Time Series Analysis & Modeling

DATS 6450, Section 1

Instructor: Dr. Reza Jafari

Term Project

Mariko McDougall

MM

05-05-2021

Table of Contents

Abstract

Introduction

Dataset

Stationarity

Time Series Decomposition

Feature Selection

Basic Models

Holt-Winters

Multiple Linear Regression

ARMA, ARIMA, SARIMA Models

Levenberg Marquardt Algorithm

Diagnostic Analysis

Final Model selection

h-step ahead Predictions

Summary and conclusion

References

Abstract

Introduction

The dataset used for this project is the Metro Interstate Traffic Volume Data Set, detailing the hourly traffic along a the I-94 interstate highway between Minneapolis and St-Paul, in the westbound direction.

The dataset consists of the hourly traffic in number of vehicles, per the Minnesota Department of Transportation, the hourly temperature, amount of rain in mm, amount of snow in mm, a numeric percentage of cloud cover, a contextual description of the weather, in both short and detailed forms, and information if the date was a holiday.

The original dataset contained hourly traffic readings from October 2, 2012 to September 30th, 2018, encompassing 48,204 individual timepoints. However, due to a large gap in the data from August 8th, 2014 to May 11th 2015 and slightly inconsistent data for the following week We have chosen to exclude the early subset of data from the analysis. Instead, the data used for this project spans from May 24th, 2015 to September 30th, 2018, and contains 32,222 datapoints.

Dataset

Several of the dataset’s original features required refinement prior to performing time series analysis.

The original weather descriptions, while likely to be informative for multivariate modeling, were originally provided as string decriptions, which are not ideal for time series models. These were converted into one hot encoding columns by creating a new binary feature column for each unique weather type, and converting the string into a 1 in the column where it appled, and a 0 for all other weather encoding colmns.

Several of these columns were grouped together into weather categories, to provide additional information to the model, with the expectation that a consolidated column may be more informative. An example of such a column would be the ‘Precipitaition’ feature, which combined ‘Thunderstorm’, ‘Drizzle, ‘Rain’, and several other similar one hot encoding columns. The original binary columns were maintained in the dataset.

Similarly, the original ‘Holiday’ feature was a string column which explicitly listed which holiday was occurring, if any. While it is possible that the specific holiday may be relevant, it seemed unlikely that there would be a large impact given each annual holiday could occur at most four times in the dataset, given the timespan. Instead, this feature was converted into a binary holiday/not holiday classification.

Finally, a boolean weekday/not weekday column was created using the timestamps of the hourly traffic report. This feature was created, as it seemed likely that traffic patterns would be strongly influenced by the day of the week, and thus if commuters would be largely present on the road.

Stationarity

Time Series Decomposition

Feature Selection

Basic Models

Holt-Winters

Multiple Linear Regression

ARMA, ARIMA, SARIMA Models

Levenberg Marquardt Algorithm

Diagnostic Analysis

Final Model selection

h-step ahead Predictions

Summary and conclusion

References