DATA2001 Assignment 3 (Weight: 25%)

Due: 7 November 2022

The aim of this assignment is to gain practical experience in analysing timeseries data and using it for forecasting. You must complete this task in Python using a Jupyter notebook. You will need to submit a single Jupyter notebook (.ipynb file) via Blackboard.

Dataset:

A major task for stock market quant analysts is forecasting intraday stock trading volumes (i.e. the amount of each stock that is traded in any given period of time). By exploring ways to predict volume, quant analysts seek to improve the performance of their trading algorithms, because they tend to depend upon the volume of trade while a stock order is active. Traditionally, traders used historical averages to predict stock volumes, but increasingly, better predictions of volumes helps to improve the performance of automated trading algorithms.

In this assignment, you will examine US stock trading data for Amazon (AMZN), and use it to make predictions of future trade volumes, using timeseries forecasting techniques. The data contains the following series (with volume in bold):

Series name	**Description**
Date	Date of trades
Open	Opening price, price of first trade of the day
High	Highest price of all trades of the day
Low	Lowest price of all trades of the day
Close	Closing price, price of last trade before the end of day
Volume	**Total number of stocks traded during the day**
OpenInt	Open interest, the total number of outstanding derivative contracts, such as options or futures, that have not been settled for an asset at the end of day.

For your interest, the data is taken from https://www.kaggle.com/datasets/borismarjanovic/price-volume-data-for-all-us-stocks-etfs

Task description:

The submitted notebook should address the following 6 tasks (see marking grid for mark allocation):

1. Data Preparation: Read the dataset using the "pandas" library and set up the index in an appropriate way for timeseries analysis. Can you identify any useful side data or exogenous variables? If so, include them into your dataframe and handle/merge them in an appropriate way. Explain how you did it along with a justification for your choices.

2. Exploratory Data Analysis (EDA): Visualise the entire data set, and comment on the patterns you can observe with respect to the features discussed in the lectures. Include visualisations appropriate for uncertainty and correlation where appropriate.

3. Focus now on the AMZN stock volume time series.
    a. Split the data into **training** and **testing** series, selecting the testing series to be the last three months of the data.
    b. Manually step through the STR decomposition process on the training data, as described in the course material. Visualise and interpret each of the components of the STR decomposition for volume. (*Hint:* You may wish to validate the output of your manual process against an automated modelling approach.)

4. Timeseries models:
    a. Fit an ARIMA model for the trend-cycle component of your STR decomposition of the training data and interpret the estimated model parameters.
    b. Using the STR components that you estimated in tasks 3 and 4, produce forecasts of AMZN stock volumes for the test data series. Include the uncertainty in the forecasts and visualise the predictions.

5. Pure forecasters - now consider your choice of ML techniques:
    a. Select an appropriate pure forecasting method to predict the trend component of the volume training data.
    b. Using the seasonal component that you estimated in task 3 and the pure forecaster from 5.a, produce forecasts of AMZN stock volumes for the test data series. Include the uncertainty in the forecasts, and visualise the predictions.

6. Evaluate the forecast performance of your model-based and pure forecasters using the test data and compare the two forecasters. Use appropriate evaluation metrics and methods. Discuss the similarities and difference between their performance and suggest possible avenues for improvement.