Time Series Analysis – Final Project

Predicting Sales for Online Retailer Sponsored Campaigns

Ezra Kim

Agenda

- Problem Statement
- Assumption & Steps
- Data Overview
  - Time Series EDA
- Data Processing
  - Feature Engineering
- Proposed Approaches
  - 3 Models Chosen
- Results
- Future Work
- Important Links

Problem Statement

- This is an exploration of the media mixed model (MMM) concept
- MMM analysis is popular in marketing analytics due to the ease of explanations to stakeholders, relative simplicity of the modeling, and scalability of the modeling
- Predicting sales is the holy grail of MMM
- Every company does MMM differently, the purpose of this project was to explore different models and evaluate them
- Developing three models to forecast sales will serve as a prime example of time series analysis

Assumptions & Steps

- Data Preparation
- Exploratory Data Analysis
- Modeling
- Diagnostics/Forecasting
- Conclusion

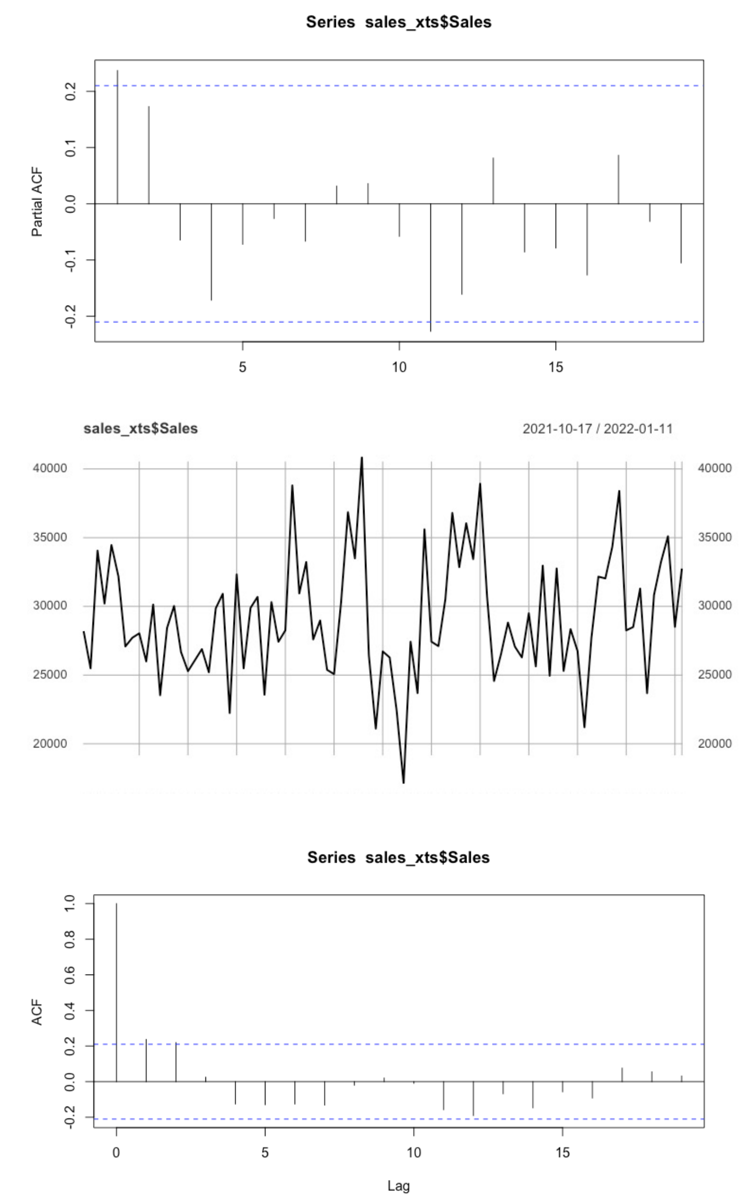
- This is an exploratory analysis so the time series being observed are Sales and Impressions
- The hypothesis is that the BSTS model will outperform the other models

Data Overview

- Data Source:
  - From Kaggle: https://www.kaggle.com/datasets/saicharansirangi/adanalyse?resource=download
- Dataset:
  - This is anonymized campaign data from various online retailer sponsored product ads, by day
- Observation Count & Time Period
  - 9586 observations
  - From 2021-10-17 to 2022-01-11
  - Train Set: 2021-10-17 to 2021-12-23
  - Test Set: 2021-12-24 to 2022-01-11

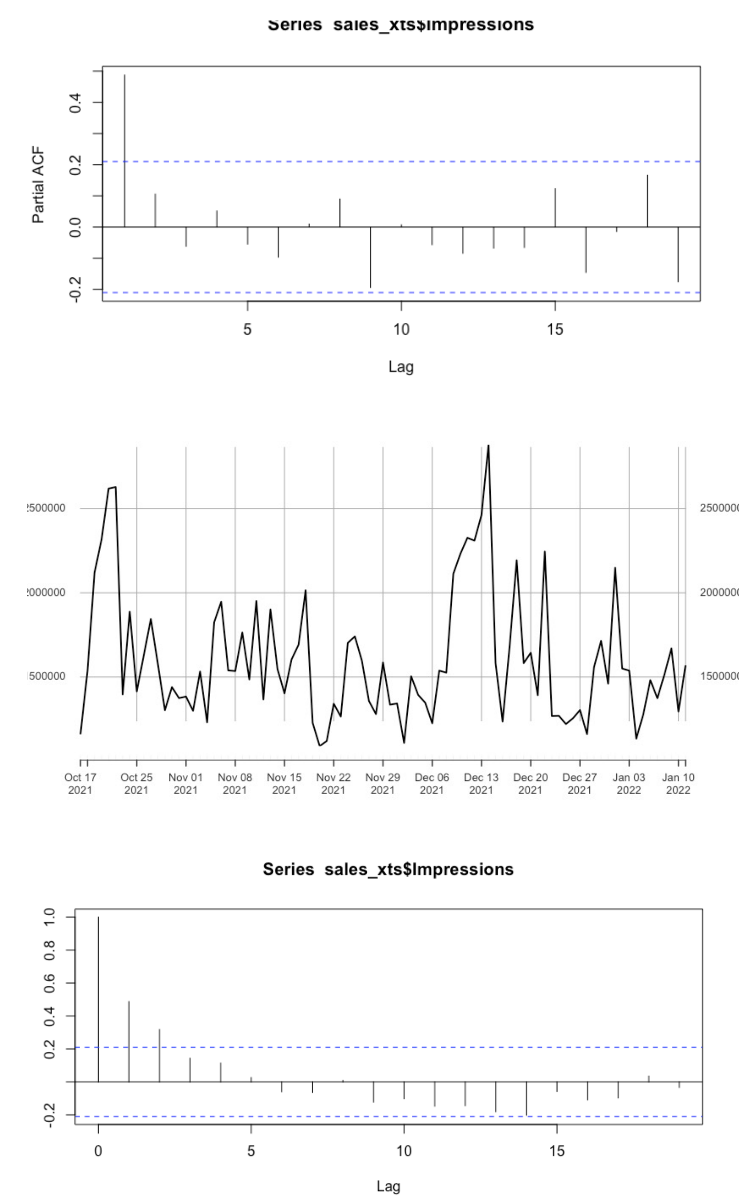
Time Series: Sales is Stationary

- Qualitative:
  - Plots show that this time series is stationary, no trend, no seasonality, and no cyclical to worry about
- Quantitively
  - The KPSS and ADF tests also support the plots that this time series is stationary
- Auto.Arima was used to determine that ARIMA(2,0,0) was the best fit.
- Due to this, there is no need to transform this variable



Time Series: Impressions

- Qualitative:
  - Plots show that this time series is stationary, no trend, no seasonality, and no cyclical to worry about
- Quantitively
  - The KPSS and ADF tests also support the plots that this time series is stationary
- Auto.Arima was used to determine that ARIMA(1,0,0) was the best fit.
- Due to this, there is no need to transform this variable
- This time series is meant to be an external regressors to Sales



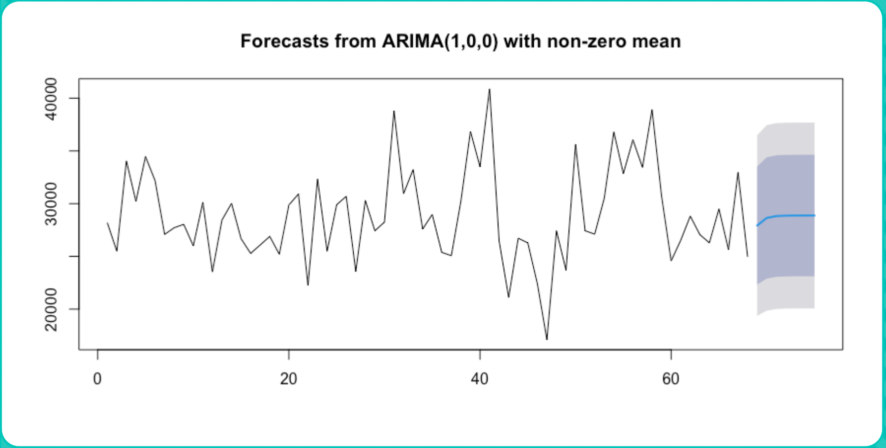
Data Processing

- Both Time Series are stationary, no seasonal, no cyclical, no trend
  - Perfect to use in modeling
- Data processing:
  - Aggregate by day since the data is split by some categorical variables
  - 87 total days
  - Train Dataset: 68 days
  - Test Dataset: 19 days
- No need for feature engineering aside from creating xts

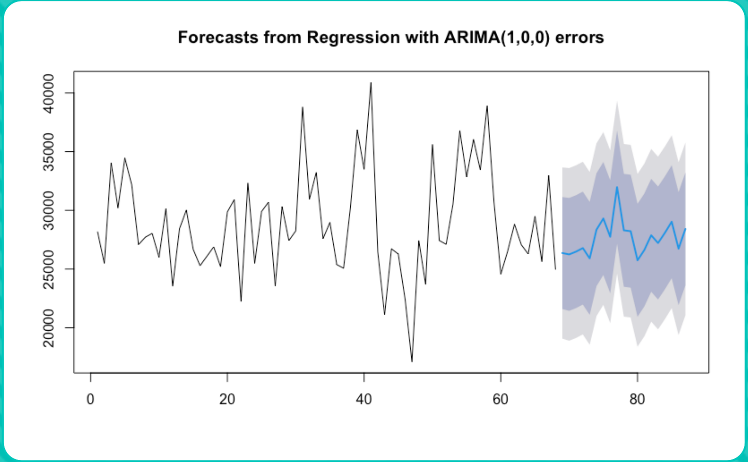
Three Models Chosen

	Auto.Arima	Arima with external regressors (Dynamic Regression)	BSTS with Dynamic Regression state specification
Model Summary	Auto.Arima is a function in the forecast package in R that automatically fits an ARIMA model to a time series by selecting the best model parameters (p, d, q) that minimize a chosen information criterion, such as AIC or BIC.	Variant of the ARIMA model where external or exogenous variables are incorporated into the model, effectively combining regression and time series techniques. Can handle data that might be influenced by factors outside the time series itself.	Incorporates uncertainty into time series forecasts. The model allows for the inclusion of external regressors through a state specification. It is very flexible and allows the data to inform the model structure.
Assumptions	• Assumes stationarity • Assumes residuals are normally distributed and independent	• Assumes relationship between external regressor and response variable is linear • Assumes stationarity • Assumes residuals are normally distributed and independent	• Has fewer assumptions than Box Jenkins models • Assumes data is generated from combination of various components and priors are specified for each component
Utility	• The key benefit is the automation of the process of identifying the order of an ARIMA model, which can be a complex and time-consuming process if done manually. • It can handle both seasonal and non-seasonal data.	• This model is useful when there are external factors may influence the time series • Incorporate these variables directly into the model and estimate their effects.	• BSTS models are very flexible and can handle complex time series structures. • Allows to incorporate uncertainty into the forecasts • Can include both known and unknown external regressors.
rade-Offs	• Does not always select the best model • May not handle irregularly spaced time series data	• Can lead to overfitting if too many unnecessary external regressors are included • Requires future values of external regressors, which may not exist	• Computationally expensive • Interpreting results requires Bayesian familiarity • Need to specify priors

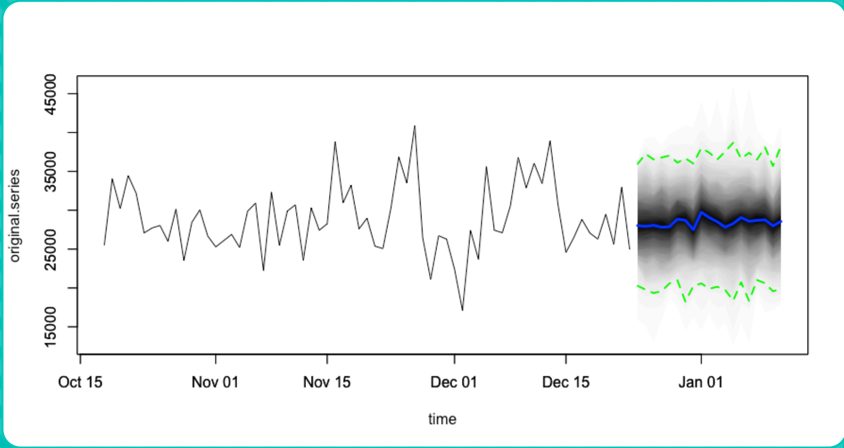
Auto.ARIMA



Dynamic Regression



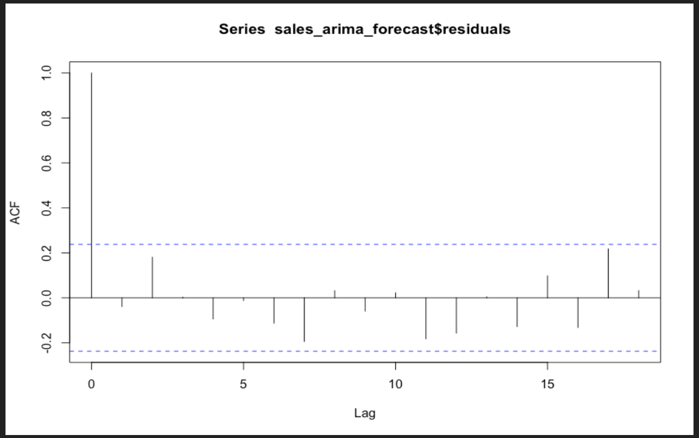
BSTS

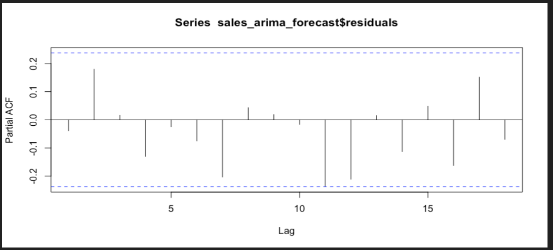


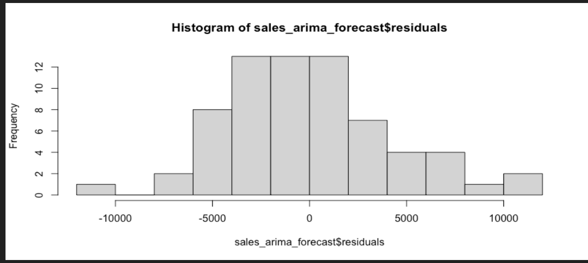
All Model Forecasts

All Models Residuals are I.I.D

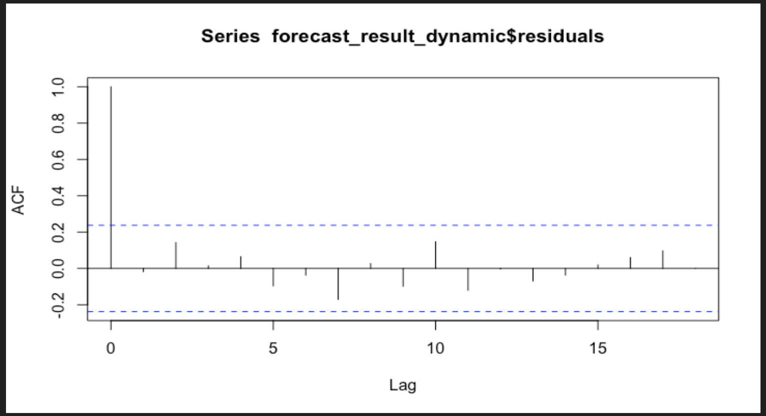
Auto.ARIMA Residual Analysis

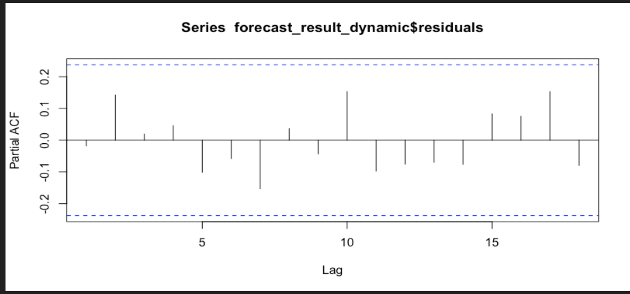


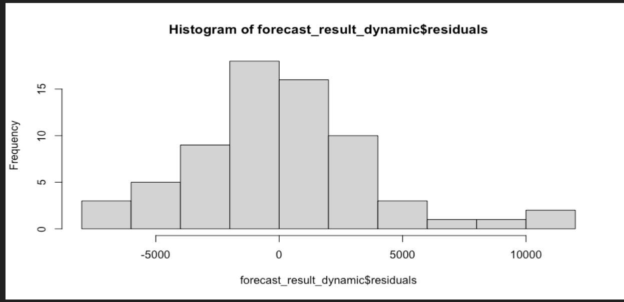




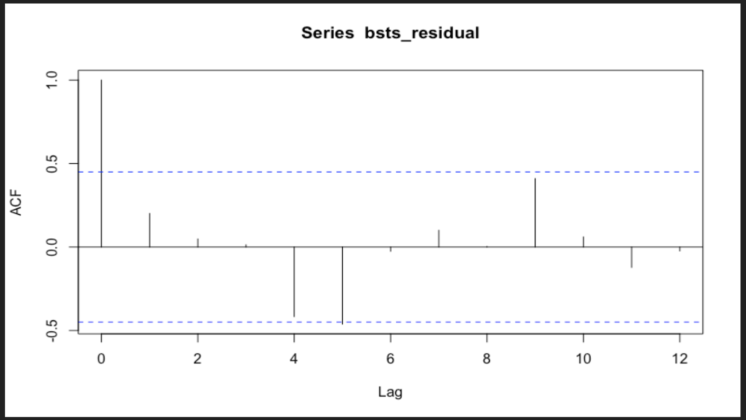
Dynamic Regression Residual Analysis

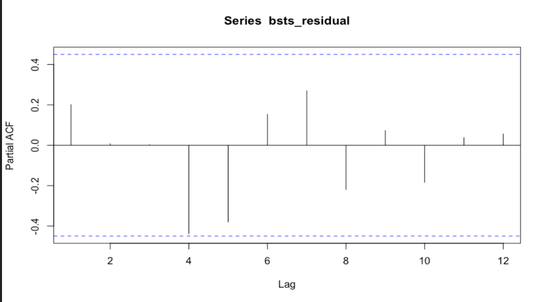


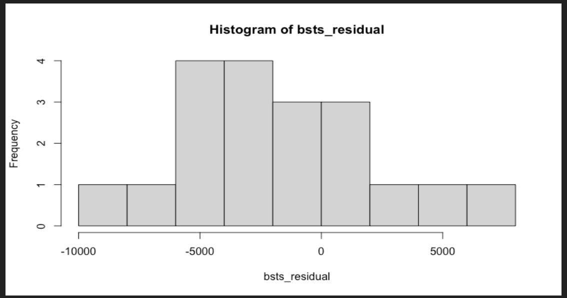




BSTS Residual Analysis







Model Performance

- Auto.Arima
  - RMSE: 4297.545
  - MAPE: 11.697%
- Dynamic Regression
  - RMSE: 4275.661
  - MAPE: 11.384%
- BSTS:
  - RMSE: 4220.212
  - MAPE: 11.469%

It is surprising to see that Dynamic Regression and BSTS performed similarly.

It would be interesting to see how robust each model could be with more regressors and feature engineering

Future Work

- Nowcasting
  - Adjusting these models to accurately forecast the short-term
- More State Specifications for BSTS
  - Black Friday, Christmas, holidays in general play a big role in retail during this time
  - Spike Slab Priors to help regularize the coefficients
- Continue to develop both ARIMA and BSTS to see which highly optimized models can outperform each other
  - Adding more regressors would help create a more complex model
- Do more feature engineering to see if there can be impressions by categorical data like retailer, campaign, ad group
  - With these new variables we can see the effect of impressions (or any media metric) from campaigns, retail stores, advertised products on sales
- Create hierarchical BSTS time series model
  - Modeling on campaign performance data would mean creating a hierarchical model to help accurately model everything

Thank you!

Important Links

- Dataset: https://www.kaggle.com/datasets/saicharansirangi/adanalyse?resource=download
- Github: https://github.com/bigtreesfallhard/Time-Series-Final