PREDICTIVE ANALYTICS

Syndicate Task #2 – Retail Sales

Syndicate 8

Tony Trinh (1099433)
Alvin Lee (1169514)
Varun Sharma (959620)
Jonno Lindsay (1026965)

Table of Contents

1 Purpose

The purpose of this report is first to analyse time-series data of retails sales of mail orders and online shopping between January 1992 and May 2020. The report then explores & compares several predictive models to forecast future sales and construct a creditors-defined loss function to measure the performance of those models further.

Being able to predict future retail sales can help retail companies and creditors. It enables retail companies to strategise, plan inventory purchases and prepare marketing campaigns. On the creditor's 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 precise information is provided.

2 Methodology

2.1 Analysing the characteristics of time series data

Over the 30-year period, retail sales appear to be growing at a steady rate (Appendix 1).

On decomposing the data further (Appendix 2), it can be observed that the seasonal effects are becoming more extreme with higher peaks and deeper troughs as time progresses. Additionally, it seems that purchase behaviour is shifting as depicted in Appendix 3 through the monthly sales trend over time. More specifically, November and especially December months have seen surges in retail sales; however, this is offset by a decline of sales throughout January, February, September and October.

Another observation is that there are specific periods in time, which cannot be explained by the seasonality or trend. The spike in the error component during 2016 is not so explainable by any world event, but the more recent spike in the error component could be attributable to the effects of Covid-19.

2.2 Fitting predictive models

The data is split into a training set (January 1992 to December 2010) and a test set (January 2011 to June 2020). After that, two sets of time-series models are fitted using the training set. The first set of models are long-horizon forecasts which will assist in initial budgeting & forecasting and the second set of models are real-time predictions which will benefit retailers and banks in managing cash flow and monthly planning.

The four models produced in each set are regression model, logistic regression model, exponential smoothing model and ARIMA model.

2.3 Long-horizon forecasts

The predictions for this category can help retailers and bankers alike to make planning decisions in advance for upcoming seasons. Retailers could plan a business expansion or how much inventory to purchase. At the same time, creditors can decide on how much to lend to the retailers, within their financial risk thresholds.

Appendix 4 shows that the linear regression model predicts growth as expected, but the difference between the predicted value and actual value widens as time passes. It underpredicts sales compared to the actual sales, which could result in lost opportunities for retailers.

Meanwhile, the logistic regression (Appendix 5) improves the prediction from the linear regression model. This is due to sales growing proportionally rather than by a fixed amount, which is something the logistic regression model accounts for. However, the model still tends to overpredict sales.

The exponential smoothing model (Appendix 6), ETS (M, A, M) also predicts a trend of sales growth. The difference in actual vs predicted sales increases as time passes, which resultantly depicts a prediction where forecasted sales are well below the actual sales achieved.

Demonstrated in Appendix 7, the ARIMA model automatically chooses (2,1,2) (0,1,0)[12] and predictions with this model are most accurate compared to other models for long-horizon predictions.

2.4 Long-horizon prediction performance

The performance of the four models and their ability to predict long-horizons are depicted in the table below. The model which performs the best across majority of metrics is the ARIMA model with its only underperforming metric relative to the other models being its ACF. The log regression has the lowest ACF score, however, the differences are not large enough to offset the strength of ARIMA model's predictive capabilities.

	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

2.5 Real-time Forecasts

Real-time prediction models benefit from the most recent information becoming available and thus, can predict more accurately.

As evident in Appendix 4 & 8, the predictive improvement of the linear regression model going from long-horizon to real-time is not significant compared to other models. As depicted in Appendix 9, 10 & 11, real-time forecasts by log regression, exponential smoothing and ARIMA, respectively, all show a remarkable improvement as compared to their respective long-horizon forecasts.

Similar to the results of the long-horizon forecast, it appears the ARIMA model provides the most accurate real-time forecast.

2.6 Real-time prediction performance

The table below highlights the performance of the models using real-time data. The ARIMA model once again outperformed the other models in all metrics.

The real-time data has also improved the ACF of all models, especially the ARIMA model with a value of -0.053, which indicates that past errors are now not systematically occurring and are less predictable.

	RMSE	MAE	MAPE	MASE	ACF
Regression	10832.854	8539.93	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

2.7 Specific loss function

A specific loss function expressed in relative term (Appendix 12 – bank loss function) has also been constructed as creditors have asymmetric losses associated with prediction errors i.e. moderate overestimation (under 20% deviation) is preferred over underestimation or severe overestimation (more than 20% deviation) of the prediction.

The result also shows that ARIMA is the best performing model, consistent with the conclusion from the statistical metrics.

	Bank Loss
Regression	94.291
Log Regression	12.762
Exp Smoothing	10.731
ARIMA	8.986

3 Analysis

Retail sales over the years is on an upward trend and the ability to forecast the trend and seasonality effects enable retail stores to strategise, plan and prepare inventory. It is just as crucial for banks and creditors to understand the effects, so they understand how much is needed to be lent out to businesses and manage their cashflow to reduce their own risks.

It is quite clear from the models constructed that generally long-horizon models tend to perform worse compared to real-time models. It is also clear from the predictions that the ARIMA model seemed to fare better in both scenarios. This could potentially be due to the fact that the changes in trends have been more extreme in the recent years. So not all historical information is necessary to predict the upcoming trends, with more weight being given to the more recent data points.

Furthermore, the introduction of an asymmetric loss metric also solidifies the fact that the ARIMA model is the one which should be used in future predictions as it outperforms the other models.

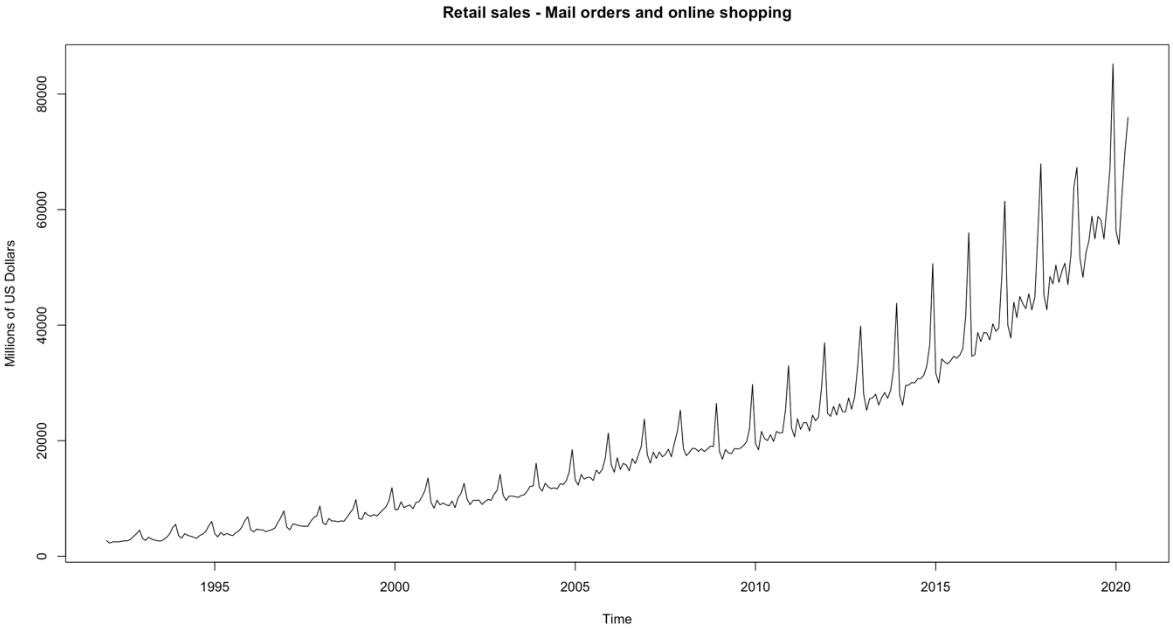
4 Limitations

The limitations of this analysis are the lack of other data points which might assist companies in their business plans. For example, adding more variables like post code or sub retail type to the data would assist both retail stores and creditors in their plans. At the moment, the analysis generalises that all retail sales follow the same pattern, and this may not be true. It is possible that online clothing retail sales perform stronger during certain times of the year compared to online homeware sales, but this is not incorporated in the current analysis.

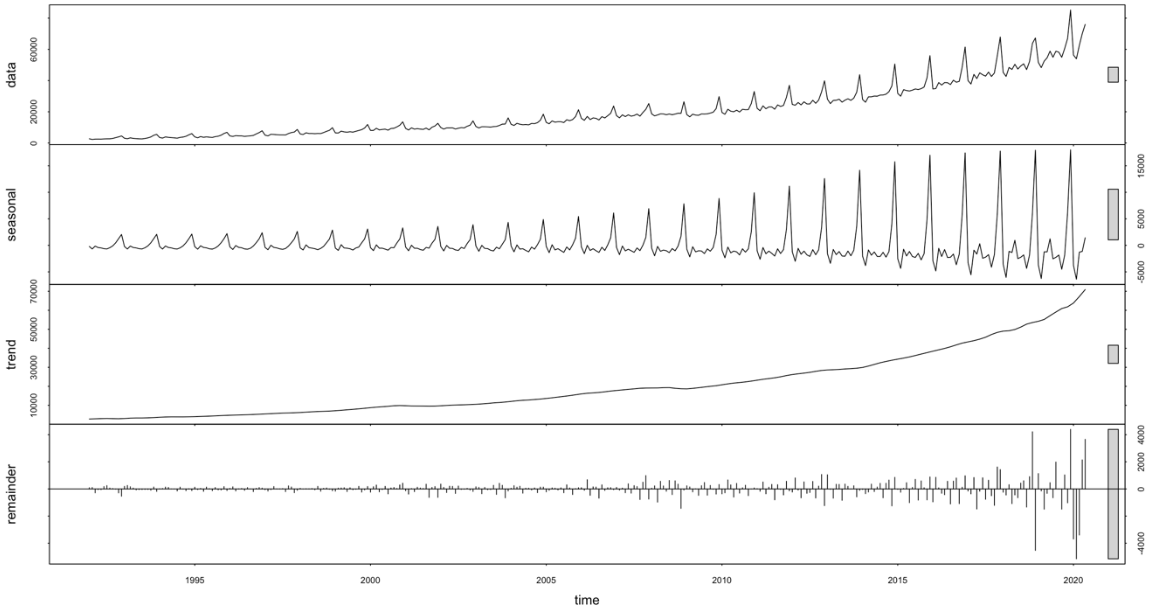
The models are also unable to cater for large one-off and structural changes that have occurred recently or are currently occurring. For example, Covid-19 is having wide reaching effects with lockdowns and generally poor customer sentiments. However, the models might not be able to account for these specific aspects as these effects would not show up in past data. If there are also large structural changes, the model may not immediately pick them up until a few years have elapsed.

5 Appendix

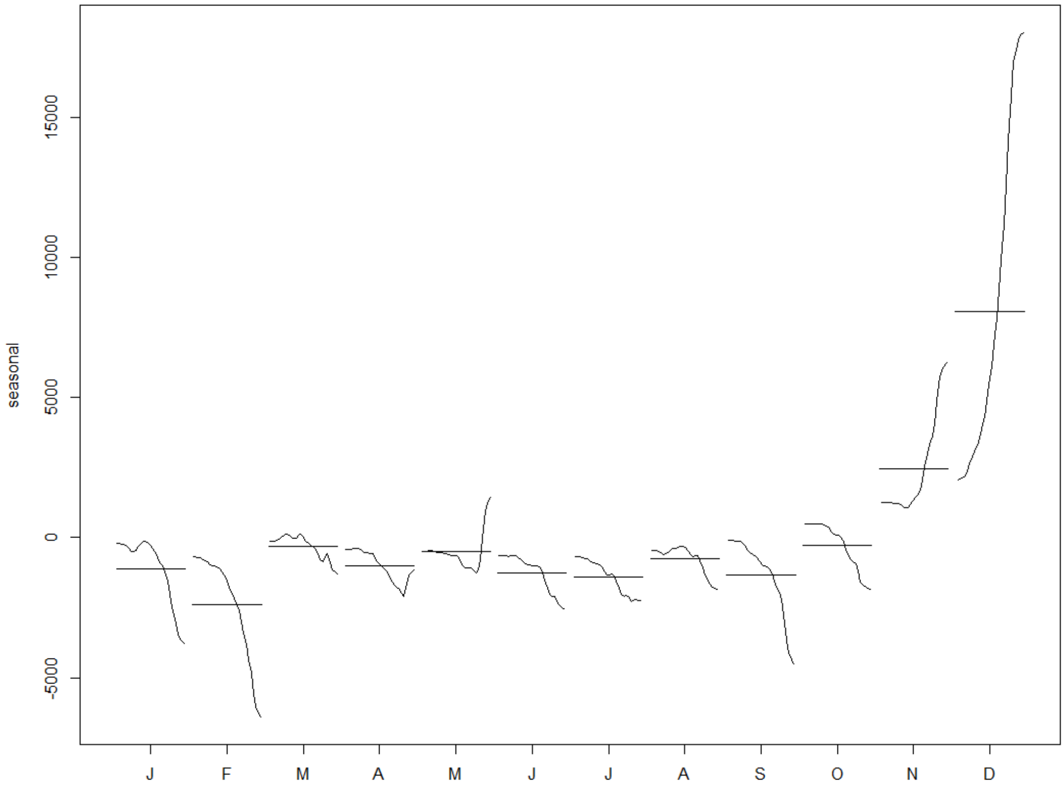
Appendix 1 – Retail Sales data over time



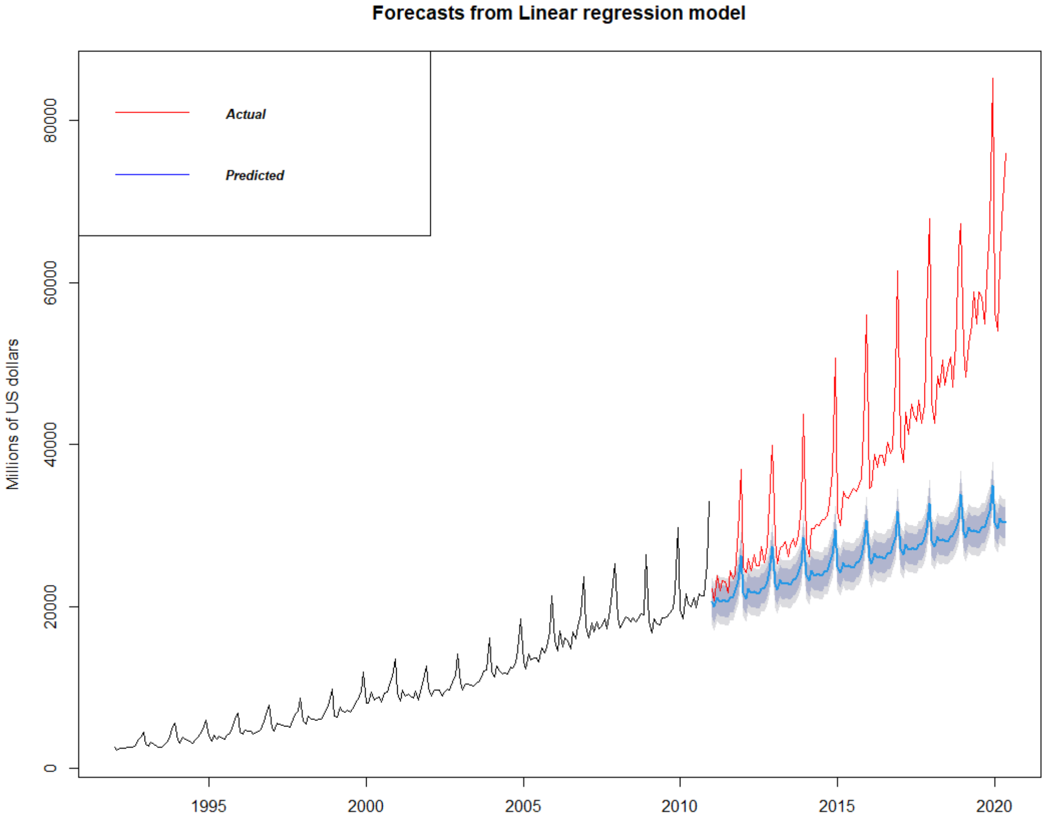
Appendix 2 – Decomposition of Data



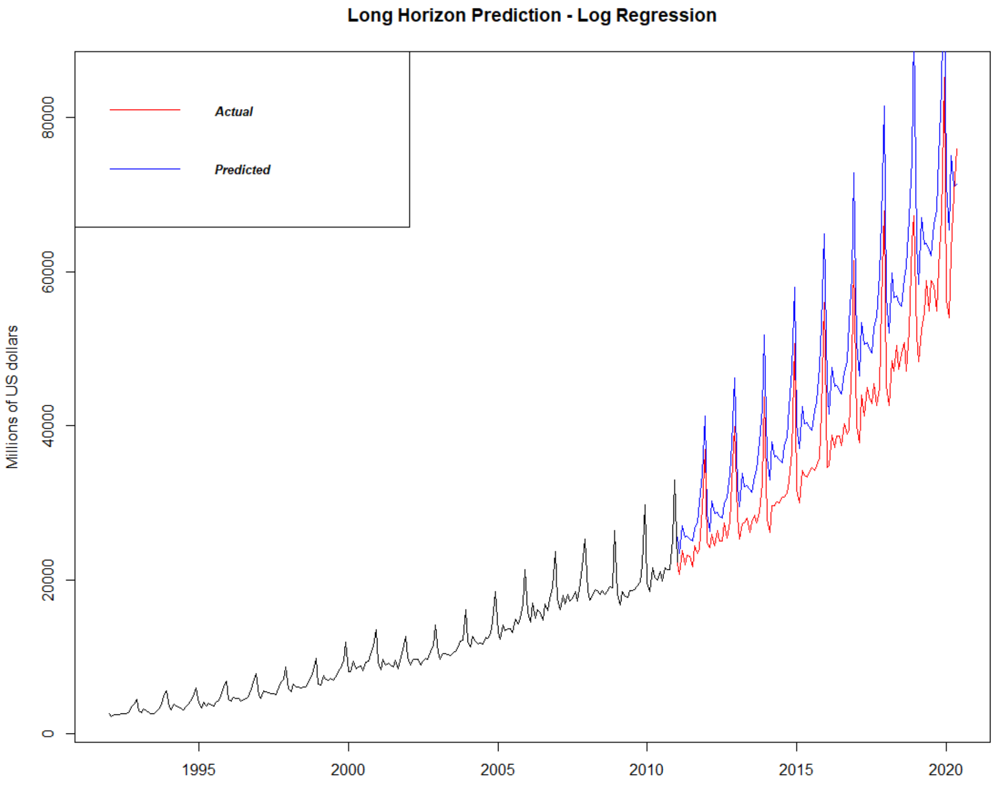
Appendix 3 – Sales trend by month over time



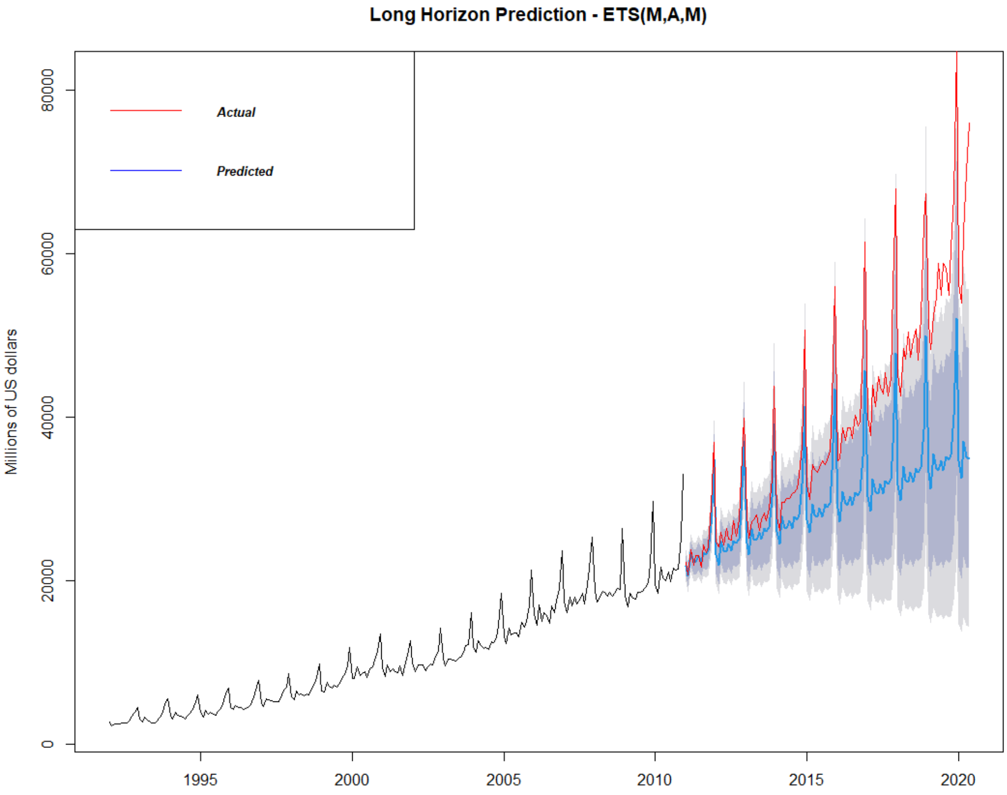
Appendix 4 – Forecast from Long-horizon Regression Model



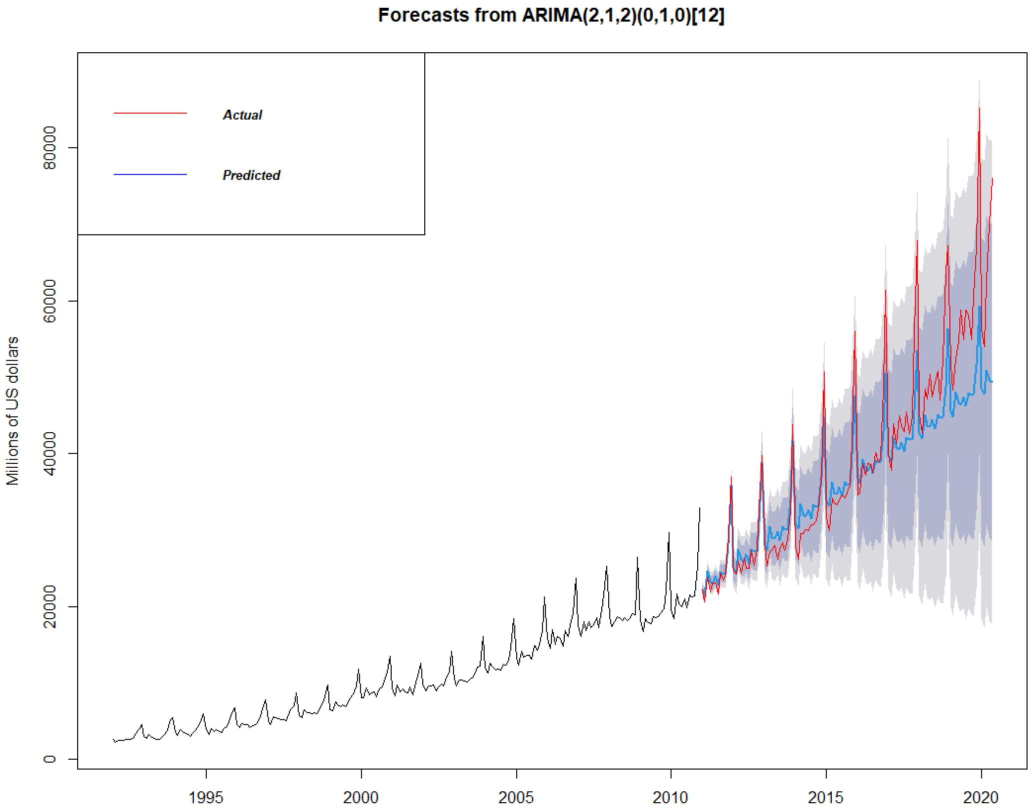
Appendix 5 – Forecast from Long-horizon Logistic Regression Model



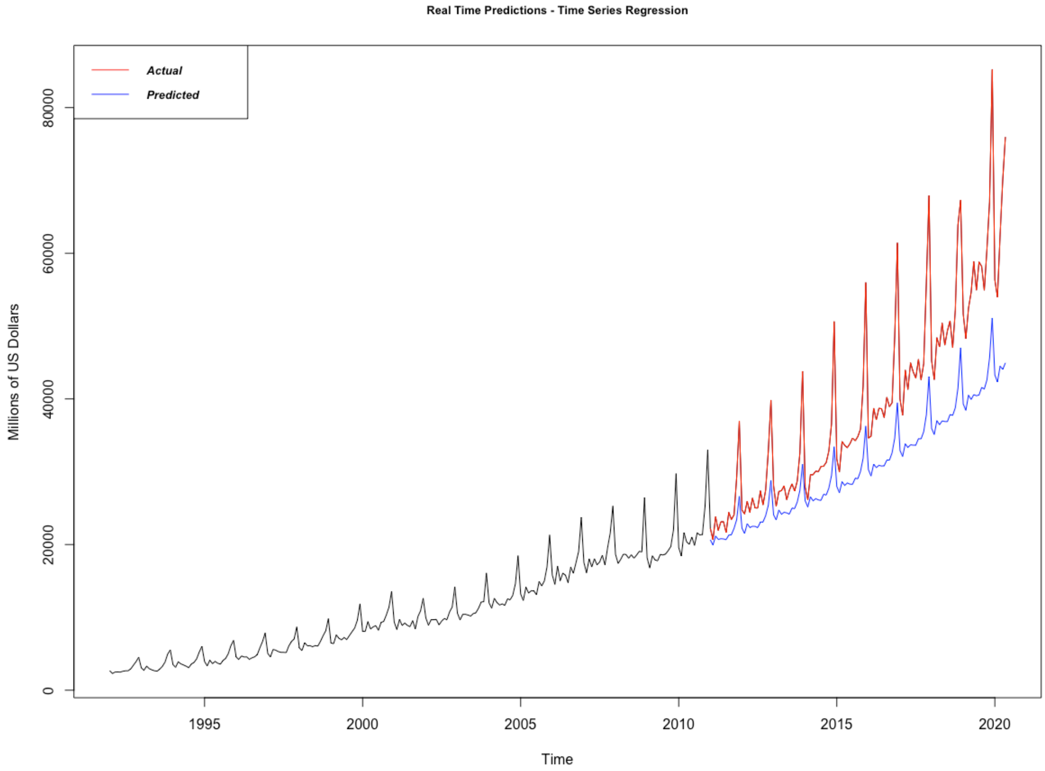
Appendix 6 – Forecast from Long-horizon Exponential Smoothing Model



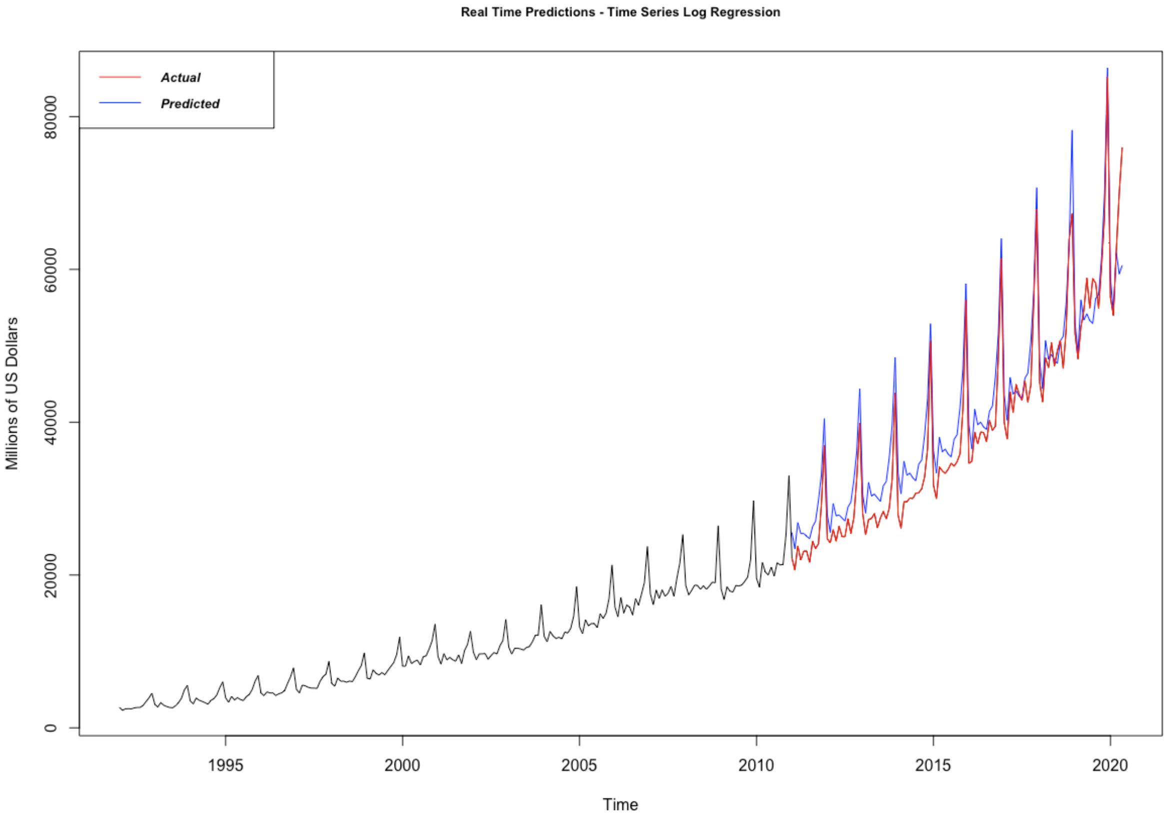
Appendix 7 – Forecast from Long-horizon ARIMA model



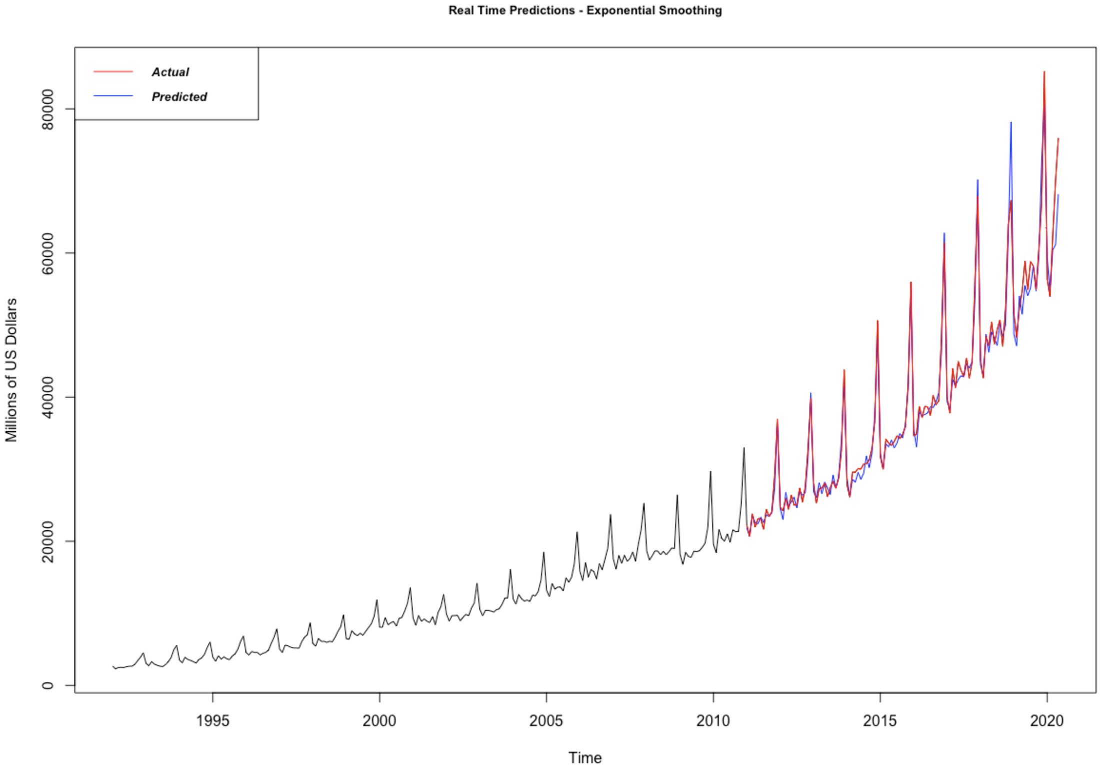
Appendix 8 – Real-Time Forecast - Time Series Regression



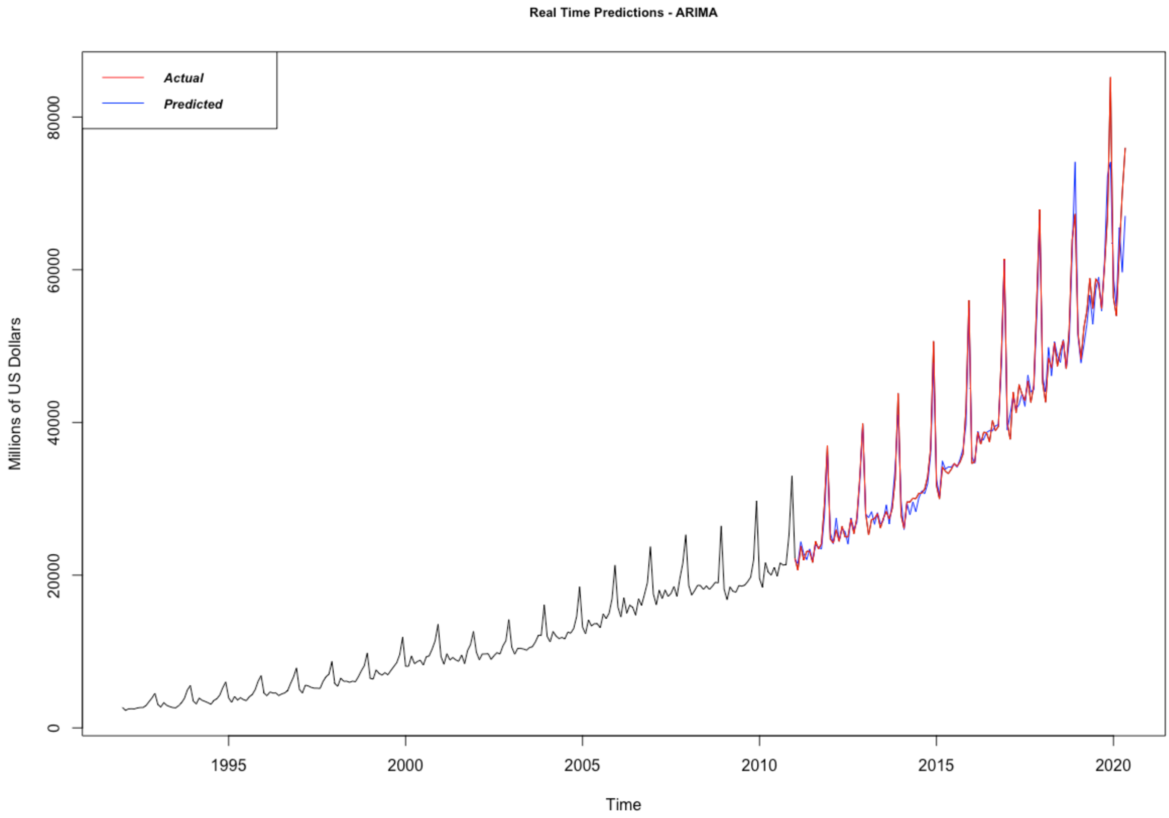
Appendix 9 – Real-Time Forecast - Time Series Log Regression



Appendix 10 – Real-Time Forecast - Exponential Smoothing



Appendix 11 – Real-Time Forecast - ARIMA



Appendix 12 – 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))
}
```