Coursework - MAT005 2022/23

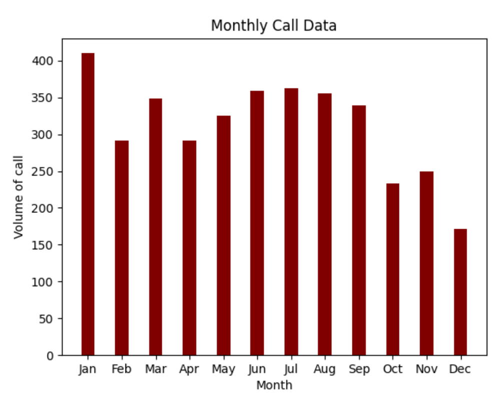
Priyanka Magar
22074213

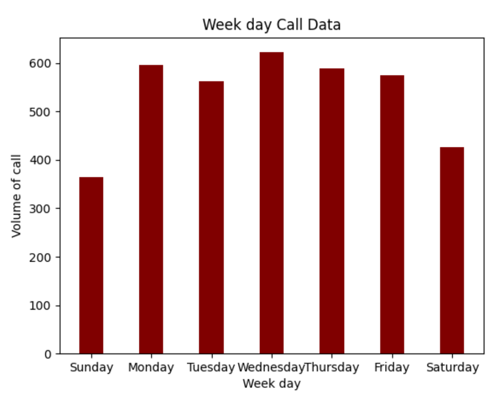
Introduction

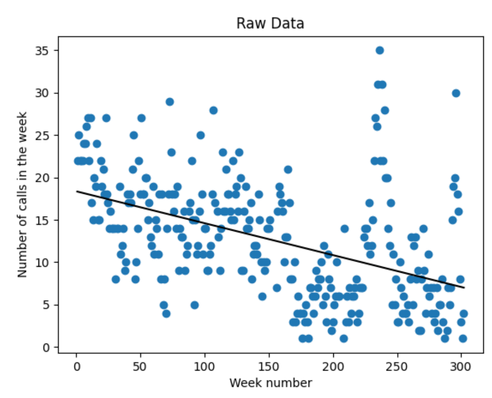
The data provided contains the date and the time an incoming call was received. The aim of this forecasting is to use multiple methods to forecast the number of calls that would be received from 24/10/22 to 18/12/22. To create the forecasts the data was initially processed and then fed through the different models. The models were then evaluated and a final conclusion drawn out.

Pre-processing

The data provided is univariant and time series regression. The dataset was checked to find any missing data. The data contained some missing data which could significantly affect the results of the forecasts. The mean number of call observed for the whole dataset was used to fill in those missing values. The aim of the forecasting is to estimate the number of weekly calls that will be received. The data provided needed to be transformed from the format provided. The volume of call received is then visualised in a monthly, weekly and week day format to make extract any obvious patterns. The raw data is also plotted in a scatter plot to extract any trend or seasonality that might be present. Using a regression line the negative trend becomes noticeable.







Methodology

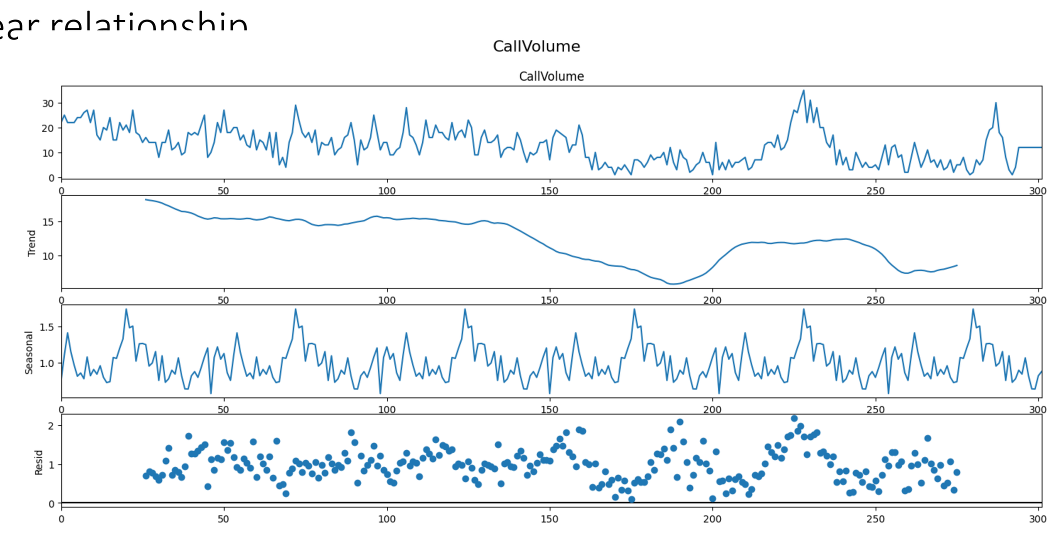
The type and quality of data can be a deciding factor when choosing which forecasting method to use. The three main models chosen to create the forecasts are simple regression, Holt Winters Additive model and ARIMA.

Simple Regression

The data is univariant meaning that simple regression is one method that can be used. The variable Y is the number of incoming calls and the X variable is the time. Simple regression checks for a linear relationship between the variables X and Y. The correlation coefficient must be used before to determine weather the two variables have a very strong linear association. The correlation coefficient calculated was -0.475(2.sf.), with p-value of 2.16e-18(2.sf) indicating that there is a moderate negative correlation between the two variables that is statistically significant.

Ordinary least squares (OLS) was used to evaluate and make a prediction. The model produced a R^2 of 0.225 indicating that only 22.5% of the variation is explained by the independent variable. The model has conducted that for each unit of increase of the independent variable the weekly call volume decreases by -0.0373. The p-value for the p test and the t test are low showing that there is a statistical significance of the model, however due to the low correlation the linear regression may not be the best model to use as the two variables could have a non linear relationship.

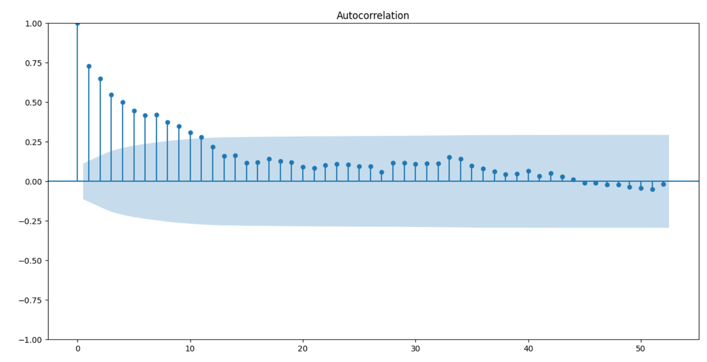
Holt Winters Additive Model

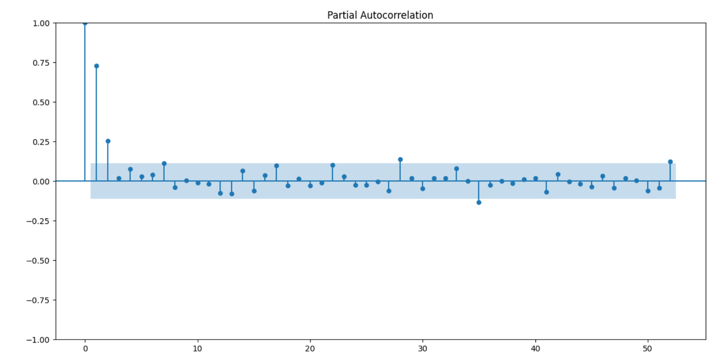


Simple regression assumes linear relationship of the variables, however it does not consider the impact of trends and seasonality. The data needed to be decomposed to withdraw those features of the data. The decomposed plot showed that the seasonal fluctuation does not vary with the level of the series therefore an additive model must be used. The independent variable are observed in a weekly time unit therefore the observations seems to have seasonal cycle of 52 weeks. The model had a SSE of 5031.077 and AIC of 961.525.

ARIMA

It was previously determined that the data has both trend and seasonality. To verify the previous findings the ACF and PACF was plotted.





The box pierce statistics was performed and the p-value associated with each lag is very small indicating that we could reject the null hypothesis that there is white noise. This also means that there is significant autocorrelation. The Box-pierce statistic shows that there is a rapid increase in the first few lags then plateaux. An Augmented Dickey-Fuller (ADF) test was performed to determine whether the data was stationary. The data was stationary with 99% confidence interval, this means that the data does not have to be further differentiated.

Results

Simple Regression forecasts:

Week Num	Forecast
303	7
304	6
305	6
306	6
307	6
308	6
309	6
310	6

Holt Winters Additive forecasts:

Week Num	Forecast
303	11
304	13
305	16
306	15
307	9
308	14
309	16
310	13

ARIMA forecasts:

Week Num	Forecast
303	11
304	10
305	10
306	10
307	10
308	10
309	10
310	10

Conclusion

The results produced by the simple regression is not the most reliable as there was not a significant linear relationship between the two variables. The low value of R^2 value shows that there are a lot of residuals that are not explained by the independent variable. This could mean that the relationship between the two variables might not be linear or there might be external variables that could be effecting the results. There are other variables that might affect the forecasting methods. An example of this is national holidays, the pandemic and other cofounding and lurking variables that could be putting people in high risk situations increasing the number of calls made during that time.
The results produced by the Holt winters additive model and the ARIMA models are much closer in range meaning that they are more reliable. To further solidify the forecasting values the mean value for each week from the two forecasts can be calculated.

Final Forecasts:

24/10/22 - 11
31/10/22 - 12
7/11/22 - 13
14/11/22 - 13
21/11/22 - 10
28/11/22 - 12
5/12/22 - 13
12/12/22 - 12

Future work

In the future the forecasts can be improved by in cooperating several techniques. The model is a direct reflection of the quality of data therefore a more reliable method should be used when filling in the missing data such as ARIMA forecasting. The data available was also very limiting as the only independent variable consisted of the time variable. Given the extra resources a research regarding which variables could be good predictors for indicating the number of calls that would be made and included in producing the forecast. Each models contain their own set of hyperparameters and given the extra time the models could be tuned to produce optimum forecasting. Regularly monitoring the accuracy and evaluating it would each change would be the next step in improving the quality of the forecasts.

References

Brownlee, J. (2020) What is time series forecasting?, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/time-series-forecasting/ (Accessed: April 10, 2023).
Brownlee, J. (2020) How to create an Arima model for time series forecasting in Python, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/ (Accessed: April 15, 2023).
Hyndman, RJ & Athanasopoulos, G 2018, Forecasting: Principles and Practice. 2nd edn, OTexts. <https://otexts.org/fpp2/>
Monigatti, L. (2022) Interpreting ACF and PACF plots for time series forecasting, Medium. Towards Data Science. Available at: https://towardsdatascience.com/interpreting-acf-and-pacf-plots-for-time-series-forecasting-af0d6db4061c (Accessed: April 15, 2023).
Prabhakaran, S. (2023) Arima model - complete guide to time series forecasting in python: ML+, Machine Learning Plus. Available at: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/ (Accessed: April 13, 2023).
Shumway, R.H. and Stoffer, D.S. (2017) Time series analysis and its applications: With R examples. Cham, Switzerland, -: Springer.