AN EXPLORATORY ANALYSIS ON TIME SERIES FORECASTING AND LOGISTIC REGRESSION MODELS

A Study On Cocoa Price And Credit Fraud Detection.

Adeola Adeniji
School of computing
National College of Ireland
X23104201@student.ncirl.ie

***Abstract*— In data analysis, the use of statistical methodologies has made it easy to make predictions. This is essential because it provides businesses with the ability to anticipate future events as well as to make well-informed decisions. In this paper, an investigation of two aspects in data analytics will be considered. This includes the use of various time series forecasting models and logistic regression models. In the first section of the paper, the use of time series models is developed based on historical data on the average cocoa price to forecast future values. Furthermore, the second section investigates classification models of fraud and non-fraud credit transactions using logistic regression. The study utilizes multi-modal methods to systematically assess several models against key performance metrics to determine which models are best suited. The evaluation metrics comprised of RMSE, MAE, Accuracy, Recall, Precision and F1- score which were used. By the means of this approach, this paper seeks to offer significant insights into effective data analytics for forecasting and classification problems.**

***Keywords*— Time Series Forecasting, Logistic Regression, Multi-Modal, RMSE, MSE, Accuracy, Recall, Precision and F1- score.**

I. INTRODUCTION

The ever-increasing amount of data presents opportunities as well as challenges for industries and businesses worldwide. In this context, the field of time series and logistical regression are harnessed to gain insights and to make informed decisions [1]. This paper focuses on two critical areas in data analytics. The first section of this paper begins with the application of the time series forecasting models, leveraging the historical data of the average cocoa price as published by the international cocoa organization from October 1995 to March 2024 [2]. By evaluating various predictive models such as Simple, Exponential Smoothing and Arima/ Serima will be highly considered. In the second section of this paper, it explores logistic regression modelling as a powerful tool for classifying data and distinguishing between fraudulent and legitimate credit card transactions. With its aim of creating a reliable model that can identify card theft with good accuracy by examining various attributes and patterns within the dataset. The overarching goal of this paper is to leverage statistical methodologies and machine learning techniques to perform predictive analysis in the domain of cocoa price forecasting and fraud detection. Through multi-model approaches, this paper systematically evaluates key performances metrics such as Root Mean Squared error (RMSE), Mean Absolute Error (MSR), Accuracy, Recall, Precision and F1- score, to determine the most suitable models for addressing forecasting and classification problems.

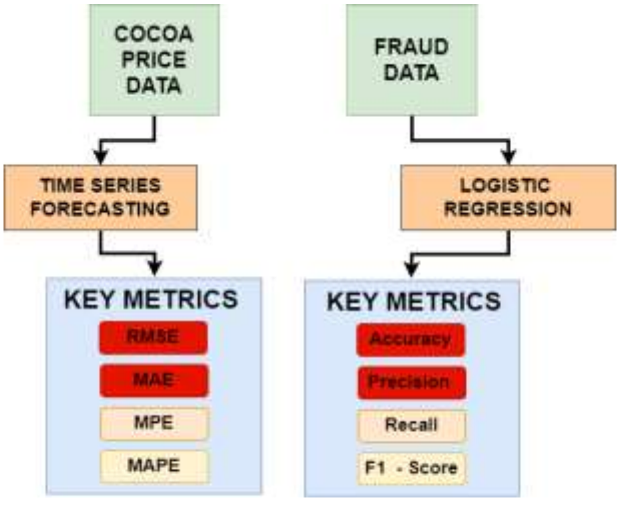


Figure 1: Overview of this Research Paper.

II. TIME SERIES ANALYSIS

One of the most important tools for drawing conclusions and forecasting trends is the time series data [3], which is made up of consecutive observations indexed by timestamps. This type of data, which is frequently measured and recorded on regular intervals, encapsulates a wide range of elements, each of which particularly contributes to the overall pattern. These elements include components of trend, seasonal, cyclical and irregularities. Durning a prolonged period, trend (T) indicates both upward and downward movements. Seasonal (S) indicates recurring changes over a defined interval of time such as daily, monthly, or seasonal variations. Cycle (C) signifies a cyclical trend or pattern that goes beyond seasonality. Data irregularity (I) considers noise or random fluctuations that occur in data making prediction difficult. [4]The observation series can be represented as y = T +C +S +I. The analysis of the time series cocoa price data would be typically unfolded through structured process methodology shown in figure 2.

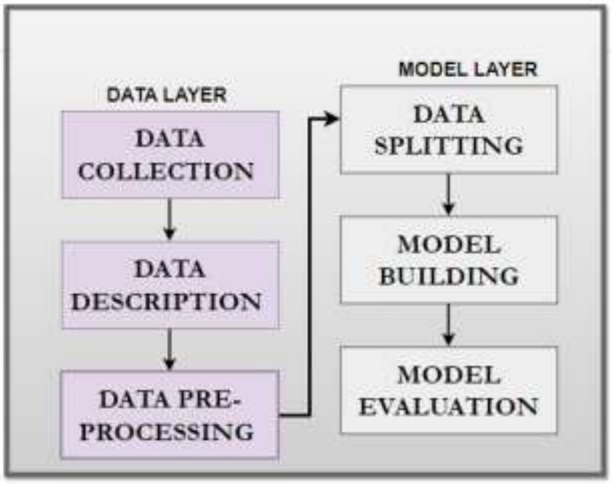
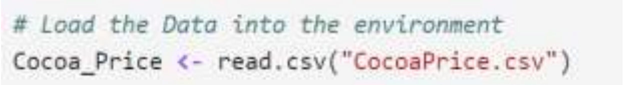


Figure 2: Roadmap of Data Analysis.

The following section of this paper delves into the time series analysis properly, where more insights into the cocoa price data would be harnessed

A. DATA COLLECTION

A historical data that represents the monthly time series of the Average Cocoa Price from the International Cocoa Organization (ICO) was collected. Covering the period from October 1994 to March 2024, in this process, a comma separated file was extracted and with the use of an integrated development environment called R, the exploration of cocoa price historical data was made possible.



```
# Load the Data into the environment
Cocoa_Price <- read.csv("CocoaPrice.csv")
```

Figure 3: Data Collection of the Cocoa Price Data into R Environment.

B. DATA DESCRIPTION

An initial preliminary assessment was conducted on the on the cocoa price data to better understand it characteristics. The dataset consists of two columns: Data and Price, with a class of 'ts' indicating time series formart. It started form the period of October 1994 and ends in March 2024 with a frequency of twelve-month intervals. Appreciate visualization techniques using plots such as histograms, boxplot, and seasonality decomposition have been employed to better understand the underlying attributes of the data. For instance figure 3. Shows the underlying trends, seasonal patterns and random fluctuations that contributes to the overall behavior of the data.

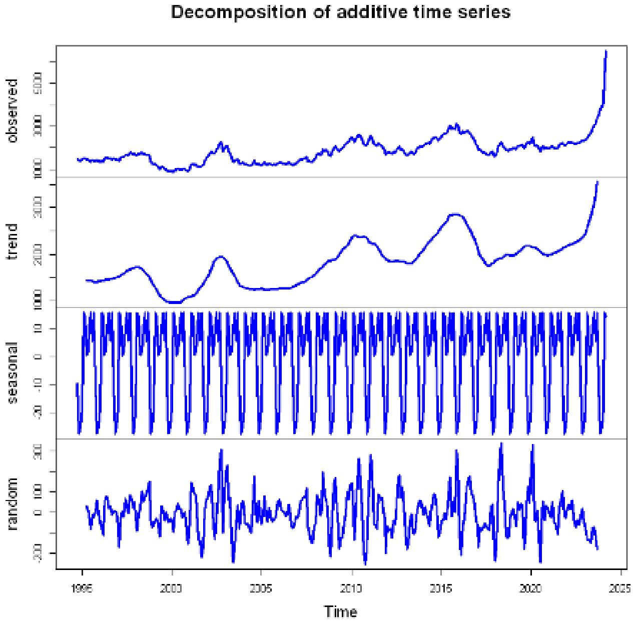
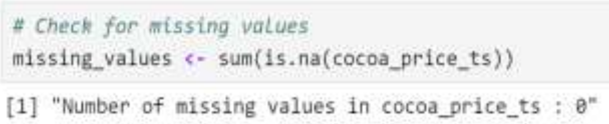


Figure 3: Decomposition of Additive time series for Cocoa Price from 1995 to 2025.

C. DATA PRE-PROCESSING

Before proceeding with the model building section, certain preprocessing steps were undertaken to ensure a quality, suitable and satisfactory analysis. The data column was converted to date formart using the "**.as.Date()**" function in the formart of " %Y-%M-%D" ensuring consistency that facilitates time-based analysis. Other steps such as handling missing values, checking for outliers and normalization were not overlooked, ensuring a clean structured and ready for modelling



```
# Check for missing values
missing_values <- sum(is.na(cocoa_price_ts))
```

```
[1] "Number of missing values in cocoa_price_ts : 0"
```

Figure 4: Checking for missing values in cocoa price data.

D. DATA SPLITTING

The time series data was divided into two parts, the "Training" and "Testing" Sets , for the advancement of model creation and evaluation. The historical data used for the training set spanned from October 1995 to September 2023. While the test sets consist of data for the last six months in October 2023 to March of 2024 to evaluate predicted ability of unforeseen data. The split allows accuracy forecasting and sound decision-making to be valid and for generalization of future periods.

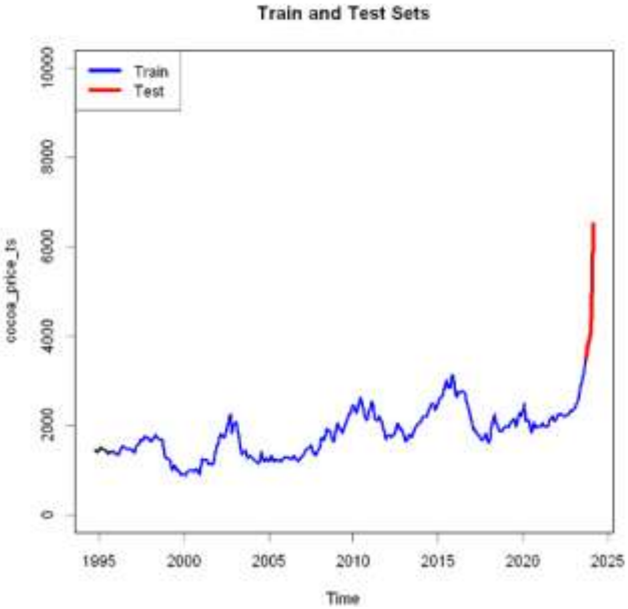


Figure 5: Spitted data for Train and Test Sets

E. MODEL BUILDING

This section focuses on building and evaluating various time series forecasting models for the cocoa price prediction. Three categories of models are taken into consideration during the investigation. The Simple Time Series Models, Exponential Smoothing Models, and ARIMA/SERIMA Models. Each of this models have a number of distinct models designed to represent various dynamics and patterns seen in the data. Through systematically testing of these models against the training data and performance evaluation, the goal is to find the most effective approach to accurately forecast and predict cocoa price.

i. Simple Time Series Models

Simple time series models as the name implies are basic elementary forecasting techniques that ignore complex relationships among data or trends. [3] . Based on historical averages and straightforward explorations, threes models offer clear cut forecasts. Four simple time series models have been used to provide a baseline for comparison against more advance techniques for the cocoa price prediction.

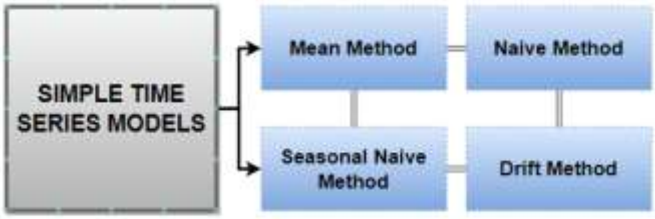


Figure 6: Simple Time Series Models used for Cocoa Price.

1) *Mean Method*

This model uses the average of previous measurements to predict future values. The forecast outcome points predictions with confidence interval are produced for every forecast horizon. In the analysis the point projection for the mean model is 1820.454 for October 2023, with a 95% confidence range that spans from 822.5087 to 2818.4.

2) *Naïve Method*

This model uses the most recent observation as a basis to forecast future values. The point forecast for each future period is simply the last observation. In using this model, then naïve produces point forecast that are mainly dependent on the most recent observed value, which lead to quite a straightforward predictions. The point projection for October 2023 is 3395.58

3) *Seasonal Nave Method*

Comparable to the naïve model, the season naïve model forecast values for the upcoming season by use the most recent observations for the season as well. The data's seasonal trends can be captured using this type of model. The result obtain projection is 2281.01 for October 2023.

4) *Drift Method*

This model applies the linear trend into the naïve forecast by exploring the trends between the first and last observation. The forecast outcome for this model modified the previous result of the naïve forecast. The point projection is 3401.553 for October 2023.

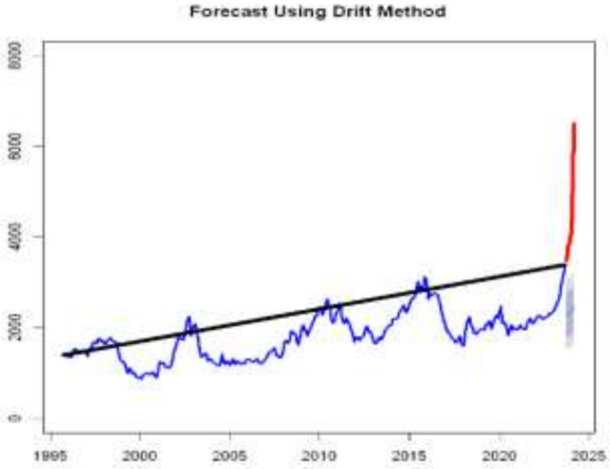
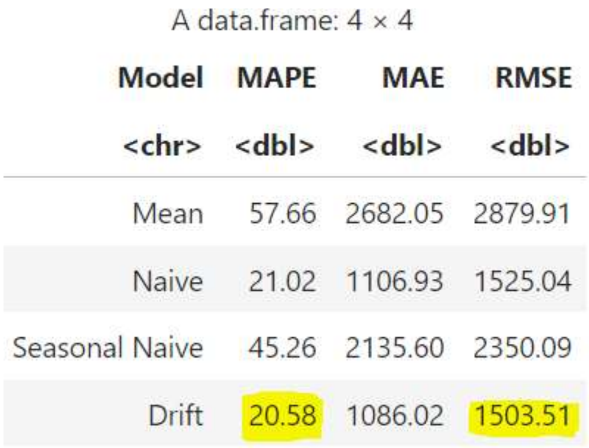


Figure 7: Forecast diagram using Drift Model.



A data.frame: 4 × 4

Model <chr>	MAPE <dbl>	MAE <dbl>	RMSE <dbl>
Mean	57.66	2682.05	2879.91
Naive	21.02	1106.93	1525.04
Seasonal Naive	45.26	2135.60	2350.09
Drift	20.58	1086.02	1503.51

Figure 8: Evaluation Results of the Simple Time Series Models.

Summary Of The Simple Time Series Models

Based on the performance parameters such as Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) on the four simple time series models namely Mean, Naïve, Seasonal Naïve, and Drift, were assessed. The drift model having the lowest MAPE of 20.58% among others models has outperformed the overall best in the simple time series category, demonstrating it greater ability to predict cocoa price in relation to the actual values. However, it worth noting that Naïve model came close with a near MARE of 21.02%. with the lowest MAE of 1086.02 and RMSE of 1503.51, the drift model is the best option among the simple time series model, closely followed by the Naïve model for forecasting cocoa price.

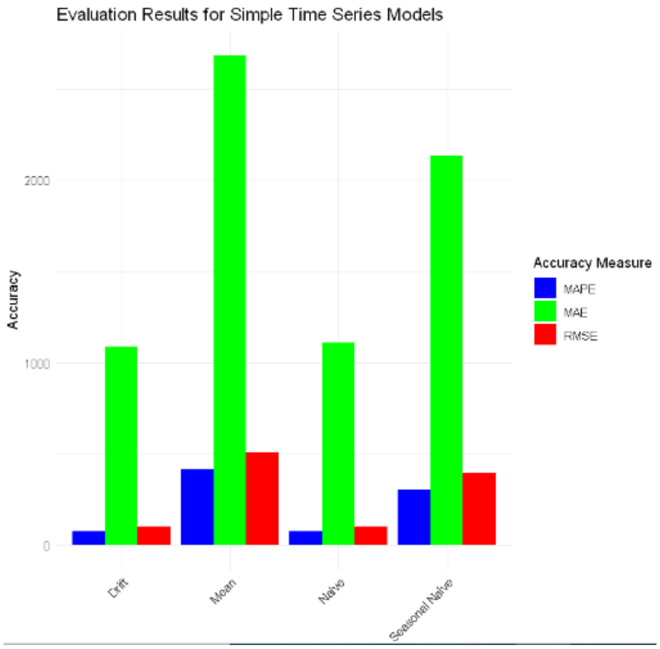


Figure 9: Histogram Results Distribution of the Simple Time Series Models.

ii. Exponential Smoothing Models

[4] is a type of time series forecasting techniques that assign exponentially decreasing weights to past observations. Exponential smoothing models are particularly useful for capturing patterns and seasonality in data. Various types of models have been investigated in this section, including Simple Exponential Smoothing, ETS models (Error, Trend, and Seasonality), and Holt-Winters models. Cocoa price forecasts can be thoroughly examined using each of the models, as each one provides a different method of identifying and predicting trends in the data.

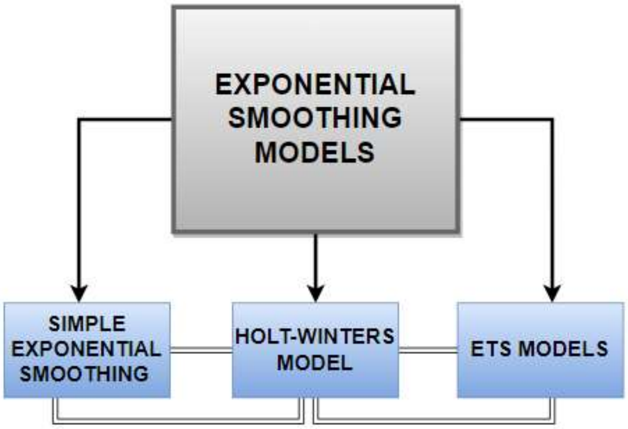


Figure 10: Exponential Smoothing Models used for Cocoa Price.

1) Simple exponential smoothing method (SES).

By averaging past observations exponentially, this method produced a smooth forecast. October 2023 has a point estimate of 3395.556. The prediction takes past data into account and smoothest out volatility to provide a fair estimate of future cocoa prices.

2) Simple exponential Smoothing method (SES- 0.05)

SES-0.05 employs a smaller smoothing value (alpha = 0.05), which makes it more sensitive to recent changes. October 2023 is anticipated to have a point value of 2400.231. Reacting faster to current price movement may produce a prediction that more prominently reflects short-term variations.

3) Simple exponential smoothing method (SES- 0.7)

However, SES – 07 places more emphasis on using a larger smoothing value (alpha = 0.7) to eliminate short-term oscillations. The point forecast for October is 3306.4448. Because the prediction is less vulnerable to recent changes because it is based on past data, it is smoother.

4) Holt- Winters method

This method is also known as the triple exponential smoothing method. By considering the trend and seasonal components it produces a better forecast than the SES. The point projection for October 2023 is around 3401.576, having a more improved model.

5) *ETS (ANN) method*

With the ANN (Additive Error, No Trend, No Seasonality) approach, which is only focused on error component, the ETS (Error, Trend, Seasonality) technique is compared. The prediction of points as of October 2023 is 3393.336. merely examining the data's random or error component. Without accounting for trends or seasonality, it offers a prediction.

6) *ETS (AAN) method*

ETS (Error, Trend, Seasonality) with AAN (Additive Error, Additive Trend, No Seasonality) method brings both error and trends components into the forecasting. It has a point projection for October 2023 to be 339.556. It also considers a linear trend in addition to the error component.

7) *ETS (AAA) method*

In AAA (Additive Error, Addictive Trend, Additive Seasonality), its method accounts for error in trends and seasonality in data. Having a projection for October 2023 to be 3388.3365. It has more accuracy because it considers all its components to have a comprehensive forecast.

8) *ETS (MMM) method*

ETS with MMM (Multiplicative Error, Multiplicative Trend, Multiplicative Seasonality) uses a makes use of an approach that uses multiplicative components for errors, trends, and seasonality to represent time series data. Its projection for October 2023 is 33799.947, offering a forecast that captures multiplicative in relationships within the data. It has proven to be a more flexible approach to modelling seasonal and trend patterns.

Summary Of The Exponential Smoothing Models

The exponential smoothing models made use of up to eight distinct models to explore the best fit model for forecasting cocoa prices. All the models exhibited various levels of accuracy. Among these methods, ETS (MMM) performed best with the lowest MAPE of 18.941 indicating the smallest average percentage error. This suggests that ETS (MMM) provided the most accurate prediction for both the simple time series and exponential smoothing models. Figure 11 shows the results of the exponential smoothing models.

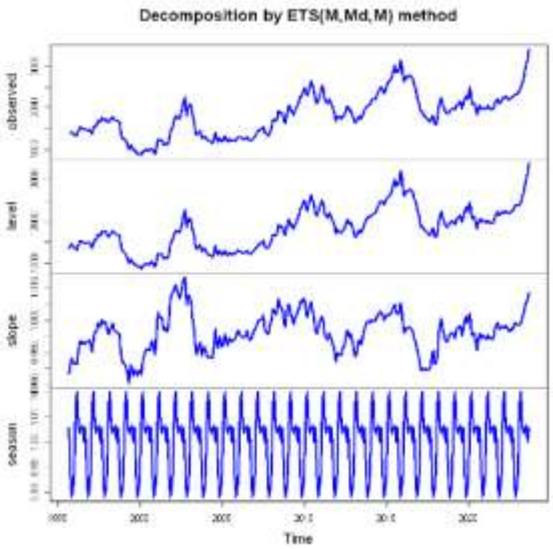
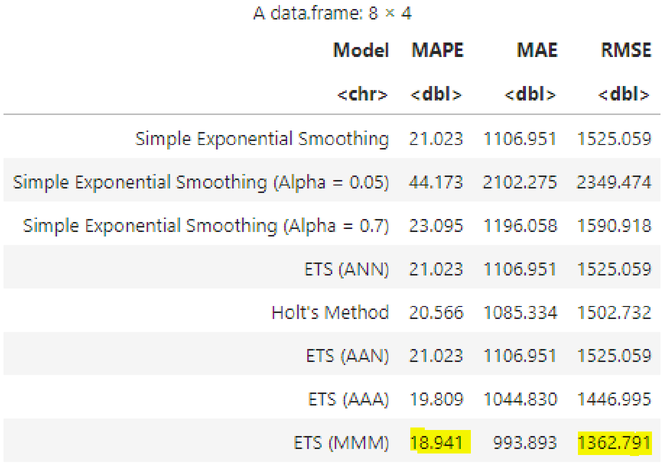


Figure 10: Forecast using ETS (MMM) Model .



A data.frame: 8 × 4

Model <chr>	MAPE <dbl>	MAE <dbl>	RMSE <dbl>
Simple Exponential Smoothing	21.023	1106.951	1525.059
Simple Exponential Smoothing (Alpha = 0.05)	44.173	2102.275	2349.474
Simple Exponential Smoothing (Alpha = 0.7)	23.095	1196.058	1590.918
ETS (ANN)	21.023	1106.951	1525.059
Holt's Method	20.566	1085.334	1502.732
ETS (AAN)	21.023	1106.951	1525.059
ETS (AAA)	19.809	1044.830	1446.995
ETS (MMM)	18.941	993.893	1362.791

Figure 11: Evaluation Results for Exponential Smoothing Models used for Cocoa Price.

iii. ARIMA/SERIMA Models

ARIMA(Autoregressive Integrated Moving Average). It combines autoregression, differencing and moving average components. While the SARIMA (Seasonal Autoregressive Integrated Moving Average) is an extension of Arima. Below are the models that were used to forest cocoa price.

1) *ARIMA (0,1,1)*

This model is a type of autoregressive integrated moving average model that accounts for the first order differencing and moving average components. Having a projection in October 2023 to be 3438.114, captures the trends for the cocoa price forecasting.

2) ARIMA (3,1,0)

This model includes autoregressive terms of up to lag 3, and first other differencing on moving average components. Having a projection of cocoa price in October to be 3429.137, capturing both trends and seasonality in the data.

3) ARIMA (2,10)

This model includes an autoregression term of 2 and first-order differencing with a projection for October 2023 to be 3429.809. with a moving average that both captures trends and seasonality of the cocoa price data.

4) ARIMA (4,1,0)

This model includes autoregressive terms up to lag 4 and first-order differencing, with a projection of 3441.276, on a moving average that both captures trends and seasonality of the cocoa price data.

5) ARIMA (3,1,1)

This model includes autoregressive terms up to lag 3 and with a projection for October to be 3429.058 on a moving average that both captures trends and seasonality of the cocoa price data.

6) ARIMA (4,1,1)

This model includes autoregressive terms up to lag 4 and with a projection for October to be 3443.973 on a moving average that both captures trends and seasonality of the cocoa price data.

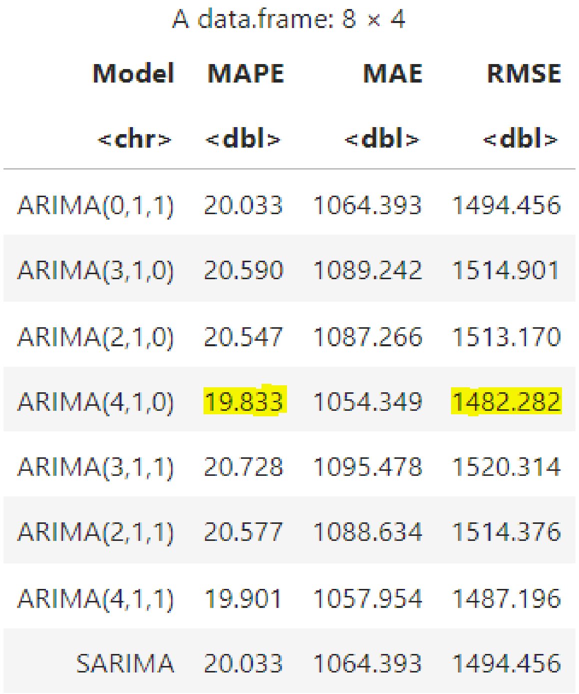
7) ARIMA (2,1,1)

This model includes an autoregression term of 2 and first-order differencing with a projection for October 2023 to be 3429.254. with a moving average that both captures trends and seasonality of the cocoa price data

8) SARIMA

The seasonal ARIMA (SARIMA) is an extension of the Arima model as it includes seasonal components , capturing both trends and seasonal data on the cocoa price forecasting. It had a projection in October to be 3438.114. It made use of multiplicative error, trend, and season components.

When predicting the cocoa price, using ARIMA and SARIMA models , a demonstration of consistency results was noticed. The mean absolute percentage error (MAPE) falls between 19.83% and 20.73%. Out of all the models used in this section, the ARIMA (4,1,0) and ARIMA (4,1,1) shows marginally lower MAPE values, indicating a superior performance. This can be seen in the figure below.



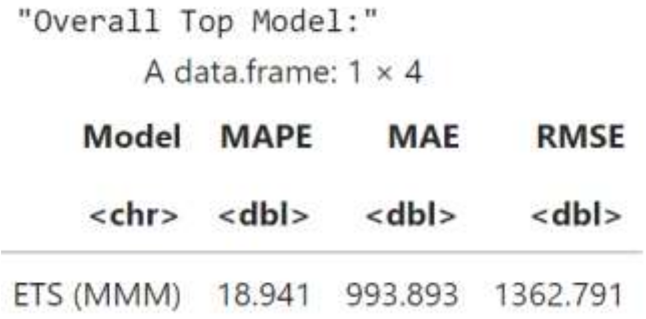
A data.frame: 8 × 4

Model	MAPE	MAE	RMSE
<chr>	<dbl>	<dbl>	<dbl>
ARIMA(0,1,1)	20.033	1064.393	1494.456
ARIMA(3,1,0)	20.590	1089.242	1514.901
ARIMA(2,1,0)	20.547	1087.266	1513.170
ARIMA(4,1,0)	19.833	1054.349	1482.282
ARIMA(3,1,1)	20.728	1095.478	1520.314
ARIMA(2,1,1)	20.577	1088.634	1514.376
ARIMA(4,1,1)	19.901	1057.954	1487.196
SARIMA	20.033	1064.393	1494.456

Figure 12: Evaluation Results for ARIMA AND SARINA Models used for Cocoa Price.

F. MODEL EVALUATION

In general the overall best model and top model for forecasting cocoa price is the ETS (MMM) model. It achieved a mean absolute error percentage error (MAPE) OF 18.41%, a mean absolute error (MAE) of 993.893 and a root mean squared error (RMSE) of 1362.79. This model indicates it superior accuracy in forecasting cocoa price.



"Overall Top Model:"

A data.frame: 1 × 4

Model	MAPE	MAE	RMSE
<chr>	<dbl>	<dbl>	<dbl>
ETS (MMM)	18.941	993.893	1362.791

Figure 13: Evaluation Results for the best Models for Cocoa Price.

III. Logistic Regression

In the world of financial security, the effectiveness of fraud detection techniques is paramount. In this paper, logistic regression has been used for categorizing transactions as fraudulent or not. With the use of predictive power as a technique to analyze credit card transaction data. The exploration analysis will begin by investigating logistic regression models and classifying its fundamental ideas with the use of fraud detection data. The aim of using this technique is to build a strong predictive model that can handle complex structure of transaction information and spot fraudulent activity by making use of its abilities to represent binary outcomes. During the exploration and usage of this technique, rigorous analysis of various attributes such as transaction time, amount, as well as other additional parameters would be considered.

With the use of an integrated development environment, focused to achieve the aim of this research, R language on Jupyter notebook platform has been instilled. In building the model, certain steps are being carried out for an effective analysis and evaluation.

A. DATA COLLECTION

A fraud.csv dataset was obtain containing anonymized data of over 283,726 credit cards transaction. In each transaction, details such as amount, time, and classification of fraud (Cass 1) or non-fraud (Class 0) were included. The dataset is highly imbalanced, with only 473 occurrences of fraud compared to non-fraud. This poses a challenge and would be addressed in subsequent analysis and model building.

B. EXPLORATORY DATA ANALYSIS

In this section, a proper investigation of the data was carried out, to better understand the structures and each variables properties, as well as to analyze their relationship to enable a smooth analysis.

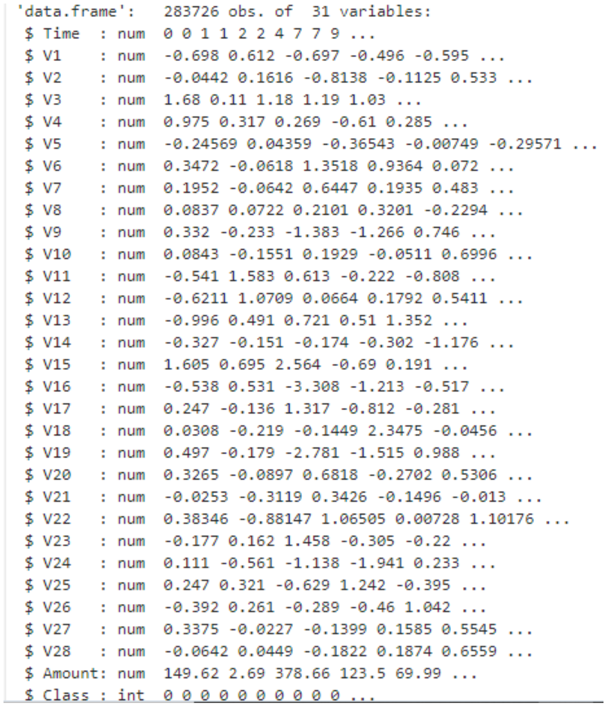
1) *Size and Attributes of fraud data.*

The fraud dataset consists of 283,726 unique rows and 31 columns. The column names include:

- Time: Time of each transaction (measured in seconds for two days)
- V1 – V28: Additional parameters that have been normalized to (0,1)
- Amount Each transaction amount in Euros.
- Class: A set of binary classifications of fraud and non-fraud.

2) *Data structure*

The structure of the dataset reveals that all attributes represent a numerical value of various aspects in the credit card transaction. Each attribute exhibits different ranges and distribution with Amount representing various amount and Class which indicates fraud classification.

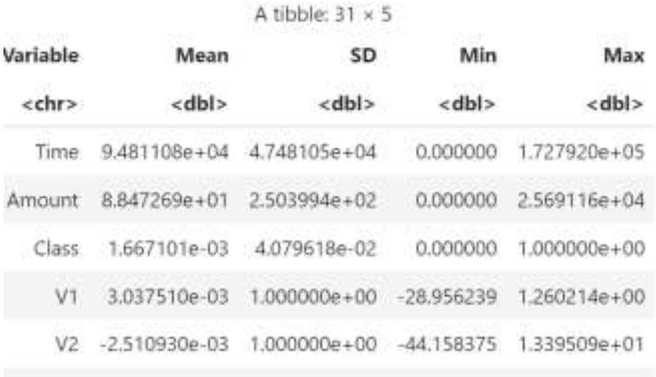


```
'data.frame':	283726 obs. of  31 variables:
 $ Time  : num  0 0 1 1 2 2 4 7 7 9 ...
 $ V1    : num  -0.698 0.612 -0.697 -0.496 -0.595 ...
 $ V2    : num  -0.0442 0.1616 -0.8138 -0.1125 0.533 ...
 $ V3    : num  1.68 0.11 1.18 1.19 1.03 ...
 $ V4    : num  0.975 0.317 0.269 -0.61 0.285 ...
 $ V5    : num  -0.24569 0.04359 -0.36543 -0.00749 -0.29571 ...
 $ V6    : num  0.3472 -0.0618 1.3518 0.9364 0.072 ...
 $ V7    : num  0.1952 -0.0642 0.6447 0.1935 0.483 ...
 $ V8    : num  0.0837 0.0722 0.2101 0.3201 -0.2294 ...
 $ V9    : num  0.332 -0.233 -1.383 -1.266 0.746 ...
 $ V10   : num  0.0843 -0.1551 0.1929 -0.0511 0.6996 ...
 $ V11   : num  -0.541 1.583 0.613 -0.222 -0.808 ...
 $ V12   : num  -0.6211 1.0709 0.0664 0.1792 0.5411 ...
 $ V13   : num  -0.996 0.491 0.721 0.51 1.352 ...
 $ V14   : num  -0.327 -0.151 -0.174 -0.302 -1.176 ...
 $ V15   : num  1.605 0.695 2.564 -0.69 0.191 ...
 $ V16   : num  -0.538 0.531 -3.308 -1.213 -0.517 ...
 $ V17   : num  0.247 -0.136 1.317 -0.812 -0.281 ...
 $ V18   : num  0.0308 -0.219 -0.1449 2.3475 -0.0456 ...
 $ V19   : num  0.497 -0.179 -2.781 -1.515 0.988 ...
 $ V20   : num  0.3265 -0.0897 0.6818 -0.2702 0.5306 ...
 $ V21   : num  -0.0253 -0.3119 0.3426 -0.1496 -0.013 ...
 $ V22   : num  0.38346 -0.88147 1.06505 0.00728 1.10176 ...
 $ V23   : num  -0.177 0.162 1.458 -0.305 -0.22 ...
 $ V24   : num  0.111 -0.561 -1.138 -1.941 0.233 ...
 $ V25   : num  0.247 0.321 -0.629 1.242 -0.395 ...
 $ V26   : num  -0.392 0.261 -0.289 -0.46 1.042 ...
 $ V27   : num  0.3375 -0.0227 -0.1399 0.1585 0.5545 ...
 $ V28   : num  -0.0642 0.0449 -0.1822 0.1874 0.6559 ...
 $ Amount: num  149.62 2.69 378.66 123.5 69.99 ...
 $ Class : int  0 0 0 0 0 0 0 0 0 0 ...
```

Figure 14: The structure of the Fraud Dataset.

3) *Descriptive Statistics*

The datasets have various ranges in their distribution. The transaction time spans from 0 to172,792 seconds with a mean average of about 94,811 seconds. However, the transaction amount ranges between 0 to 25,691.16 euros, or 88.87 euros. Due to the imbalance nature of the dataset, it showed only 0.167% of transactions classified as fraud. While other variables (V1-V28) also showed various distributions, reflecting the data complexity.



A tibble: 31 × 5

Variable	Mean	SD	Min	Max
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
Time	9.481108e+04	4.748105e+04	0.000000	1.727920e+05
Amount	8.847269e+01	2.503994e+02	0.000000	2.569116e+04
Class	1.667101e-03	4.079618e-02	0.000000	1.000000e+00
V1	3.037510e-03	1.000000e+00	-28.956239	1.260214e+00
V2	-2.510930e-03	1.000000e+00	-44.158375	1.339509e+01

Figure 15: Summary Statistics of few attributes.

4) *Class Distribution*

As stated earlier, the dataset exhibits a significant distribution of imbalance class with over 283,253 occurrences in the majority class (non-fraud) and 473 instances in the minority class(fraud). Figure 16 displays the distribution of a picture of this.

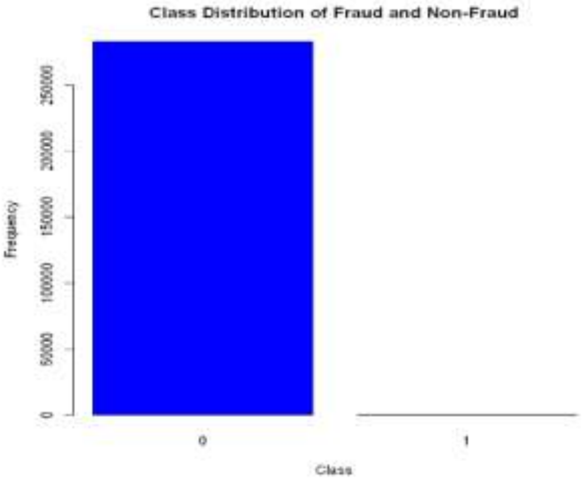


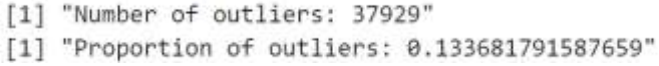
Figure 16: Class distribution of Fraud data.

5) *Missing Values*

With the use of "colSums(is.na)" made it possible to check for missing values in the dataset. At the end, each variable contained no missing entries.

6) *Outliers Detection*

A check was made to identify outliers within the dataset and a total of 37,929 approximately 13.37% of the observation is detected. These anomalies represent a proportion in the data, as in subsequent sections, they would be carefully examined to assess how they affect the modelling.



```
[1] "Number of outliers: 37929"
[1] "Proportion of outliers: 0.133681791587659"
```

Figure 17: Number and Proportion of outliers.

7) *Correlation Metrics*

The metrics for correlations provide a great deal of information with the relationships between independent and dependent variables. It shows a strong link to help when applying feature selection to find the best suitable model.

8) *Normalization*

This is a very critical part in data preprocessing. Normalization helps guarantee that every single feature contributes equally to the modelling. In the fraud data, certain variables were normalized such as Time, V1-V28 and Amount, as this would make them stable for the application of the machine learning model.

9) *Label Encoding*

The dependent variable also knows as the target variable used in the modelling process is the "Class" variable within fraud data. The use of "**. asfactor**" as a label encoding technique to transform the numerical variable class into a categorical level, to enable subsequent interpretation of modelling task to be easier.

10) *Data Balancing*

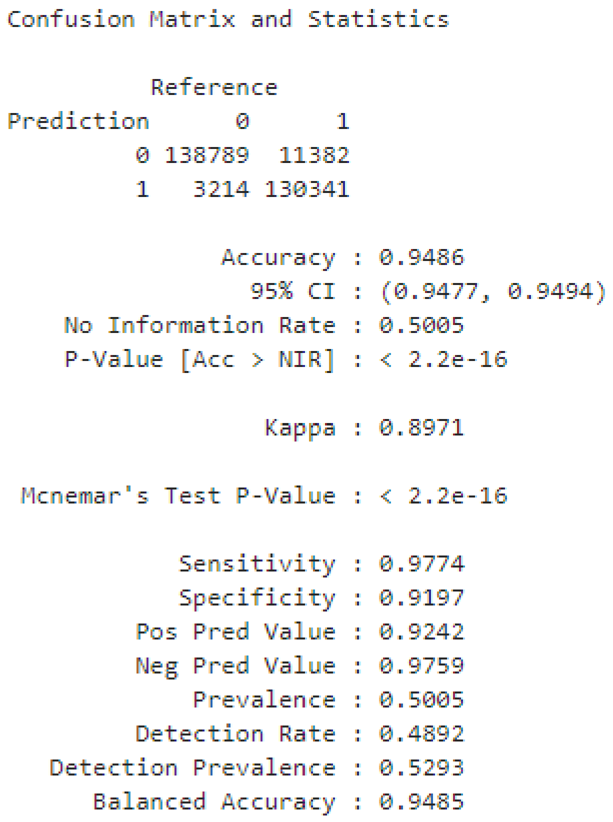
As discussed earlier, the imbalance of distribution of the dataset between the majority class (0) and the minority class (1). The utilization of Over-Sampling technique was applied. In this case, a balance distribution achieved by creating a synthetic instance of the minority class with the synthetic minority of the over-sampling technique (SMOTE). After the technique has been applied, both classes now contain about 142,003 and 141,723 distributed instances.

C. MODELLING

Building upon the findings from the precious section, this section delves into developing predictive models to get the best fitted model to clarify credit card transactions as fraudulent or non-fraudulent. The generated models would be evaluated using different parameters including accuracy, precision, recall and F1-score. As the selected model would be the best performed model, ensuring the development of an effective fraud detection system.

1) First Model

With the use of logistic regression to build the first model, all variables in the fraud dataset were considered to enable predict the "Class" variable, to distinguish between fraudulent or non-fraudulent credit card transactions. This model produced an impressive accuracy of about 94.86%. Also having a sensitivity of 97.74% and specificity of 91.97%. However, more evaluation would be considered to evaluate the significance of some specific variables as the model can be further improved.

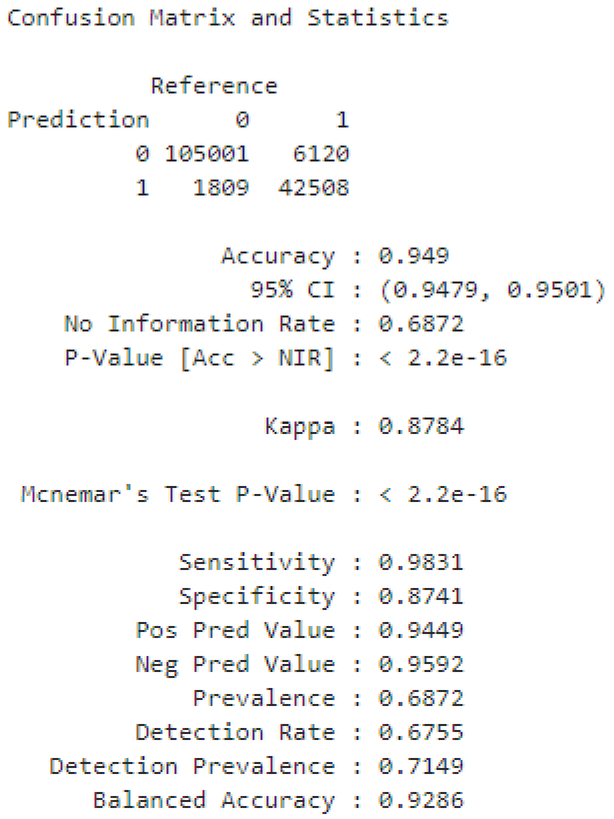


```
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 138789  11382
         1   3214 130341

               Accuracy : 0.9486
                 95% CI : (0.9477, 0.9494)
    No Information Rate : 0.5005
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8971

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9774
            Specificity : 0.9197
         Pos Pred Value : 0.9242
         Neg Pred Value : 0.9759
             Prevalence : 0.5005
         Detection Rate : 0.4892
   Detection Prevalence : 0.5293
      Balanced Accuracy : 0.9485
```

Figure 18: Overall performance of the First Model.

2) Second Model

To achieve a second model, two vital issues had to be addressed.: Outliers anomalies and second is the feature redundancy and complexity in the data. Earlier during the exploratory analysis carried out, outliers were discovered. Over 13% identified as anomalies in the data. In addressing this issue, the IQR approach was implemented across the data. As a result of this, 37929 instances were eliminated ensuring robustness in subsequent analysis. Additionally, the second issue addressed was the feature complexity and redundancy. The solution used was features selection. The reduction of subsets features from the original fraud data, and this was done based on their correlation with the target variable “Class”. This with the aim of improving the model’s performance by selecting only the most relevant features. A threshold of 0.05 was defined resulting in 28 variables out of 31.



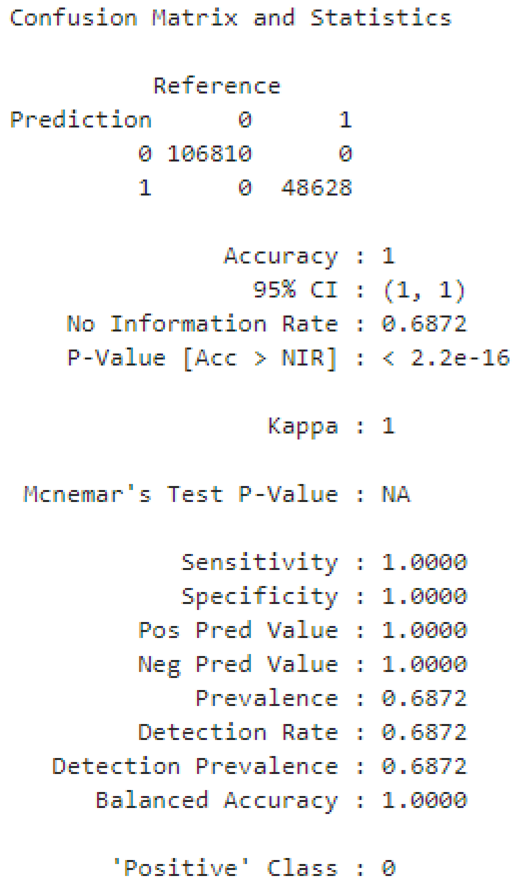
```
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 105001   6120
         1   1809  42508

               Accuracy : 0.949
                 95% CI : (0.9479, 0.9501)
    No Information Rate : 0.6872
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8784

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9831
            Specificity : 0.8741
         Pos Pred Value : 0.9449
         Neg Pred Value : 0.9592
             Prevalence : 0.6872
         Detection Rate : 0.6755
   Detection Prevalence : 0.7149
      Balanced Accuracy : 0.9286
```

Figure 19: Overall performance of the Second Model

The second model has an accuracy of about 94.9%, sensitivity of 98.31 and specificity of 87.41%, demonstrating a little enhancement in the accuracy from the first model. Further analyses were explored to investigate whether accuracy can be improved.

3) Third Model

The second model performed slightly better than the first model in terms of accuracy, true negative predictions, and Kappa values as both exhibited limitations in precision and recall values. The third model is created by using a novel approach model. This method is known as dimensionality reduction through the application of Principal Component Analysis (PCA). The use of this approach is to improve the previous models. This method helps by reducing the dimensions of fraud dataset. This can be achieved by retaining the maximal data variance while minimizing information loss. The third model explored the machine learning elements in which new components were created from the original data. As effective as this method can get for reducing dimension and eliminating autocorrelation from data, it cannot be applicable to every dataset. To guarantee that the fraud dataset is suitable in using PCA, prior investigations were carried out. Firstly, the check to verify if the sample data (fraud) was large enough. Secondly, Bartlett's test of sphericity (P, < 0.05), Kaiser-Meyer-Olkin (KMO) > 5, and Pearson coefficient (0.3) were also investigated. These tests were applied on the data as all outcomes satisfied the PCA conditions.



```
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 106810      0
         1      0  48628

               Accuracy : 1
                 95% CI : (1, 1)
    No Information Rate : 0.6872
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.6872
         Detection Rate : 0.6872
   Detection Prevalence : 0.6872
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

Figure 20: Overall performance of the Third Model

The figure above shows the statistical results of the third model. With the application of principal component analysis, a flawless performance was achieved. The result had perfect accuracy, sensitivity, specificity as well as predictive values. Having perfect precision, the model exhibited an ideal classification, effectively distinguishing t\between transactions that are fraudulent and those that are not.

D. EVALUATION.

In the final assessment using three distinct predictive models the third model employing PCA showcased an impeccable performance with 100% accuracy. However, the first and second model fall short, underscoring the efficacy of PCA. The table below shows an overview of the result.

LR Model	Accuracy	Sensitivity/ Recall	Specificity	Kappa
1st Model	0.948%	0.9774	0.9197	0897
2nd Model	0.95%	0.9831	0.8741	0.8784
3rd Model	100%	100%	100%	100%

***Table1**: Overall Evaluation using Logistic Regression.*

IV. CONCLUSION

In conclusion, this research was successful in achieving its goals. First, in exploring different forecasts on cocoa Price using time series models and secondly using logistic regression analysis to create various classification models on fraud and non-fraud card transactions. A range of statistical tests and techniques were used to create intermediate models for time series forecasting as well as model construction utilizing a logistic regression algorithm. The final models were then selected from among the several created models by use of various metrics and criteria for performance evaluation. These models, however, are just one potential solution. Making room to research, comprehend, test, and use the many other algorithms and techniques that may be used with larger datasets to conduct time series forecasting and logistic regression analysis.

V. References

[1] N. V. C. Y. J. G. J. W. Zhi-Hua Zhou, "Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives," IEEE Standards Online, 2014 .

[2] T. I. C. O. (ICCO), "Cocoa Market Report Daily Prices," International Cocoa Organization, February 2024. [Online]. Available: https://www.icco.org/.

[3] C. Horn, "TIME SERIES II," Horn, Christian, 2024.

[4] S. solutions, "An Introduction to Exponential Smoothing for Time Series Forecasting in Python," 2023. [Online]. Available: https://www.simplilearn.com/exponential-smoothing-for-time-series-forecasting-in-python-article.

[5] D. P. P. Dr. Avishek Pal, Practical Time Series Analysis, Packt Publishing Ltd. Copyright., 20117.

[6] F. K. L. a. N. M. Diebold, "TIME SERIES ANALYSIS," American Journal of Orthodontics and Dentofacial Orthopedics, 2026.