Predictive Analytics Term 4, 2020

Associate Professor Ole Maneesoonthorn

Predictive Analytics Syndicate Task #2

In this session, we will deviate from the wine quality data, and explore the time series prediction techniques.

The data file "RetailEShopMail.csv" contains time series data on retail sales of mail orders and online shopping. We will use this data set to explore the topic of time series predictions.

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.
2. Split the data into the training set: January 1992 to December 2010; and the test set: January 2011 to June 2020.
3. Construct the following time series predictive models for the training set:
   a. Time series regression with trend and seasonality components.
   b. Exponential smoothing using ets() function.
   c. 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.

4. 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.

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.
    a. Evaluate your real-time predictive models using this user specific loss.
    b. Comment on which predictive model perform best according to this loss. Is this consistent with the conclusion from the statistical metrics?

Produce a 3-page report that summarizes your analysis.

This task is due at 6pm on Saturday 3rd October 2020.