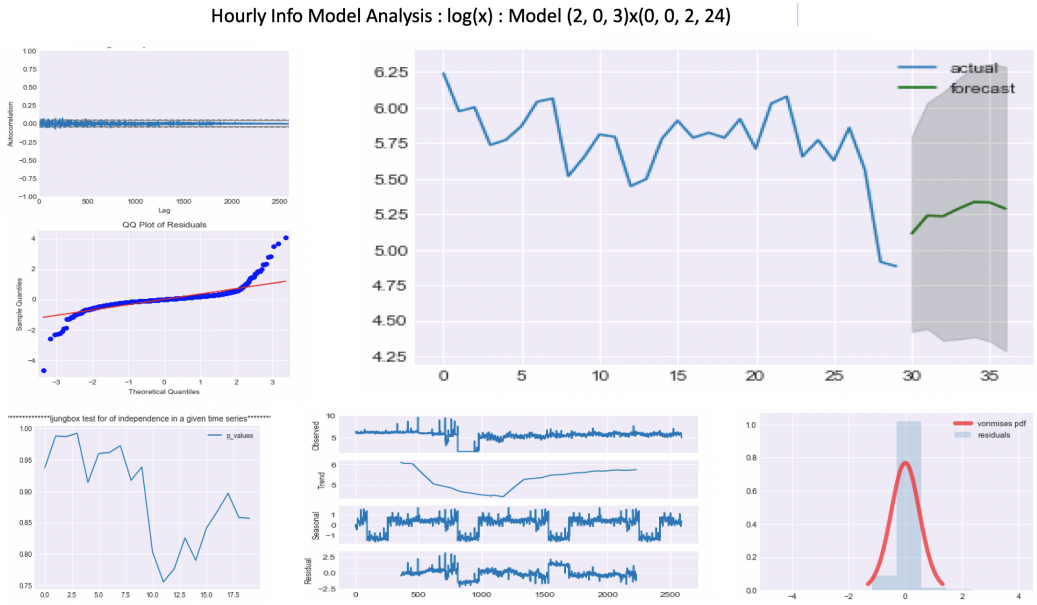

Figure 5.2: Info Hourly ARIMA Analysis log(x)

The results of the analysis as per table 5.4 conclude that the natural log(x) transformation was the best fit model with an order of (2,0,3)x(0,0,2,24) an AIC of (1955.72) and an RMSE of (0.45). The Unit Root tests do conflict with each other and only agree on 1st difference transformation. There is evidence to suggest that seasonality does exist via most of the transformations using the Canova-Hansen and Seasonal Decomposition tests. On testing for heteroscedasticity, there was evidence to suggest that it existed when no transformation was performed on the data but after transformation, it was smoothed out. Passing the LB test confirmed that the time series data was not dependant on previous time series lags which is also evident in the correlogram of the model as per figure 5.2 as no values are outside of the confidence interval boundaries. The quantile plot in figure 5.2 does indicate a fair fit model based on the fitted regression line. The data is not of a normal distribution as it deviates from the regression line on the quantile plot and which is evident in the CVM graph.

5.2.2 Warn Hourly ARIMA Analysis

Warn type messages were transformed based on natural log, first difference and square root to determine which transformation fitted the model best.

Type	**Test**	**Untransformed**	**Log(x)**	**1st Diff**	**Sq Root**
No Unit Root					
	ADF	True	True	True	True
	KPSS	True	True	False	True
No Trend					
	Ndiff ADF	True	True	True	True
No Seasonality					
	CH	False	True	True	True
	STL	False	False	False	False
No ARCH Effect					
	LM	False	True	True	True
No TS Indep.					
	LB	True	True	True	True
Normal Dist					
	SW	False	False	False	False
	A-D	False	False	False	False
Model					
	SARIMA	(2, 0, 2)	(1, 0, 2)	(3, 0, 2)	(1,0,2)
Model Score					
	RMSE	25.60	0.03	0.48	0.31
	AIC	35060.71	-1717.81	4357.86	11618.60
	Accepted or Rejected	Reject	Accept	Reject	Reject

Table 5.5: Hourly Warn Model Transformation Analysis

Warn Hourly ARIMA Analysis

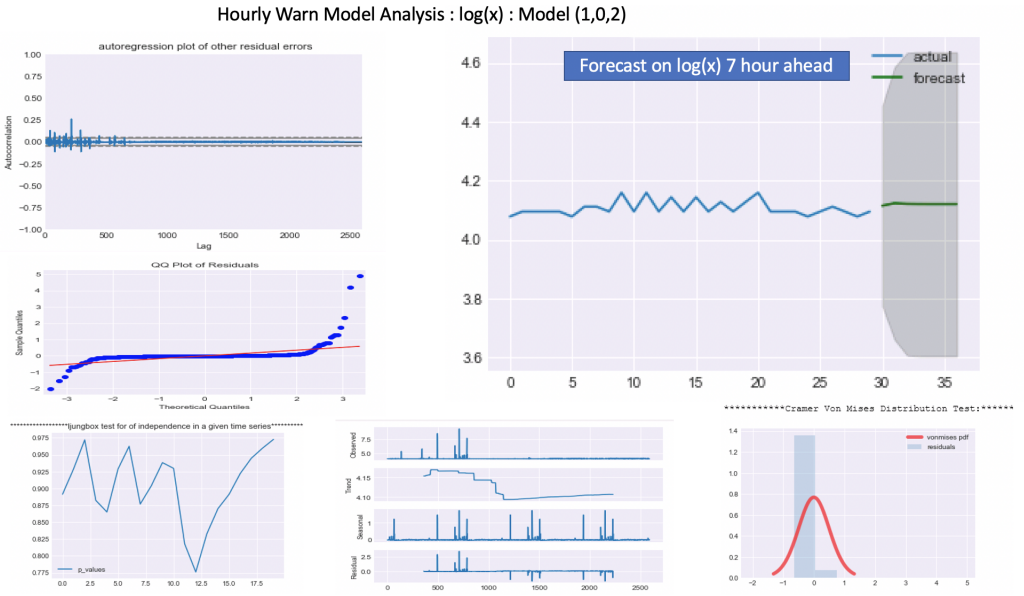

Figure 5.3: Warn Hourly ARIMA Analysis log(x)

For warn the model analysis table 5.5 shows that the natural log transformation produced the best model fit of order (1,0,2) with an AIC (-1717.81) and RMSE (0.03) score. Seasonality and trend were detected within the dataset but showed conflicting evidence. The test for heteroscedasticity does indicate that the data is volatile when no transformation is performed but after transformation, the ARCH element no longer exists. Throughout all different model analysis, there is no indication that the time series is dependant on previous time series values. Figure 5.3 shows the quantile plot produces a fair fitted model as the data does generally fit the regression line very well except when it deviates from the line at the ends. There still appears to be some residual noise in the autocorrelation plot which is an indication that correlation does exist with previous time lags in the model but it is not strong enough to fail the LB test. Seasonal decomposition does show evidence of trend and seasonality. The CM test does indicate that the data is not normally distributed after the model has been fitted

5.2.3 Error Hourly ARIMA Analysis

Error Hourly ARIMA Analysis

Type	**Test**	**Untransformed**	**Log(x)**	**1st Diff**	**Sq Root**
No Unit Root					
	ADF	True	True	True	True
	KPSS	True	True	True	False
No Trend					
	Ndiff ADF	True	True	True	True
No Seasonality					
	CH	True	True	True	True
	STL	False	False	False	False
No ARCH Effect					
	LM	True	Not completed	Not completed	Not completed
TS Indep.					
	LB	True	False	True	True
Normal Dist					
	SW	False	True	False	False
	A-D	False	False	False	False
Model					
	SARIMA	(0, 0, 1)	SARIMAX	(1, 0, 0)	(3, 0, 3)
Model Score					
	RMSE	363.26	1313.78	5.99	0.82
	AIC	56281.57	41611.88	56290.17	21743.40
	Accepted or Rejected	Rejected	Rejected	Rejected	Accepted

Table 5.6: Hourly Error Model Transformation Analysis

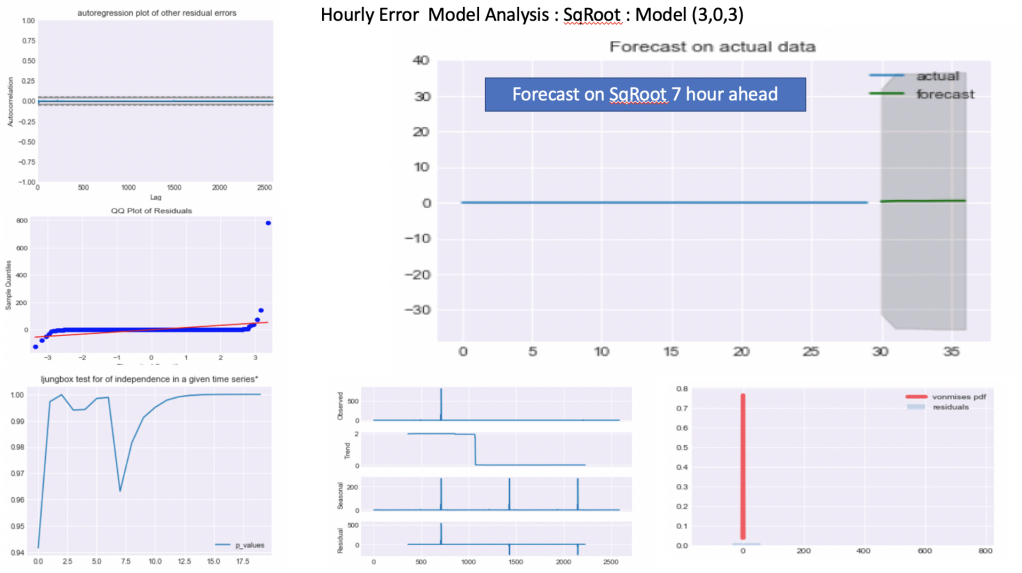

Figure 5.4: Error Hourly ARIMA Analysis Square Root

Error

For error the model analysis reference to table 5.6 observes that the square root transformation was the best fit model (3, 0, 3) with an AIC score of (21743.40) and an RMSE (0.82) score. For most of the transformations the model did not contain unit root but there was conflicting evidence to support this. Seasonality and trend do exist via the seasonal decomposition charts. The correlation on the residuals passed the LB test which indicates that the time series is not dependent on past time series data.

5.2.4 Info Hourly Prediction Analysis

Further analysis was conducted on Informational type messages. Warn and Error type messages were excluded from the analysis due to time constraints.

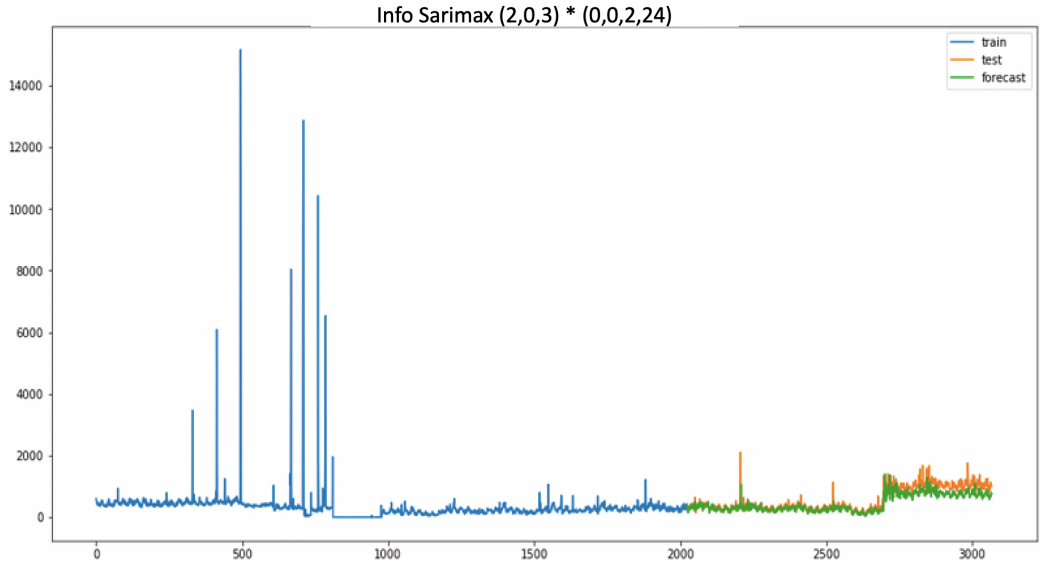

Figure 5.5: Info Hourly Train And Prediction Result

From figure 5.5 we observe the trained, actual and prediction data values. Visually the model appears to fit the data very well but there appears to be an apparent slight shift in the data from series 2700 onwards. This shift shows a very slight underprediction of the ARIMA model. Based on the visual inspection of the graph, this model does appear to be a fair fit model.

A closer inspection was conducted on the model with the trained data removed from the graph. A 2 standard deviation window was added to see how far the model stayed within the confines of the upper and lower confidence interval.

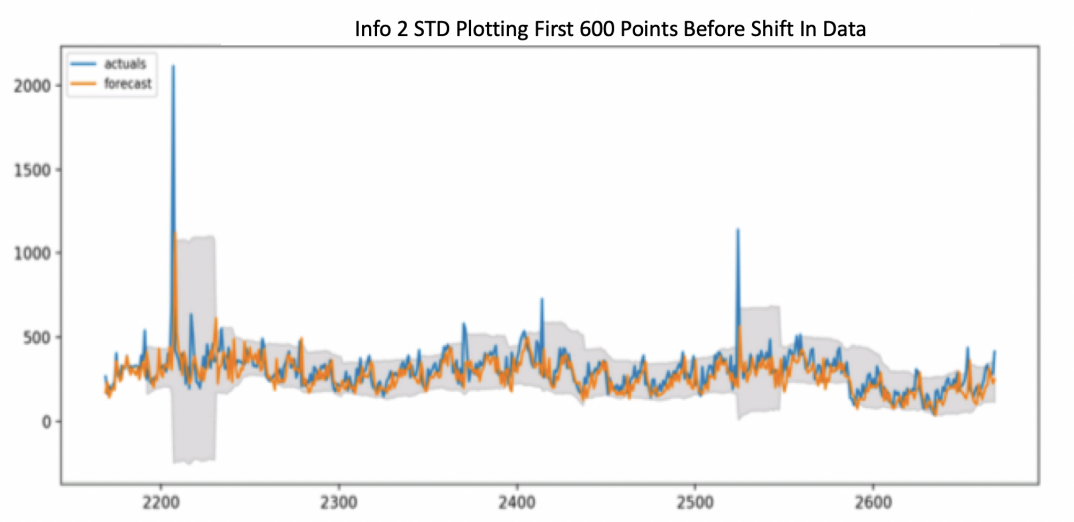

Figure 5.6: Info ARIMA Predictions Before Shift In Data

Figure 5.6 shows a closer look at the data before the shift in the data occurred. We can see from the graph that the predictions are quite close to the actual values but some predictions are underpredicting where there appears to be a higher than normal increase in messages.

On looking at figure 5.7 we can see that a lot of the predictions are outside of the lower confidence interval boundary. This graph gives a clearer indication that this model is not the best fit due to the inaccurate predictions of the data when a shift occurs.

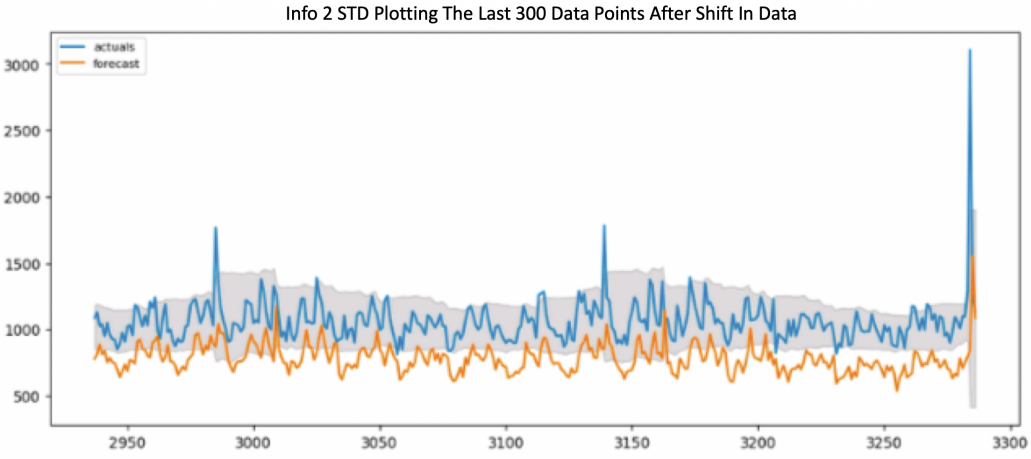

Figure 5.7: Info ARIMA Predictions After Shift In Data

5.3 LSTM

LSTM Modelling was conducted on Info type events. No modelling was done on warn and error type messages due to time constraints. The analysis was conducted based on the first difference and no other transformations were done on the data.

5.3.1 Info Hourly LSTM Analysis

A univariate sequential LSTM model was built using a walk forward model. Its parameters were tested with four memory neurons using a loss function of mean squared error and an Adam algorithm. Each test was repeated ten times and the average RMSE value was returned. This is because each time an independent model is run it produces different RMSE values. Getting the mean of the RMSE provides more confidence that the test result was not a statistical fluke. For each type of test, only the epoch value was changed. The values tested were 1, 10, 50 and 100. The batch size remained at a constant of 1. The train and test size were kept the same as the ARIMA Model to help align the parameters as close as possible to each other.

Descriptive Statistics and Box plots graphs were conducted on each of the tests for LSTM.

Epoch Test	**Count**	**RMSE**	**STD**	**Min**	**25%**	**50%**	**75%**	**Max**	**Status**
1	10	155.82	7.36	146.63	150.56	154.78	160.36	168.77	Reject
10	10	147.80	3.52	142.85	145.50	146.70	149.88	153.92	Reject
50	10	145.43	3.46	138.10	144.03	145.98	148.11	149.30	Accept
100	10	2542.56	7573.09	143.98	145.70	147.63	149.22	24095.96	Reject

Table 5.7: Hourly Info LSTM Model Descriptive Statistics

Table 5.7 shows the results of each test. The status column determines which model was accepted or rejected. Running ten iterations with a batch size of one using fifty

epochs produced the lowest mean RSME with a value of 145.43. An epoch size higher than 50 contained the worst result with an epoch size of ten being the 2nd best walk forward model. This can be seen in figure 5.8.

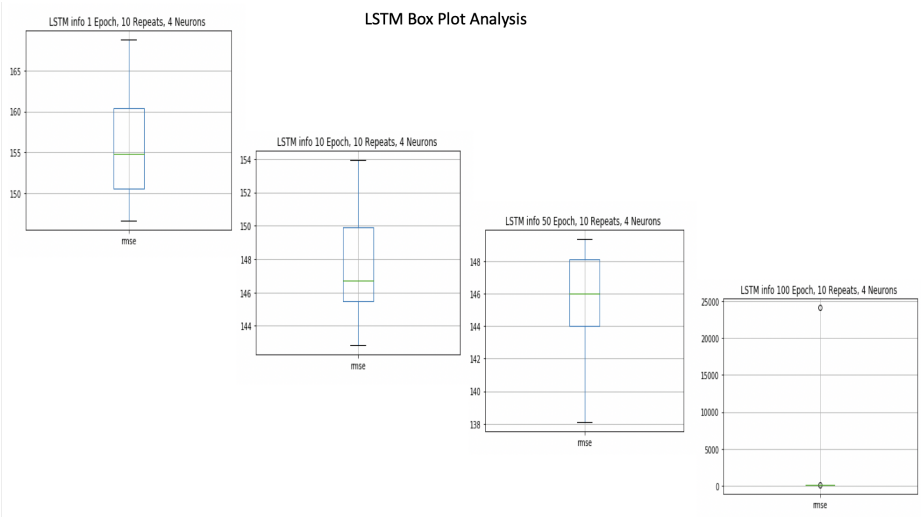

Figure 5.8: Info LSTM Model Univariate Walk Forward Box Plot Analysis

5.3.2 Info LSTM Prediction Analysis

Figure 5.9 shows the train and test model results of the last iteration of the fifty epoch model. The initial set of values for training the model ie that is the 66% of the data has been removed. We can see from the graph that LSTM has handled the shift in the data very well. The test values which are the prediction values are quite close to the test values.

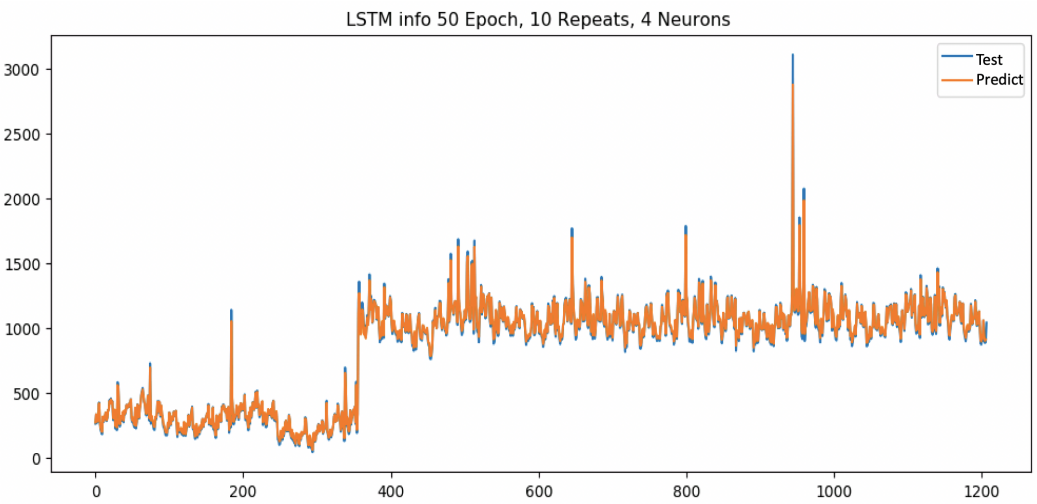

Figure 5.9: Info Hourly LSTM 50 Epoch Prediction Analysis

To ensure consistency in approach the LSTM model was zoomed in on the first three hundred and fifty points. Figure 5.10 shows that the model is a very good fit model. two standard deviation confidence interval boundaries were set on the upper and lower limits.

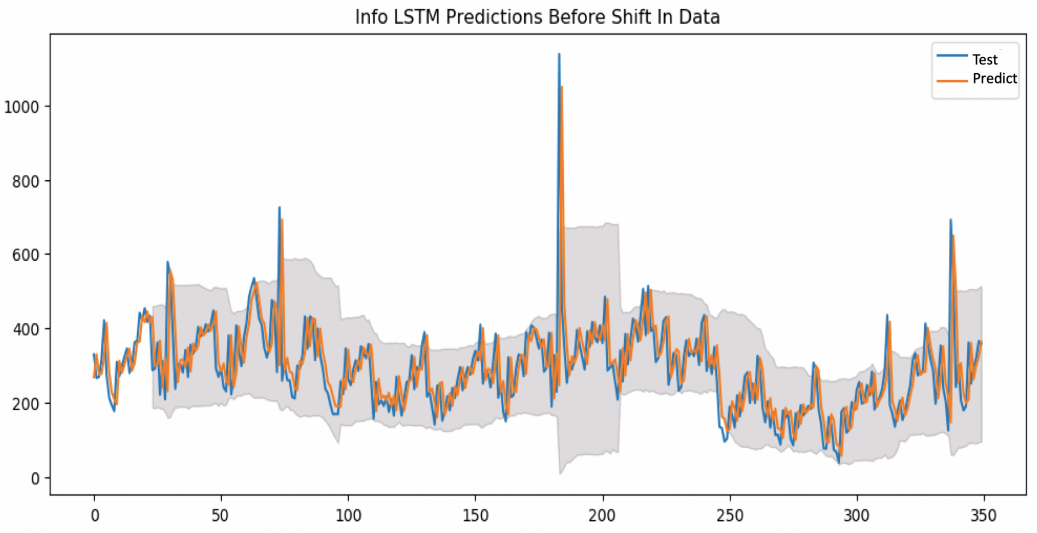

Figure 5.10: Info LSTM Before Shift - Plotting First 350 Data Points

Zooming in on the shift in the time series in figure 5.11 we can see that the model is a very good fit.

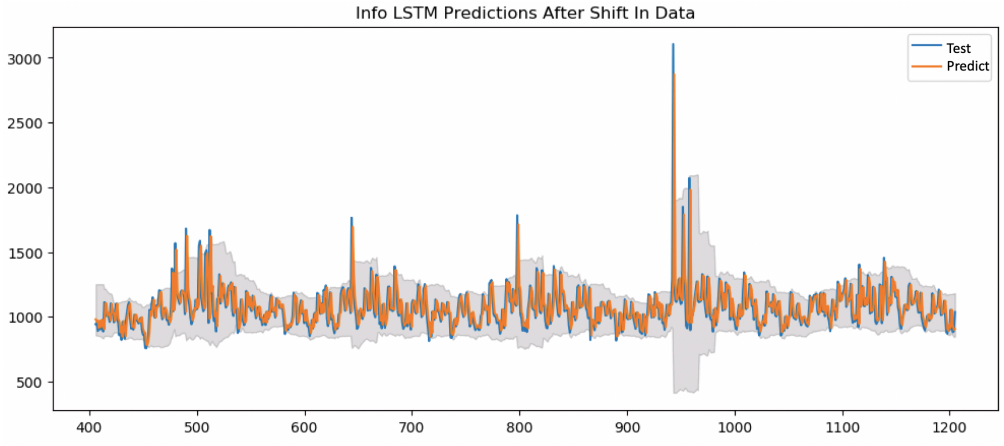

Figure 5.11: Info LSTM After Shift - Plotting last 800 Data Points

Chapter 6

Anomaly Detection

To find anomalies in data one needs to look at extreme values or values that deviate from the norm that are not reflective of cyclic seasonality or trends. For anomaly detection residual error, principal component analysis, cooks distance and level shift are some of the tools used to determine if data deviating from the norm is an actual anomaly or not. These anomalies are based on unexplained observations and are also known as outliers and both these words are used quite interchangeably in the research papers. Types of anomalies are a point, contextual and collective. Point anomalies are also known as additive outliers which are defined by (Fox, 1972) and his interpretation on how to capture them is via a likelihood ratio test. These anomalies are a sudden sharp increase in value followed by a sudden change back to normal. Collective anomalies are when a consecutive number of anomalies occur throughout observations also known as transient change outliers. These collective anomalies can be caused by a seasonal shift in the data which is known as a level shift.

Collective anomalies are the scope of this project. Our investigation is to identify collective anomalies and compare them against that of the ARIMA and LSTM models. A simple approach used to detect if anomalies occurred is to evaluate how many points the data deviated from the mean using a standard deviation (STD) function. A

twenty-four-hour rolling window for the STD was used. The sigma levels were based on two STD's so that the anomaly was not limited to only identifying really large spikes in the data. Due to their being no domain experts involved, no outliers in the data were removed and data was analysed based on all data points. The residual errors were graphed to see if there was any visual observation of anomalies in the data based on the two standard deviation confidence level. Anomalies were only conducted on Informational type messages due to the constraints of time.

6.0.1 ARIMA

Figure 6.1 shows the residual errors from the ARIMA model. It is observed that anomalies have occurred in the model based on the points that deviate outside of the upper and lower two STD confidence interval boundaries. Visually it is hard to tell if collective anomalies have occurred.

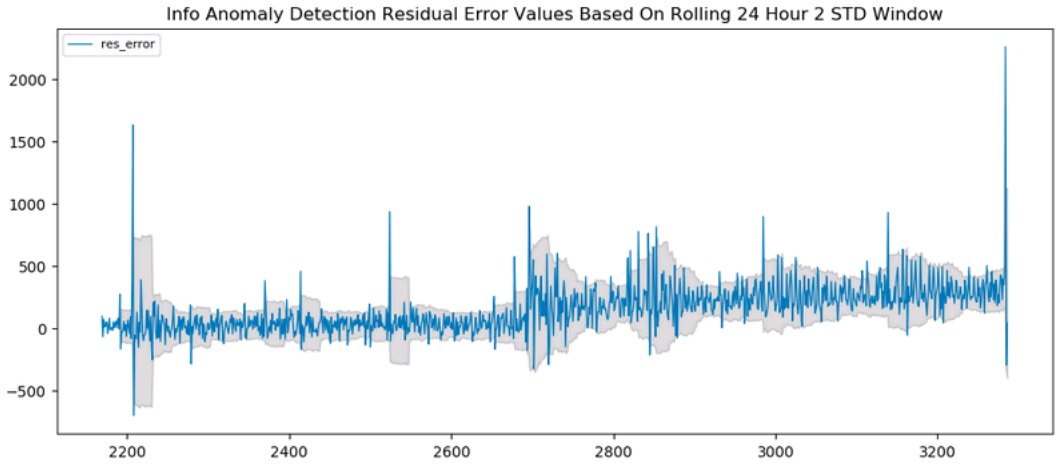

Figure 6.1: Info ARIMA Residual Errors

Two graphs have been created. Here collective anomalies have been detected. Two graphs have been plotted. These graphs are filtered to show a reduction in the dataset that is concentrated in showing detected collective anomalies. From figure 6.2 we can

visually see more clearly the anomalies detected. Three collective anomalies have occurred within the full dataset.

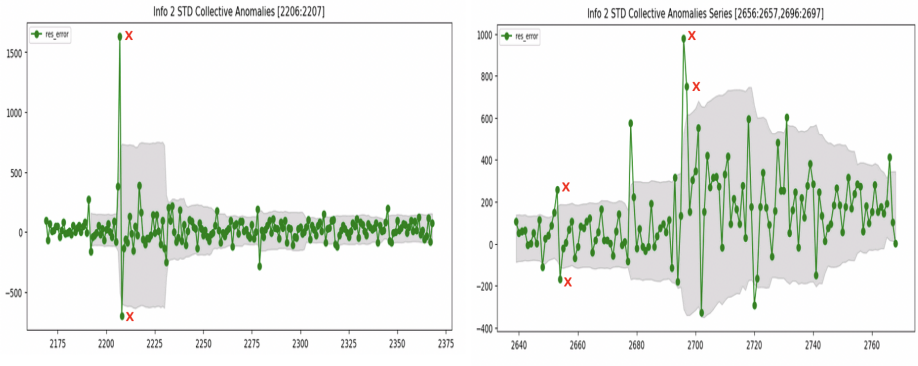

Figure 6.2: Info : Two STD Collective Anomalies

Table 6.1 identifies the amount of point and contextual anomalies detected. It also identifies the series of where the anomalies occurred. As expected there is more one STD anomalies than that of two and three STD's. Two STD's for Informational type messages observed three collective anomalies between series 2206 and 2207 and series 2652 to 2653 and series 2696 to 2697.

Deviation	**Point**	**Collective**	**Series**
Three STD	33	2	[2206:2207,2696:2697]
Two STD	43	3	[2206:2207,2652:2653, 2696:2697]
1 STD	65	6	

Table 6.1: Info - Anomaly Count

There is significant variance in the STD around series 2220. The data was then further analysed to see if some sort of a pattern existed that caused the significant spike to occur. We can see that the data reached its peak very sharply over one hour and was not, in fact, a gradual incline as per figure **??**. It may be determined that this is due

to missing data and a further check was done to determine if the data was indeed missing.

Series	Date	Value
2206	2019-03-26 14:00:00	689.0
2207	2019-03-26 15:00:00	2115.0
2208	2019-03-26 16:00:00	420.0

Table 6.2: Info Anomaly Detection Missing Data Check for Spike

We can see from table 6.2 that this is not in the case, that there was, in fact, no missing data for that period. A domain expert would need to asses this incline to give a better indication as to the reason for the significant increase.

6.0.2 LSTM

The residuals of the LSTM model were graphed in figure 6.3 with the two STD boundaries added. Most of the residuals are centered around zero except for the residuals near-series 900. The residual graph appears stationary and does not show the level shift that occurs in the ARIMA residual graph 6.1.

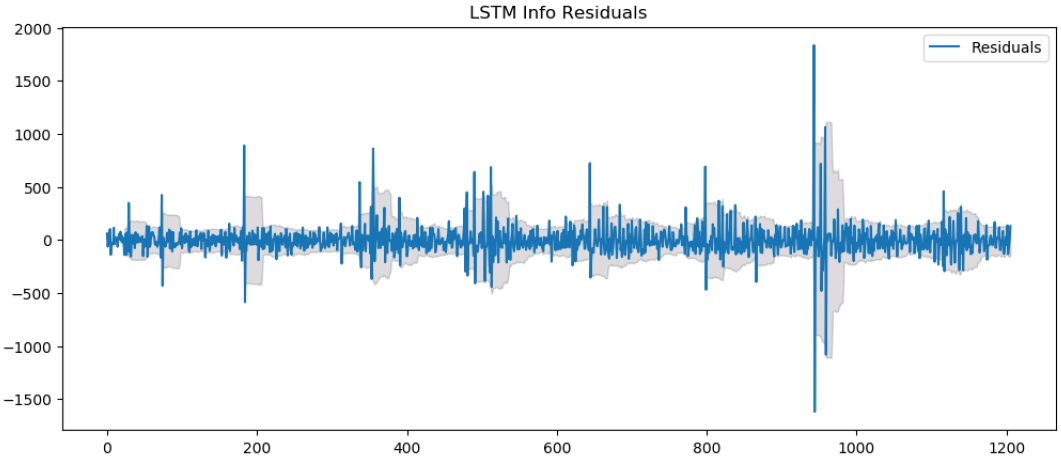

Figure 6.3: Info LSTM Residuals

Anomalies have been plotted with an x on figure 6.4. Eighty-four point and fifteen collective anomalies have been detected. For each collective anomaly detected all's it anomalies have been plotted. Out of the fifteen collective anomalies detected thirteen occurred within a two-hour window and two occurred within a three-hour window. The green x's represent the three-hour window and the red x's indicate the two-hour window.

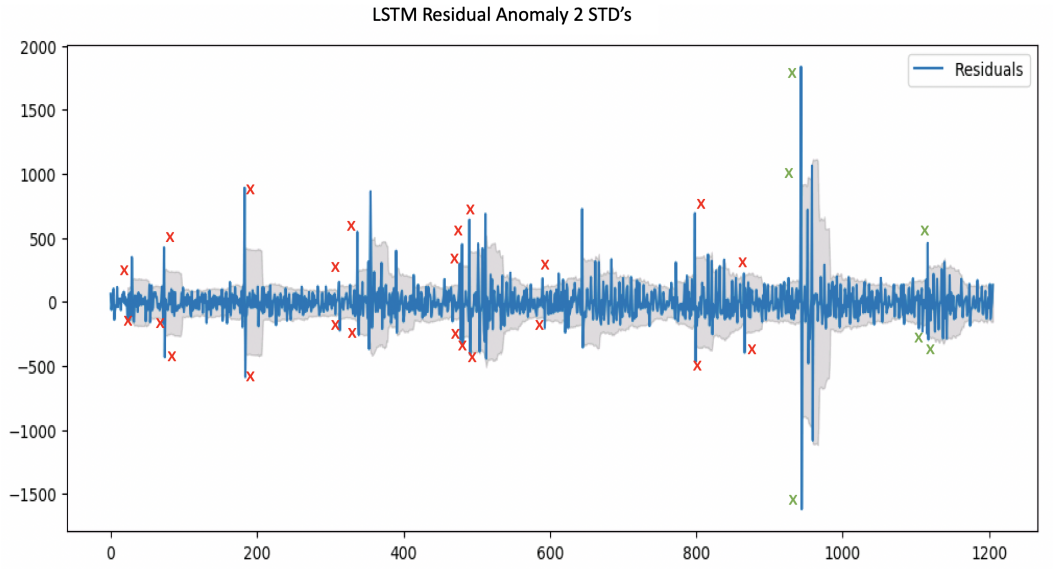

Figure 6.4: Info LSTM Collective Anomalies On The Residuals

Anomaly Comparison

For ARIMA it detected three collective anomalies while LSTM detected fifteen as per table 6.3

Model	point	Collective
ARIMA	43	3
LSTM	84	15

Table 6.3: Info LSTM and ARIMA Anomaly Count

Chapter 7

CPU-Memory - Performance Analysis

Memory and CPU metrics were analysed to see if they have any correlation with increased log message output. Pearson's, Kendal and Spearman are Goodness of fit tests. Pearson's cross-correlation statistic was used to test the linear relationship between the variables. A correlation coefficient of one indicates a positive high correlation. -1 indicates a negative correlation. A correlation of zero indicates no correlation.

7.1 CPU

For CPU metrics it was identified that the server contained thirty-one CPU's. A "Percent Total CPU Used" metric was used for analysis. Figure 7.1 shows the person's cross-correlation coefficients matrix results. With a correlation value of r=0.1, this correlation coefficients indicates that there is a low correlation between CPU usage and Informational log message output.

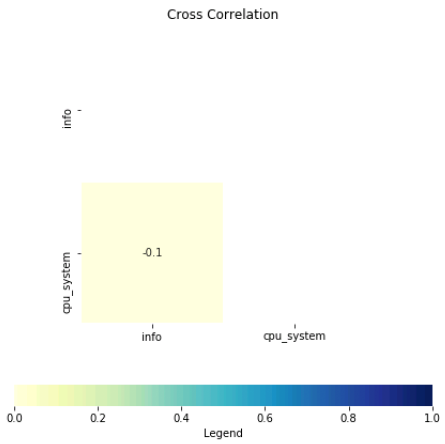

Figure 7.1: Info and CPU Hourly Pearson's Correlation Analysis

Figure 7.2 shows that seasonality exists within the data as we can see a pattern emerging. The correlation is negative at its highest point at lag 8 and then drifts off. A correlation coefficient of r=1.5 suggests that the evidence is not strong enough to indicate correlation exists.

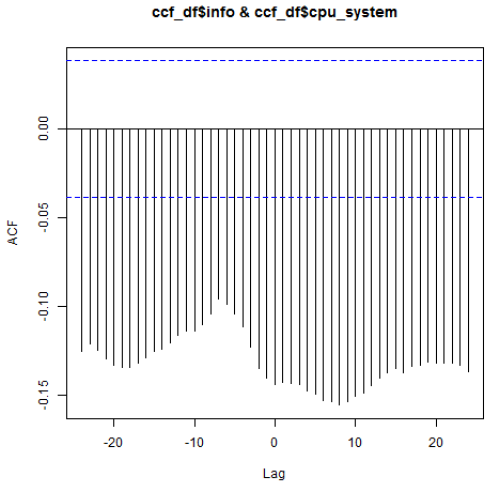

Figure 7.2: Info and CPU Hourly Correlation Analysis

7.2 Memory

For memory metrics, the server contained 100gb of memory. A "free memory" metric was used for analysis. Figure **??** shows the results of the Pearson test which indicates a correlation value of r=0.2. This indicates a low correlation between memory usage and Informational log message output.

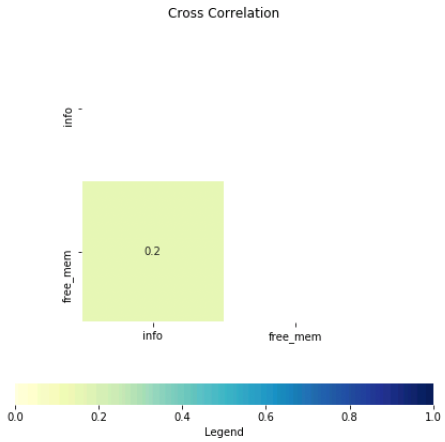

Figure 7.3: Info and Memory Hourly Pearson's Cross Correlation Analysis

In figure 7.4 the correlation graph between info and free memory indicate a shift in the data from lag zero onwards. The correlation value is quite low at 0.15 and suggest that there is no evidence to suggest a correlation between memory and info log messages.

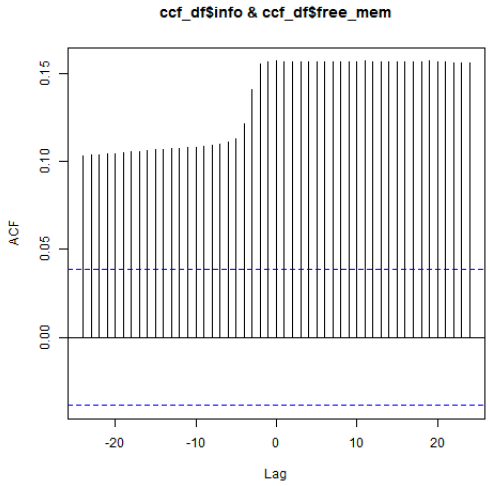

Figure 7.4: Info and Memory Hourly Correlation Analysis

Chapter 8

Evaluation

For Daily data, it was observed that 55% of the events were generated by the error severity event and only 8% were generated by the warn severity event. A 55:35 split was detected between error and info event types. From these statistical counts, it would appear that an error event may have occurred over a considerable amount of time that caused it to surpass the info type message count. Observationally from these values, it would appear that no correlation exists between the warn and error type events or it may be the case that the error events that occurred may have been stuck in an iterative loop over a considerable period.

8.1 Daily

For time series modelling we need to conclude from the data if it fits a certain pattern or shape. The results of these tests may indicate the need for further tests or transformations to be done before the data can be modelled. Those types of tests are normality, unit root, stationarity, volatility, trend, seasonality and time series dependence tests. The majority of these tests have been conducted on the daily data.

8.1.1 Info

Testing For Normality:

We reject the null hypothesis of the SW test p = 0.0. The data is not normally distributed. We reject the null hypothesis of the AD test (test statistic=0.07 > critical values at 5%= 0.74. The data is not normally distributed. Skewness=2.5 indicates a heavy right-tailed distribution with a platykurtic kurtosis=9.1 both of which indicates variance in the data. Both the quantile plot and the histogram do show that the data was not of a normal distribution. Based on the combined tests there is strong evidence to suggest the data is not of a Gaussian distribution.

Testing For Stationarity:

When testing to see if a shock in the data has an impact on the time series the ADF unit root test p=0.20 indicates unit root does exist and implies non-stationarity. The KPSS test for unit root (test statistic =0.38 < critical value 0.46) shows evidence that unit root does not exist and implies stationarity. Mean and variance tests on the data using two sample populations from the same dataset indicated a high degree of variance and mean. These results do not hold for ARIMA which looks for conditional mean and constant variance. Using the combined tests there was strong evidence to suggest that the info type event data did not present stationarity.

Testing For Trend And Seasonality:

For trend, info type events do show patterns of variance change in the data. Seasonality was also evident in the seasonal decomposition chart. The correlogram did show that trend exists. The results of the CH test indicated that no trends existed in the data. Although statistically there was no evidence to suggest that seasonality existed there was too strong an evidence in the visualization charts to reject the hypothesis that seasonality or trend did not exist.

8.1.2 Warn

Testing For Normality:

Warn type events did display volatility in the data. With the SW test p=3.2, we fail to reject the null hypothesis, the data is normally distributed. The AD test (test statistic =21.6 > critical value at 5% = 0.74) rejects the null hypothesis. There is evidence to suggest the data is not normally distributed. Skewness = 6.7 shows a heavy right-tailed distribution with a leptokurtic kurtosis = 48.7 both of which indicates variance in the data. The histogram and the quantile plot show that the data is not of a normal distribution as it does not fit anywhere along the regression line and the majority of the values in the histogram occur within the zero to one thousand range. Based on the combined tests the evidence is conflicting. If the low number of high outliers were removed from the dataset this may change the results of the skewness and kurtosis test. It may also change the shape of the histogram and the distribution of the fit along the regression line. Further analysis would need to be conducted with the outliers removed to see if they occurred by random chance and are not seen to be part of the normal observation.

Testing For Stationarity:

Testing for stationarity the ADF unit root test p=0.00 implies that the time series has no unit root and is stationary. For KPSS unit root (test statistic =0.12 < critical value 0.46) provides evidence to suggest that the time series is stationary so we fail to reject the null hypothesis. A high degree of variance and mean are an indication that the time series is not stationary. Using the statistical KPSS and ADF tests their is strong evidence to suggest that the warn type event data is stationary. The high variance and mean in the data may be partially due to the high outlier values detected in the dataset.

Testing For Trend And Seasonality:

A transient type of change was observed in the trend chart. Seasonality does exist over repeat observations. The correlogram does not show any trend or seasonality. The CH test failed to detect seasonality or trend in the dataset. Based on the visual and

statistical evidence more tests will need to be conducted on the data to provide more solid reasoning for accepting or rejecting the hypothesis that seasonality and trend exist.

8.1.3 Error

Testing For Normality:

Error type events show volatility in the data. This can be seen in its histogram where a high degree of low values frequently occurs with a low degree of high values. Its quantile plot shows that the data is not of Gaussian distribution as none of the data fits along the regression line. It is observed from the quantile plot and the histogram that a significant outlier occurred that may have contributed to the data not fitting a normal distribution. The test for normality using the SW test p=0.0. shows strong statistical evidence that the data is not of a normal distribution. The AD test (test statistic=585.9 $>$ critical values at 5%=0.74) rejects the null hypothesis, the data is not normally distributed. With skewness=34.7 and kurtosis=1261 this indicates that the data contains a heavy right-tailed distribution and a leptokurtic shape. The results of the normality test for error may be due to the same reasons as that of the warn tests. The significant outlier in the data may have an impact on the data's shape and distribution. Removal of this outlier if it occurred by chance and was not seen to be a normal observational pattern will give a better indication to the true shape of the data. It is suggested that this outlier be removed before any further analysis is conducted.

Testing For Stationarity:

Testing for stationarity the ADF unit root test p=0.0 implies that the time series is stationary. For KPSS unit root (test statistic =0.12 $<$ critical value 0.46) also provides evidence to suggest that the time series is stationary. Running a two population sample mean and variance test on the dataset confirms that they both contain high variance and mean. The results of the test show strong statistical evidence that the time series is stationary through the ADF and KPSS test. It may be the case that the significant

outlier has an impact on the mean and variance results.

Testing For Trend And Seasonality:

The results of the CH test confirmed that no trend or seasonality exists. The seasonal decomposition chart, on the other hand, does visually provide evidence of trend and seasonality. The correlogram does show that no trend or seasonality exists in the ACF plot but there appears to be a negative trend occurring between lag forty-two and forty-eight in the PACF plot. Using the results of the correlogram and the seasonal decomposition there is evidence to suggest that seasonality and trend do exist.

8.2 Hourly

No high-level aggregation analysis was implemented on the hourly data. The analysis was done to ascertain if the hourly data could be time series modelled.

8.2.1 Info

Testing For Normality:

There was evidence to suggest that info type events were not normally distributed by the results of the AD and SW test. Its histogram and quantile plot also showed graphical evidence that the data is not Gaussian. The quantile plot regression line does indicate a not so good fitting. A high number of outliers are deviating from the tail end of the regression line. A non-symmetrical shape is displayed on the histogram with a heavy right-tailed distribution. A kurtosis of 272.0 and skewness of 215.5 also indicate the data may not be of a normal distribution.

Testing For Stationarity:

Their is evidence to suggest that info type events are time series stationary with the ADF test(p=0.00, test statistic (-19.27) $>$ critical values at 95% (-2.86)). The KPSS

test also provides evidence to suggest that the time series is stationary (test statistic (1.28) < critical values at 95% (0.46))

Testing For Trend And Seasonality:

The STL chart does show evidence of trend and seasonality. A positive and negative variance change is detected. The results of the CH, ACF and PACF tests fail to identify seasonality or trend in the data although the ACF and PACF do observe some volatility. A deeper dive on the correlogram for ACF does confirm non-existent trend or seasonality as per figure 8.1

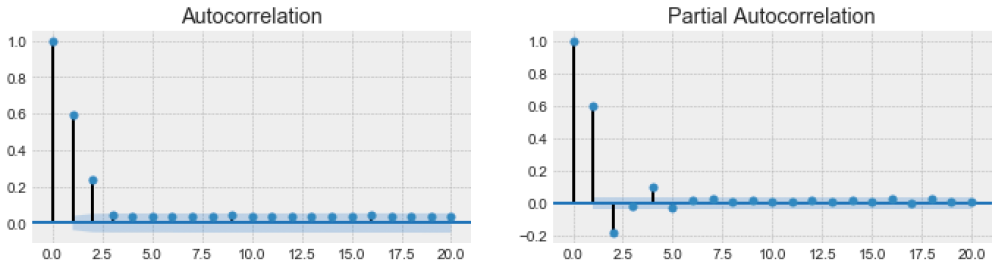

Figure 8.1: Info ACF Filtered On 1st Twenty Lags

8.2.2 Warn

Testing For Normality:

Warn type events were not normally distributed based on the evidence provided by the AD and SW tests. The quantile plot indicates that the data does not fit along the regression line. We also observe a high number of outliers deviating from the tail end of the regression line which may be affecting the shape of the distribution. The histogram does not show a symmetrical shape to support a normal distribution and displays a heavy right tail. Its kurtosis=680.9 and its skewness=25.1 also indicate the data may not be of a normal distribution.

Testing For Stationarity:

Warn type events pass the unit root test for ADF (p=0.00, test statistic =-13.49 > critical values at 95% (-2.86)) implying the time series is not stationary but KPSS

(test statistic = 0.27 < critical values at 95% (0.46)) passes the test for stationarity. This indicates a conflict of results. The ADF test for warn type events may suffer from near observation equivalence and further tests would need to be conducted before a judgement could be made.

Testing For Trend And Seasonality:

The STL graph shows that trend and seasonality do exist. The CH test identified seasonality on the monthly frequency of the hourly data but did not observe seasonality on the hourly or weekly frequency. The ACF graph does not display any trend or seasonality but there appears to be a pattern emerging in the PACF graph as per figure 8.2. This can be observed as a range of values that are spanning in-in time. We can see the negative values at lag two, five, seven and ten showing a slight linear shift.

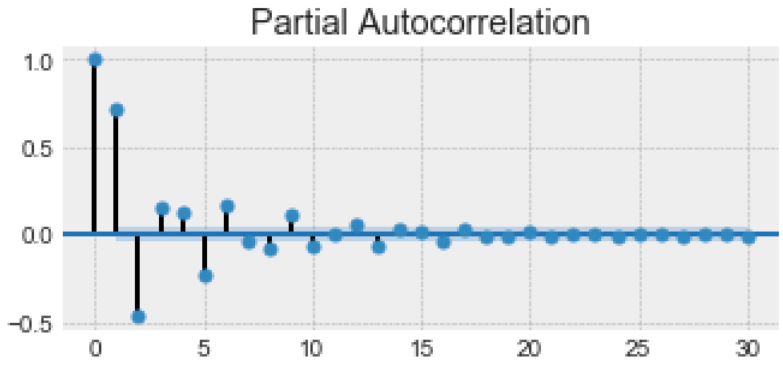

Figure 8.2: Warn PACF First 30 Lags Filtered Observation

8.2.3 Error

Testing For Normality:

The error type events were not normally distributed. Both tests rejected the null hypothesis for the AD and SW tests. The quantile plot shows that the data does not fit along the regression line. Two outliers appear to deviate from the regression linen. A non-symmetrical heavy right-tailed distribution is showing on the histogram. With

a kurtosis=2144.2 and skewness=45.2, it is also evidence to suggest that the data is not normally distributed.

Testing For Stationarity:

Error type events pass the unit root tests for ADF (p=0.00, test statistic=-27.00 > critical values at 95% (-2.86)). The KPSS test also indicates that the time series is stationary (Test statistic (-0.13) < critical values at 95% (0.46))

Testing For Trend And Seasonality:

Tend and seasonality is observed within STL. The CH test does not identify seasonality for any of the daily, weekly or monthly tests conducted which is further supported by the correlogram for ACF which does not indicate any seasonality.

8.3 Daily - Hourly Recap

Daily

For info type events there is evidence to suggest that the data is not normally distributed. The evidence for stationary did not hold. Trend and seasonality were observed in the data.

Warn type events are displaying volatility. The AD test fails on a normality test and SW passes the test for a Gaussian distribution. Other statistical tools provide evidence to suggest that the data is not normally distributed. Unit root tests passed the ADF and KPSS tests and both tests provided evidence that the data was stationary. Trend and seasonality do exist in the data. It was noted that there is a significant outlier in the data. This outlier may have an impact on some of the test results. It is suggested that this outlier be removed or analysed to see if it happened by chance or is a normal observation pattern.

Error type events also show volatility in the data. Both the AD test and the SW test failed for normality. Both the KPSS and ADF test confirm that the data is stationary. Trend and seasonality do exist in the data. As per the suggestions for warn, there is

a significant outlier in the data that may have an impact on some of the test results. It would be suggested that this outlier be removed or further analysis.

The strength of the relationship between two variables was conducted on each of the severity type events. This was done so that it could identify if a causation relationship existed and to what extent was the strength of that relationship. The test showed a significant relationship exists between warn and error and info and warn with a lesser significant relationship with info and error.

Hourly

There is strong evidence to suggest that info type events are time series stationary. This was evident with the results of the ADF and KPSS test. Info type events were not normally distributed. The AD and SW tests showed evidence to suggest this. This was further confirmed with the visual observations from the histograms and quantile plots.

There was conflicting evidence on warn type events for unit root. The ADF test implies non-stationarity while the KPSS implies stationarity. It is known from the literature that ADF suffered from type 1 errors. More analysis would need to be done on the data to confirm if the ADF test suffers from near observation equivalence. A recommendation would be to also try and ADF-GLS test. Warn type events were not normally distributed based on the evidence provided by the AD and SW tests.The high number of outliers observed deviating from the tail end of the regression line would need further analysis to understand the story behind their occurrence.

The error type events were not normally distributed. Both tests rejected the null hypothesis for the AD and SW tests. Their is evidence to suggest that error type events are time series stationary with the passing of the ADF test and KPSS tests. Seasonilty was slightly detected in the STL graphs and all other tests did not detect seasonality or trend existed. Looking at all the combined tests, their is evidence to suggest that seasonality or trend do not exist.

8.4 Time Series Modelling

8.4.1 Info

For time-series modeling, the data went through multiple transformations to detect the best fit model for predictions. It was first noted that the untransformed hourly data suffered from trend and seasonality via STL but that was not detected in the correlogram. It is evident from the tests in table 5.4 that both KPSS and ADF conflicted with their results for stationarity for three of the four tests. The only time the CH test detected seasonality was on its 1st difference transformation. All models presented evidence that the data did not suffer from heteroskedasticity and was suitable for ARIMA modeling. On looking at the results of all the transformations there was never a case where all tests equally passed.
The lowest RMSE of 0.45 of the natural log transformation with a model of (2,0,3)*(0,0,0,24) was used as the best fit model. The (p,d,q) parameters (2,0,3) reflect the ACF and PACF correlogram shown in figure 8.1, which indicate that the ACF p value = 2 and PACF q value=3 with zero for no difference. This no difference may indicate that ADF holds out on this test more than KPSS as the ADF test passed for stationarity but the KPSS test failed for stationarity. The Seasonal values(0,0,0,24) indicate white noise. When we look back at autocorrelation of the residuals from the model as per figure 5.2 this indicates no white noise and there is no evidence to suggest that the model is time series dependant based on the results of the LB test. It may be the case that the time-series does in-fact not contain seasonality, therefore, an ARIMA model may be better suited.

8.4.2 Warn

To recap, the warn data for hourly analysis did not present normality and has conflicting results for stationarity. Trend and seasonality were detected in the dataset. Table 5.5 shows all the results of the model analysis for each of the transformations. Het-

eroskedasticity was not observed after the data was transformed. The lowest RMSE was recorded at 0.03 with an AIC of -1717.81 from the natural log transformation. On testing for seasonality within the ARIMA model the results show that ARIMA did not identify any seasonality or trend, this conflicts with the results of the STL test but does not conflict with the results of the CH test. The (p,d,q) parameters of (1,0,2) were observed as the best parameters from auto ARIMA. Looking at figure 8.3 may suggest that (3,0,2) may be a better fit model.

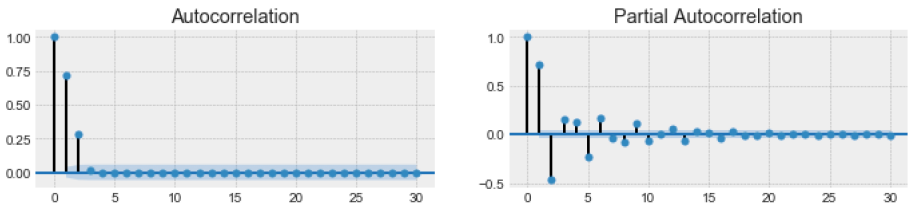

Figure 8.3: Warn ACF - PACF Filtered Observation

8.4.3 Error

The initial analysis of the hourly error data indicated volatility in the dataset. The data was not of a normal distribution and trend and seasonality were not detected. One significant outlier was detected in the data that may have had an impact on the results of the statistical tests. It would be recommended that this outlier be removed from the system as a temporary measure as it is so significant until further analysis can be conducted to see why it occurred and under what conditions caused this behavior. Although the best fit model for info and warn was from the natural log transformation, the accepted model for the error type events was square root. As the data did present volatility before the transformation was conducted we observe from the time-series graph in 8.4 that most of the observations are at zero with one significant spike. As these are error type events - they may not occur as often as info or warn type events. It is recommended that a GARCH model be tested on the data before and also an ARIMA and GARCH combined be conducted.

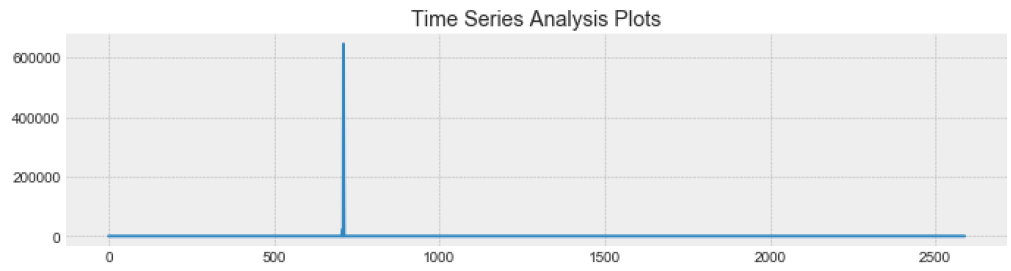

Figure 8.4: Error Hourly Time Series Observation

8.5 Anomaly Detection

8.5.1 SARIMA

The SARIMA model for info type events shows that the test data does fit the train data up to a period where the data does not suffer variance. The predictions against those of the test data do seem to fit the data quite well but it is evident that the predictions are nearly always linear upward trend in the residual errors. A level shift occurred in the data that saw the predictions weakening straight from the point of shift. Their is a slight linear upward trend detected in the residual errors. This level shift would confirm that trends still exist in the data and this was proven from the statistical KPSS test which failed the stationarity test as it should be able to detect the change in mean and variance when the shift occurs. The CH test was conducted to test for trends and seasonality but this test rejected the hypothesis for trend and seasonality. After some investigations, a limitation was identified in the CH test. CH first needs the data to be transformed before it can make its assumption. It can also only detect seasonality or trend at the lowest level of data (Taylor, 2003). On looking back at the hourly transformations this identified limitation does not hold. It is observed that the CH test was able to detect that seasonality existed on the warn type events on the untransformed data as per figure 5.5

After the shift in the data, it becomes quite apparent how far the prediction deviates

from the observed values. It is constantly under predicting by around 200 values at each point. Although the SARIMA model was not able to predict the data it was further used as an analysis for anomaly detection using a two standard deviation approach.

For anomaly detection, a three STD approach was initially used but was then reduced to two STD's to reduce only capturing the extreme outliers and missing the lower impact outliers. As our research was looking at collective anomalies the number of outliers detected was further reduced. It was observed from table 6.1 that out of forty-three point anomalies only three were collective anomalies. These anomalies occurred over 2 periods.

With the SARIMA model detecting a structural pattern change in the data, further research was done in this area. Different algorithms exist for different patterns or shapes identified in data. As our data suffered from level shift it is worth investigating if a level shift algorithm can handle this level shift? A level shift is when there is an abrupt change in the mean level (Balke, 1993) *"Outliers, level shifts, and variance changes are commonplace in applied time series analysis. However, their existence is often ignored and their impact is overlooked."* (Tsay, 1988) A level shift and a transient change outlier algorithm would be a more suitable approach than that of the STD mechanism. Lasisi et al studied outlier detection on airport data. They looked at Innovation (IO), Level Shift(LS), Additive (AO) and Transient Change (TC) Outlier algorithms. Their findings concluded that combined usage of AO, LS, and TC captured 60% of they're outliers with LS producing the best results. (Lasisi & Shangodoyin, 2014) These algorithm's are best suited for level shifts in the data set.

8.5.2 LSTM

LSTM was implemented to see if a deep learning neural network model could better detect and forecast anomalies than that of a classical SARIMA or GARCH model.

Multiple models were tested to bring back the lowest mean RMSE. It was confirmed that running ten iterations with a batch size of one using fifty epochs produced the best results (RSME=145.43). Any iterations above fifty epochs caused a decline in model performance. The LSTM residuals of the model appear to be stationary. There does not appear to be much variance in the data. The results of the tests for LSTM provide a near perfect fit. The observations and the actual values are so close to each other the difference is hardly recognizable. From the anomalies, it detected LSTM identified eighty-four point anomalies and out of those eighty-four anomalies fifteen of them were collective. Some of these anomalies also appear to occur over three periods which means it existed for 1.5 hours.

8.5.3 Comparison

Our initial research aim was to compare SARIMA, GARCH and LSTM models for anomaly detection. We confirm that SARIMA was not suitable for the info type events for anomaly detection due to the existing level shift in the data. LSTM, on the other hand, was able to give more accurate predictions even with the level shift in the data.

Chapter 9

Conclusion and Future Work

A Box-Jenkins SARIMA model and a highly sophisticated neural network LSTM model were analyzed. Log messages with a severity type of info, error and warn was tested. SARIMA was tested on untransformed data, 1st difference, natural log, and square root transformations.

Different parameter factors were taken into consideration before deciding which model to use. Those factors came from the results of the unit root, normality, heteroskedasticity, time series dependency, and seasonality tests. RMSE was used for the model accuracy measures. A 1st difference transformation was applied to the LSTM model.

Unit root tests for KPSS and ADF showed conflicting results for unit root. The ADF test always failed to reject the null hypothesis and concluded that that unit root existed through all of the tests. This, however, was not the case for KPSS which did show it both reject and accept its hypothesis. As the ADF test suffers from type 1 errors and near observation equivalence, it is recommended that another test like the PP test or the ADF-GLS test is implemented instead of the ADF test. The ADF test was initially chosen due to it being so popular in the research papers

When testing for seasonality it was evident from the results of the test in comparison to the results of the STL tests that the CH test was not able to detect seasonality or trend for the majority of cases. It was, however, a little better at predicting seasonality at the higher frequency level for monthly data over the hourly periods. The CH for seasonality needs the first transformation to be done on the data before it can be applied. It is not able to handle higher level seasonal dimensionality in the data. This test should be eliminated from the study as it was not the best tool of choice. It was unfortunate that the limitations of the CH test were not evident in the research papers first read and only after questioning the results of the tests did I find the necessary papers.

For time-series prediction the results of the models concluded that the SARIMA model was not suitable for modeling predictions due to the existing shift in the data after the first principle transformation was done. The LSTM model was far more superior and better suited to handle the shift in the data. It is recommended that a further transformation is done on the data to remove the existing seasonality or trend in the data.

A rolling twenty-four window two STD approach was used for anomaly detection. The LSTM model was able to better predict anomalies than that of SARIMA. It is recommended that a better-suited algorithm that supports a level shift in the data should be implemented like LS or TC. Other recommendations would be to try Principal Component Analysis (PCA) or Cooks distance.

A hybrid model of SARIMA and LSTM could be implemented so that the classical model can be able to better handle the seasonality in the data. As only info type events were analyzed for anomalies error and warn event type events should be tested in future studies.

For ARIMA model parameters a periodicity of twenty-four was only implemented. It is recommended that different periodicity values should be implemented that can catch the higher dimensional levels of trends and seasonality. As was observed from the seasonal decomposition graphs the weekly and monthly graphs are more pronounced for trend and seasonality than that of the daily graphs.

For the LSTM model parameters, it is recommended that further analysis be conducted by increasing the number of repeats that the model cycles through. It is also recommended to use a batch size greater than one to help the model predict better. As was noted after fifty epochs the model started to degrade. It is recommended that no further epoch increases are recommended. It is also recommended that further transformations are done on the data and applied to the LSTM model. Currently, only a first difference transformation was applied to the model.

The correlation on info type messages, CPU and memory was quite low and the evidence suggested that this should be rejected. Further correlation tests on CPU, memory and disk usage should be tested against the warn and error type events. There were thirty-two CPU's on the server. Correlation analysis should be further refined by looking at the correlation between each CPU and each log event type message as an overall percentage metric might hide a potential load on these anomalous events.A Pearson correlation test was used for the analysis. As the CPU metric did display seasonality while the info type events also displayed non-stationarity it would be better if a Kendal or Spearman's correlation was implemented instead.

References

Ahmad, S., Lavin, A., Purdy, S., & Agha, Z. (2017). Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, *262*, 134–147. Retrieved 2019-02-09, from https://www.sciencedirect.com/science/article/pii/S0925231217309864

Ahrens, H. (1988). Stigler, stephen m.: The history of statistics. the measurement of uncertainty before 1900. the belknap press of harvard university, cambridge, mass., & london 1986; xvi, 410 s. *Biometrical Journal*, *30*(5), 631-632. Retrieved 2019-05-05, from https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.4710300527 doi: 10.1002/bimj.4710300527

Aldrich, J., et al. (1997). Ra fisher and the making of maximum likelihood 1912-1922. *Statistical science*, *12*(3), 162–176. Retrieved 2019-05-05, from https://projecteuclid.org/download/pdf_1/euclid.ss/1030037906

Amer, M., & Goldstein, M. (2012). Nearest-neighbor and clustering based anomaly detection algorithms for rapidminer. In *Proc. of the 3rd rapidminer community meeting and conference (rcomm 2012)* (pp. 1–12). Retrieved 2019-01-02, from https://www.researchgate.net/publication/230856452_Nearest-Neighbor_and_Clustering_based_Anomaly_Detection_Algorithms_for_RapidMiner

Amineh Aminia, T. Y. W. (2008). Adaptive density-based clustering algorithms for data stream mining. In *Third international conference on theoretical and mathematical foundations of computer science* (pp. 620–624). Retrieved 2019-

03-01, from https://www.researchgate.net/publication/258442503_Adaptive_Density-based_Clustering_Algorithms_for_Data_Stream_Mining

Arendt, W. (1987). Vector-valued laplace transforms and cauchy problems. *Israel Journal of Mathematics*, *59*(3), 327–352. Retrieved from https://link.springer.com/article/10.1007/BF02774144

Ashot Vagharshakyan, J. A. (1999). *On hidden periodicities*. Birkhäuser-Verlag. Retrieved 2019-03-15, from https://doi.org/10.1007/BF01229990

Babu, C. N., & Reddy, B. E. (2014). A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data. *Applied Soft Computing*, *23*, 27–38. Retrieved 2019-05-18, from https://doi.org/10.1016/j.asoc.2014.05.028 doi: 10.1016/j.asoc.2014.05.028

Balke, N. S. (1993). Detecting level shifts in time series. *Journal of Business And Economic Statistics*, *11*(1), 81–92. Retrieved 2019-03-10, from http://www.jstor.org/stable/1391308

Berthold, M. R., & Höppner, F. (2016). On clustering time series using euclidean distance and pearson correlation. *CoRR*, *abs/1601.02213*, XXXXXX. Retrieved 2019-05-01, from http://arxiv.org/abs/1601.02213

Bishop, C. M. (1998). Latent variable models. In *Learning in graphical models* (pp. 371–403). Springer. Retrieved 2019-02-10, from https://link.springer.com/chapter/10.1007/978-94-011-5014-9_13

Bollerslev, T. (1986, april). Generalized autoregressive conditional heteroscedasticity. In Elsevier (Ed.), *Journal of econometrics* (Vol. 31, p. 307-327). Elsevier. Retrieved 2019-04-13, from https://www.sciencedirect.com/science/article/pii/0304407686900631?via%3Dihub

Bontemps, L., McDermott, J., Le-Khac, N.-A., et al. (2016). Collective anomaly detection based on long short-term memory recurrent neural networks. In *International*

conference on future data and security engineering (pp. 141–152). Retrieved 2019-05-02, from https://link.springer.com/chapter/10.1007/978-3-319-48057-2_9

Box, G., & Jenkins, G. (1970, 01). Time series analysis forecasting and control. *Journal of Time Series Analysis*, *3*. Retrieved 2019-04-13, from https://www.researchgate.net/publication/37877248_Time_Series_Analysis_Forecasting_And_Control doi: 10.2307/1912100

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*, *13*(2), 195–212. Retrieved 2019-06-01, from https://link.springer.com/article/10.1007/BF01246098

Chai, T., & Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature. In *Geoscientific model development.* doi:10.5194/gmd-7-1247-2014: Geoscientific Model Development. Retrieved 2019-04-13, from https://pdfs.semanticscholar.org/11c9/aefb2fa45b9fd3292454ff8de134cfd1c6b1.pdf?_ga=2.200847278.1423443195.1560210368-1258299768.1558470258

Chan, S., Chen, L., Chow, N., & Liu, H. (2005). An ancova approach to normalize microarray data, and its performance to existing methods. *Journal Bioinformatics and Computational Biology*, *3*(2), 257–268. Retrieved 2019-01-14, from https://doi.org/10.1142/S0219720005001041 doi: 10.1142/S0219720005001041

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 7. Retrieved 2019-04-10, from https://dl.acm.org/

Chen, X.-y., & Zhan, Y.-y. (2008). Multi-scale anomaly detection algorithm based on infrequent pattern of time series. *Journal of Computational and Applied Mathematics*, *214*(1), 227–237. Retrieved 2019-02-07, from https://www.researchgate.net/publication/222315042_Multi-scale_anomaly_detection_algorithm_based_on_infrequent_pattern_of_time_series

Cleveland, R., & Cleveland, W. (1990, 01). Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official ...*, *6*, 3-73.

Cook, R. D. (1977, 02). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18. Retrieved from https://doi.org/10.1080/00401706.1977.10489493 doi: 10.1080/00401706.1977.10489493

Davidson, R., & Mackinnon, J. (2012). *Review of econometric theory and methods*. Retrieved 2019-03-15, from https://www.researchgate.net/publication/246006084_Review_of_Econometric_Theory_and_Methods

Delany, S. J., Cunningham, P., & Coyle, L. (2005). An assessment of case-based reasoning for spam filtering. *Artificial Intelligence Review*, *24*(3-4), 359–378. Retrieved 2019-05-04, from https://link.springer.com/article/10.1007/s10462-005-9006-6

Demos, A., & Sentana, E. (1998). Testing for garch effects: a one-sided approach. In *Journal of econometrics* (Vol. 86, p. 97-127). Journal of Econometrics. Retrieved 2019-04-17, from https://www.sciencedirect.com/science/article/pii/S0304407697001103?via%3Dihub

Devi, M. I., Rajaram, R., & Selvakuberan, K. (2008). Generating best features for web page classification. *Webology*, *5*(1), 52. Retrieved 2019-02-06, from http://www.webology.org/2008/v5n1/a52.html

Dickey, D., & A. Fuller, W. (1979, 06). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*, 6. doi: 10.2307/2286348

Dominique Ladiray, B. Q. (2001). *Seasonal adjustment with the x-11 method*. Springer, New York, NY.

Dominique Ladiray, G. M., Jean Palate, & Proietti, T. (2018). Seasonal adjustment of daily data. *16th Conference Of IAOS*. Retrieved 2019-03-15, from http://www.oecd.org/iaos2018/programme/IAOS-OECD2018_Item_1-A-1-Ladiray_et_al.pdf

Elliott, G., Stock, J., & J. Rothenberg, T. (1996, 02). Efficient tests for an autoregressive unit root. *Econometrica*, *64*, 813-36. Retrieved 2019-03-15, from https://www.researchgate.net/publication/4898638_Efficient_Tests_for_an_Autoregressive_Unit_Root doi: 10.2307/2171846

Encyclopedia.com. (n.d.). *Karl pearson.* Retrieved 2019, from http://www-history.mcs.st-andrews.ac.uk/DSB/Pearson.pdf

Engle, R. (2001). Garch 101: The use of arch/garch models in applied econometrics. In *Journal of economic perspectives* (Vol. 15, p. 157-168). Journal of Economic Perspectives. Retrieved 2019-04-13, from https://www.aeaweb.org/articles?id=10.1257/jep.15.4.157

Eskin, E. (2000). Anomaly detection over noisy data using learned probability distributions. In *In proceedings of the international conference on machine learning* (p. 1-8). Retrieved 2019-04-01, from https://www.researchgate.net/publication/221345596_Anomaly_Detection_over_Noisy_Data_using_Learned_Probability_Distributions

Fisher, R. A. (1937). Professor karl pearson and the method of moments. *Annals of Eugenics*, *7*(4), 303-318. Retrieved 2019-05-05, from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1937.tb02149.x doi: 10.1111/j.1469-1809.1937.tb02149.x

Flores, J. H. F., Engel, P. M., & Pinto, R. C. (2012). Autocorrelation and partial autocorrelation functions to improve neural networks models on univariate time series forecasting. In *Neural networks (ijcnn), the 2012 international joint conference on* (pp. 1–8). Retrieved 2019-04-05, from https://ieeexplore.ieee.org/abstract/document/6252470

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society Series B (Methodological)*, *34*(3), 350–363. Retrieved 2019-03-10, from http://www.jstor.org/stable/2985071

Fu, Q., Lou, J.-G., Wang, Y., & Li, J. (2009). Execution anomaly detection in distributed systems through unstructured log analysis. In *Data mining, 2009. icdm'09. ninth ieee international conference on* (pp. 149–158). Retrieved 2019-06-10, from https://www.researchgate.net/publication/220765301_Execution_Anomaly_Detection_in_Distributed_Systems_through_Unstructured_Log_Analysis

George E. P Box, G. C. R., Gwilym M. Jenkins. (1976). *Time series analysis: Forecasting and control (holden-day series in time series analysis)* (4th ed.). Wiley-Blackwell. Retrieved 2019-03-15, from https://www.wiley.com/en-us/Time+Series+Analysis%3A+Forecasting+and+Control%2C+5th+Edition-p-9781118675021

He, P., Zhu, J., He, S., Li, J., & Lyu, M. R. (2018). Towards automated log parsing for large-scale log data analysis. *IEEE Transactions on Dependable and Secure Computing*, *15*(6), 931–944. Retrieved 2019-04-03, from https://ieeexplore.ieee.org/document/8067504

Hochreiter, S., & Schmidhuber, J. (1997, 12). Long short-term memory. *Neural computation*, *9*, 1735-80. Retrieved 2019-06-10, from https://www.researchgate.net/publication/13853244_Long_Short-term_Memory doi: 10.1162/neco.1997.9.8.1735

Hodge, V. J., & Austin, J. (2004, Oct 01). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126. Retrieved 2019-04-13, from https://doi.org/10.1007/s10462-004-4304-y doi: 10.1007/s10462-004-4304-y

Hu, W., Liao, Y., & Vemuri, V. R. (2003). Robust anomaly detection using support vector machines. In *Proceedings of the international conference on machine learning* (pp. 282–289). Retrieved 2019-04-04, from https://www.researchgate.net/publication/2890287_Robust_Anomaly_Detection_Using_Support_Vector_Machines

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, *2*(3), 283–

304. Retrieved 2019-06-06, from https://link.springer.com/article/10.1023/A:1009769707641

Hylleberg, S. (1992). Modelling seasonality. In *Advanced texts in econometrics (paperback)* (p. 10). Oxford University Press. Retrieved 2019-03-15, from https://www.bookdepository.com/Modelling-Seasonality-Svend-Hylleberg/9780198773184

Ibidunmoye, O., Hernandez-Rodriguez, F., & Elmroth, E. (2015, 06). Performance anomaly detection and bottleneck identification. *ACM Computing Surveys*, *48*. Retrieved 2019-04-13, from https://www.researchgate.net/publication/280111583_Performance_Anomaly_Detection_and_Bottleneck_Identification doi: 10.1145/2791120

Jayathilake, D. (2012). Towards structured log analysis. In *Computer science and software engineering (jcsse), 2012 international joint conference on* (pp. 259–264). Retrieved 2019-03-03, from https://www.researchgate.net/profile/Dileepa_Jayathilake/publication/261165831_Towards_structured_log_analysis/links/55a3f2bf08aef604aa03c65b/Towards-structured-log-analysis

Jennifer Castle, D. H. (2010). A low-dimension portmanteau test for non-linearity. *Journal of Econometrics*, *158*(471), 231-245. Retrieved 2019-03-15, from 10.1016/j.jeconom.2010.01.006

Kang, D., Gangal, V., Lu, A., Chen, Z., & Hovy, E. (2017). Detecting and explaining causes from text for a time series event. *Conference on Empirical Methods in Natural Language Processing*, 2758–2767. Retrieved 2019-05-02, from https://www.semanticscholar.org/paper/Detecting-and-Explaining-Causes-From-Text-For-a-Kang-Gangal/2ed0beae20c2a30782ddc5d585ca144f3ae28aac

Ladiray, D., & Quenneville, B. (2001). Seasonal adjustment with the x-11 method. In *Seasonal adjustment with the x-11 method* (pp. 13–22). New York, NY: Springer New York. Retrieved 2019-03-15, from https://doi.org/10.1007/978-1-4613-0175-2_3 doi: 10.1007/978-1-4613-0175-2_3

Lasisi, T., & Shangodoyin, D. (2014, March). Arima methods of detecting outliers in time series periodic processes. *International Journal of Modern Mathematical Sciences*, *11*(1), 40–48. Retrieved 2019-03-10, from http://www.modernscientificpress.com/journals/ViewArticle.aspx?XBq7Uu+HD/8eRjFUGMqlReQBjnm7DCwFW6Zvbv+nCJUdav/pvZTKI4iEqzyhAtOr

Lee, M., Kang, Y. S., & Seok, J. (2018). The estimation of probability distribution for factor variables with many categorical values. *PloS one*, *13*(8), e0202547. Retrieved 2019-05-01, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6108477/

Ljung, G., & Box, G. (1978, 08). On a measure of lack of fit in time series models. *Biometrika*, *65*. Retrieved 2019-03-15, from https://www.researchgate.net/publication/246995234_On_a_Measure_of_Lack_of_Fit_in_Time_Series_Models doi: 10.1093/biomet/65.2.297

Lu, C.-T., Chen, D., & Kou, Y. (2003). Algorithms for spatial outlier detection. In *Third ieee international conference on data mining* (Vol. ICDM'03, pp. 597–600). IEEE Computer Society. Retrieved 2019-04-13, from https://pdfs.semanticscholar.org/7d5d/842df75a348f350d3f178c6f930be1cd02e1.pdf

Manevitz, L. M., & Yousef, M. (2001). One-class svms for document classification. *Journal of machine Learning research*, *2*(Dec), 139–154. Retrieved 2019-01-01, from https://dl.acm.org/citation.cfm?id=944808

Mehdiyev, N., Enke, D., Fettke, P., & Loos, P. (2016, 12). Evaluating forecasting methods by considering different accuracy measures. *Procedia Computer Science*, *95*, 264-271. Retrieved 2019-03-10, from https://www.researchgate.net/publication/309587915_Evaluating_Forecasting_Methods_by_Considering_Different_Accuracy_Measures doi: 10.1016/j.procs.2016.09.332

Mohd Razali, N., & Yap, B. (2011, 01). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21-33. Retrieved 2019-03-10, from https://www

.researchgate.net/publication/267205556_Power_Comparisons_of_Shapiro-Wilk_Kolmogorov-Smirnov_Lilliefors_and_Anderson-Darling_Tests

Nelson, C. R., & Plosser, C. R. (1982). Trends and random walks in macroeconmic time series: Some evidence and implications. *Journal of Monetary Economics*, *10*(2), 139 - 162. Retrieved 2019-03-15, from http://www.sciencedirect.com/science/article/pii/0304393282900125 doi: https://doi.org/10.1016/0304-3932(82)90012-5

Nicholas, D., Huntington, P., & Watkinson, A. (2005). Scholarly journal usage: the results of deep log analysis. *Journal of documentation*, *61*(2), 248–280. Retrieved 2019-03-06, from https://www.emeraldinsight.com/doi/abs/10.1108/00220410510585214

Peart, S. (2002). *The economics of w.s.jevons* (S. Peart, Ed.). Taylor And Francis. Retrieved 2019-03-15, from https://doi.org/10.4324/9780203022498

Peter C. B. Phillips, P. P. (1988). *Testing for a unit root in time series regression* (Vol. 75). Retrieved 2019-03-15, from http://www.jstor.org/stable/2336182

Puri, M. L., & Rao, C. R. (1976). Augmenting shapiro-wilk test for normality. In W. J. Ziegler (Ed.), *Contribution to applied statistics: Dedicated to professor arthur linder* (pp. 129–139). Basel: Birkhäuser Basel. Retrieved 2019-05-21, from https://doi.org/10.1007/978-3-0348-5513-6_13 doi: 10.1007/978-3-0348-5513-6_13

Robert J Hyndman, G. A. (2018). Forecasting: principles and practice. In *Jorecasting: principles and practice.* OTexts. Retrieved 2019-03-15, from OTexts.com/fpp2

Sánchez, M., & Peña, D. (2003, 01). The identification of multiple outliers in arima models. *Communications in Statistics - Theory and Methods*, *32*, 1265:1285. Retrieved 2019-03-10, from https://www.researchgate.net/publication/40754391_The_Identification_of_Multiple_Outliers_in_ARIMA_Models doi: 10.1081/STA-120021331

Schuster, A. (1898, 01). On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Journal of Geophysical Research*, *3*, 13-41. doi: 10.1029/TM003i001p00013

Sharma, N., & Gaud, N. (2015). K-modes clustering algorithm for categorical data. *International Journal of Computer Applications*, *127*(1), 46. Retrieved 2019-05-09, from https://pdfs.semanticscholar.org/1069/2c9b80be922903526682f8fae5ad6ffb68f6.pdf

Shin, Y., Kwiatkowski, D., Schmidt, P., & Phillips, P. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series are nonstationary. *Journal of Econometrics*, *54*, 159–178. Retrieved 2019-03-15, from https://www.sciencedirect.com/science/article/pii/030440769290104Y

Shipmon, D. T., Gurevitch, J. M., Piselli, P. M., & Edwards, S. T. (2017). Time series anomaly detection; detection of anomalous drops with limited features and sparse examples in noisy highly periodic data. *arXiv: Machine Learning*. Retrieved 2019-04-12, from https://www.semanticscholar.org/paper/Time-Series-Anomaly-Detection%3B-Detection-of-drops-Shipmon-Gurevitch/9c9ce2f8fecc6053176d908e5e431db65b5617c8

Siami-Namini, S., & Namin, A. S. (2018). Forecasting economics and financial time series: Arima vs. lstm. *CoRR*, *abs/1803.06386*. Retrieved 2019-03-10, from https://www.semanticscholar.org/paper/Forecasting-Economics-and-Financial-Time-Series%3A-Siami-Namini-Namin/75e895086f91a1a212a01dd8f426e535db01979f

Singh, K., & Upadhyaya, S. (2012, 01). Outlier detection: Applications and techniques. *International Journal of Computer Science Issues*, *9*. Retrieved 2019-04-13, from https://www.researchgate.net/publication/267964435_Outlier_Detection_Applications_And_Techniques

Statsoft. (2013). *Electronic statistics textbook*. StatSoft, Inc. Retrieved 2019-03-15, from http://www.statsoft.com/textbook/

Taylor, A. M. R. (2003). Robust stationarity tests in seasonal time series processes. In *Journal of business and economic statistics* (Vol. 21, p. 156-163). Taylor And Francis, Ltd. Retrieved 2019-03-15, from https://www.jstor.org/stable/1392360

Ting, S., Ip, W., & Tsang, A. H. (2011). Is naive bayes a good classifier for document classification? *International Journal of Software Engineering and Its Applications*, *5*(3), 37–46. Retrieved 2019-02-01, from https://www.researchgate.net/publication/266463703_Is_Naive_Bayes_a_Good_Classifier_for_Document_Classification

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, *7*(1), 1-20. Retrieved 2019-03-10, from https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3980070102 doi: 10.1002/for.3980070102

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. In *Climate research.* Retrieved from https://pdfs.semanticscholar.org/581d/0024eccf55493dd7d63554063a683bef6103.pdf?_ga=2.225010586.1423443195.1560210368-1258299768.1558470258

Wooldridge, J. M. (2001). Applications of generalized method of moments estimation. *Journal of Economic perspectives*, *15*(4), 87–100. Retrieved 2019-04-13, from https://pdfs.semanticscholar.org/640d/c5b00c526a7e2424b8e13cb712b4e5a95171.pdf

Youn, S., & McLeod, D. (2007). A comparative study for email classification. In *Advances and innovations in systems, computing sciences and software engineering* (pp. 387–391). Springer. Retrieved 2019-02-02, from https://link.springer.com/chapter/10.1007/978-3-540-30115-8_22

Zwiernik, P., Uhler, C., & Richards, D. (2017). Maximum likelihood estimation for linear gaussian covariance models. *Journal of the Royal Statistical Society: Series B*

(Statistical Methodology), *79*(4), 1269–1292. Retrieved 2019-06-06, from `https://arxiv.org/abs/1408.5604`