Table of Contents

[1. Purpose 3](#_Toc52558817)

[2. Methodology 3](#_Toc52558818)

[2.1a Decomposition of Retail Sales Data 3](#_Toc52558819)

[2.2 Forecast from Long Horizon Regression Model 3](#_Toc52558820)

[2.3 Forecast from Logistic Regression Model 4](#_Toc52558821)

[2.4 Forecast from ETS 4](#_Toc52558822)

[2.5 Forecast from ARIMA 4](#_Toc52558823)

[2.6 Realtime Forecast using Linear Regression Model 4](#_Toc52558824)

[3. Analysis 4](#_Toc52558825)

[3. Limitations 4](#_Toc52558826)

# 1. Purpose

The purpose of this report is to create a predictive model to predict future retail sales utilizing data from January 1992 to June 2020. Being able to predict future retail sales can help both retail companies and creditors. It enables retail companies to strategize, plan inventory purchases and prepare marketing campaigns, and on the creditors side, it assists in determining how much credit should be provided to retail companies whilst minimising loss due to bad loans.

Long horizon forecasts and real time forecasts will both be provided. Long horizon forecasts will realistically be used to set budgets and credit limits and real time predictions will assist in updating forecasts as more actual information is provided.

# 2. Methodology

1. Exploratory Data Analysis of retail data
   1. Decomposition of Retail Sales data
2. Forecast from Linear Regression Model
3. Forecast from Logistic Regression Model
4. Forecast from ETS
5. Forecast from ARIMA
6. Realtime Forecast using Linear Regression Model

## 2.1a Decomposition of Retail Sales Data

Analysing the data by using time series techniques it is observed that the trend is growing, and raw data shows seasonality with peaks and troughs in sales (Appendix X). When data is seasonally adjusted (Appendix X), the growth trend looks smoother and growing with few anomalies representing two recessions, approx. 2001 and 2007. Another stark observation is recent spike in sales due to Covid-19. On splitting the data further (Appendix X) it can be observed that overtime the magnitude of seasonality is increasing.

The decomposition of retail data as shown in Appendix X demonstrates the trends that have been occurring since 1992. Retail sales has seen a steady seasonally adjusted growth. Seasonal effects are also becoming more extreme and it can be seen in Appendix X that purchase behavior is shifting. Consistently demonstrated over time, November and especially December see surges in retail sales however this is offset by a decline of sales throughout January, February, September and October.

## 2.2 Forecast from Long Horizon Linear Regression Model

The long horizon regression model forecast in appendix 4 visually demonstrates that it systematically underestimates retail sales. The forecasted trend is quite flat whilst the actual sales increases quite steadily.

## 2.3 Forecast from Long Horizon Logistic Regression Model

The long horizon logistic regression sees an improved predicted forecast, however it still underestimates the actual retail sales.

## 2.4 Forecast from ETS

## 2.5 Forecast from ARIMA

## 2.6 Realtime Forecast using Linear Regression Model

# 3. Analysis

1. Load your data into R and convert the data into the time series data type. Comment on the features of this time series data.

* Growing trend
* Seem to have a seasonal pattern with peak and trough

1. Split the data into the training set: January 1992 to December 2010; and the test set: January 2011 to June 2020. è done

# Limitations

-Model cant predict structural changes

1. Construct the following time series predictive models for the training set:
   1. Time series regression with trend and seasonality components.
   2. Exponential smoothing using ets() function.
   3. ARIMA model using the auto.arima() function.

The predictions for this category of retail sales is of special interest to banks and credit companies as credit cards are the most common form of payments for these purchases. They typically update their predictions as new data becomes available for timely management of cashflows and liquidity.

1. Construct real time prediction assessment for the three models above over the test set data. Summarize the predictive performance of the three models using statistical metrics and comment.

**Real time prediction plots**

Graphical user interface, chart, histogram

Description automatically generated

**Real time predictive performance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **RMSE** | **MAE** | **MAPE** | **MASE** | **ACF** |
| **Regression** | 10832.854 | 8539.930 | 18.858 | 1.851 | 0.658 |
| **Log Regression** | 3865.368 | 3171.411 | 8.965 | 0.687 | 0.491 |
| **Exp Smoothing** | 2094.822 | 1328.246 | 3.015 | 0.288 | -0.109 |
| **ARIMA** | 2176.496 | 1251.983 | 2.816 | 0.271 | -0.053 |

Banks and credit companies have asymmetric losses associated with prediction errors in this context. To be prudent, they would prefer an overestimate rather than an underestimate of the prediction. However, there are constraints on their liquidity and cashflow management such that a severe overestimate is also costly. Specifically, the credit portfolio management team would like a loss function expressed in relative term, with penalty such that

* -  An underestimate of the prediction is weighed by a factor of 5.
* -  An overestimate of the prediction that is less than 20% deviation from the

truth is weighed by a factor of 1.

* -  A severe overestimate of the prediction that is greater than 20% deviation

from the truth is weighed by a factor of 3.

5. Design a loss function that capture these preferences.

1. Evaluate your real-time predictive models using this user specific loss.
2. Comment on which predictive model perform best according to this

loss. Is this consistent with the conclusion from the statistical metrics?

**Bank loss function**

|  |
| --- |
| bank\_loss<-function(error,actual){  relative=100\*(error/actual)  underestimate=relative[(relative>=0)]  overU20=relative[ (relative<0) & (relative >= -20)]  overTheTop=relative[(relative < -20)]  loss=sum(5\*underestimate) + sum(1\*abs(overU20)) + sum(3\*abs(overTheTop))  return(loss/length(error))  } |

**Real time predictive performance combined with custom bank loss function**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **RMSE** | **MAE** | **MAPE** | **MASE** | **ACF** | **BankLoss** |
| Regression | 10832.854 | 8539.930 | 18.858 | 1.851 | 0.658 | 94.291 |
| Log Regression | 3865.368 | 3171.411 | 8.965 | 0.687 | 0.491 | 12.762 |
| Exp Smoothing | 2094.822 | 1328.246 | 3.015 | 0.288 | -0.109 | 10.731 |
| ARIMA | 2176.496 | 1251.983 | 2.816 | 0.271 | -0.053 | 8.986 |

Appendix

**Raw time series data feature**

Chart, histogram

Description automatically generated

**Seasonally Adjusted Retail Sales**

Chart, line chart

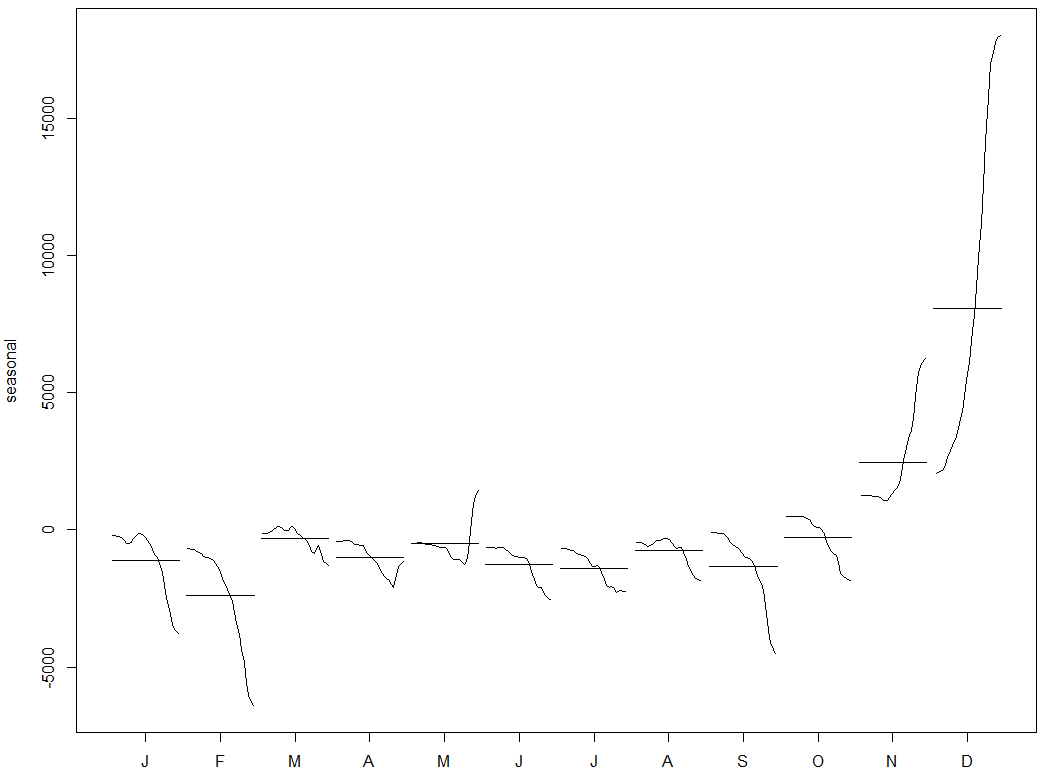
Description automatically generated

**APPENDIX 1 - Decomposition of Data**

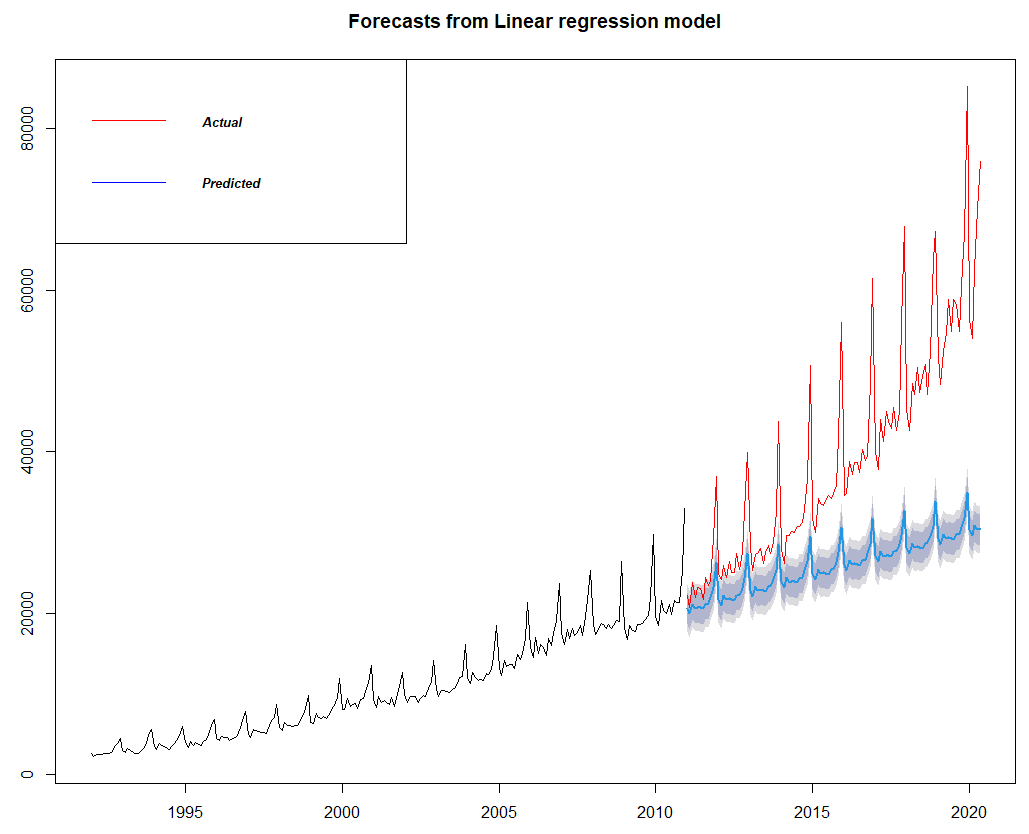
A picture containing chart, histogram

Description automatically generated

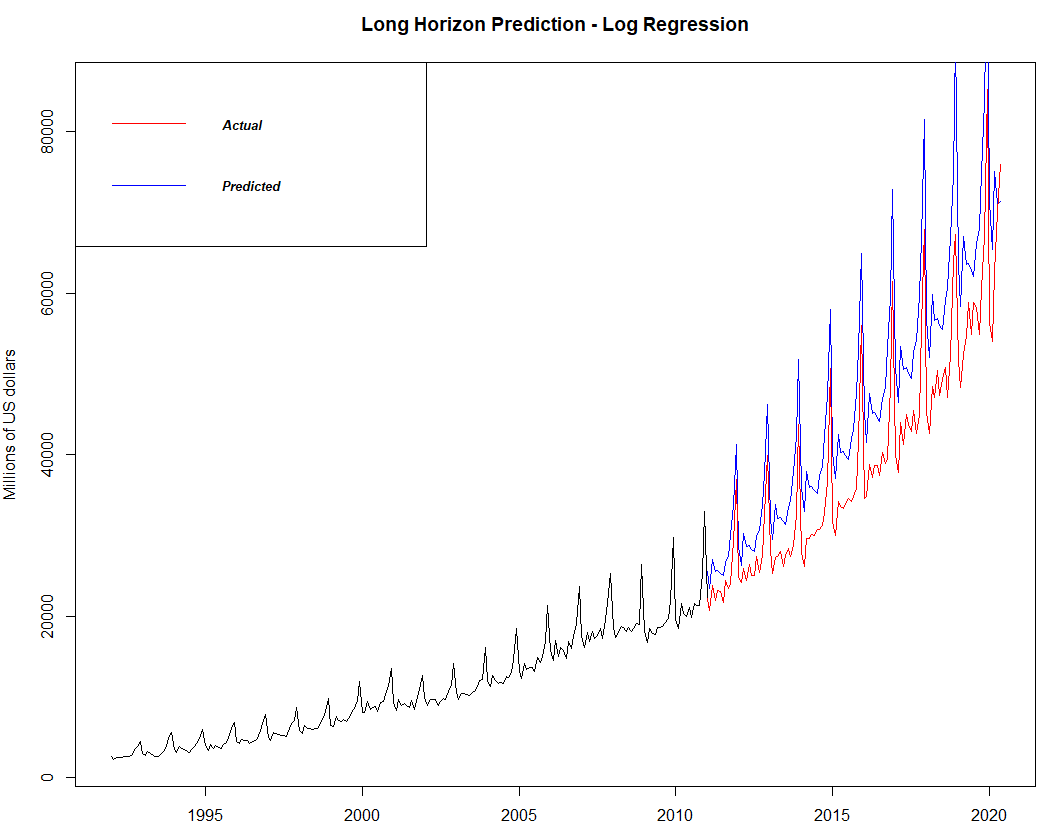
**APPENDIX 2 – Seasonal trends by month over time**



APPENDIX 3 - Forecast from Long Horizon Regression



Appendix 4 – log regression



**Long horizon forecast**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **RMSE** | **MAE** | **MAPE** | **MASE** | **ACF** |
| **Regression** | 17143.018 | 13558.705 | 29.609 | 2.939 | 0.807 |
| **Log Regression** | 8652.580 | 7750.469 | 19.550 | 1.680 | 0.669 |
| **Exp Smoothing** | 12753.083 | 9347.915 | 19.516 | 2.026 | 0.876 |
| **ARIMA** | 6286.037 | 3821.393 | 7.825 | 0.828 | 0.694 |