MSDS 413 Midterm Project
John Moderwell

Problem:

In this competition we were given a challenging time-series dataset consisting of daily sales data, provided by one of the largest Russian software firms - 1C Company.
We were asked to predict total sales for every product and store in the next month.
The evaluation metric was RMSE where True target values are clipped into [0,20] range.

Significance:

This problem is significant because of the real business implications associated with it. If a company is able to develop an accurate predictive model for future sales, they can allocate their resources much more effectively, meet demand more efficiently and gain a competitive advantage in their market.

Data:

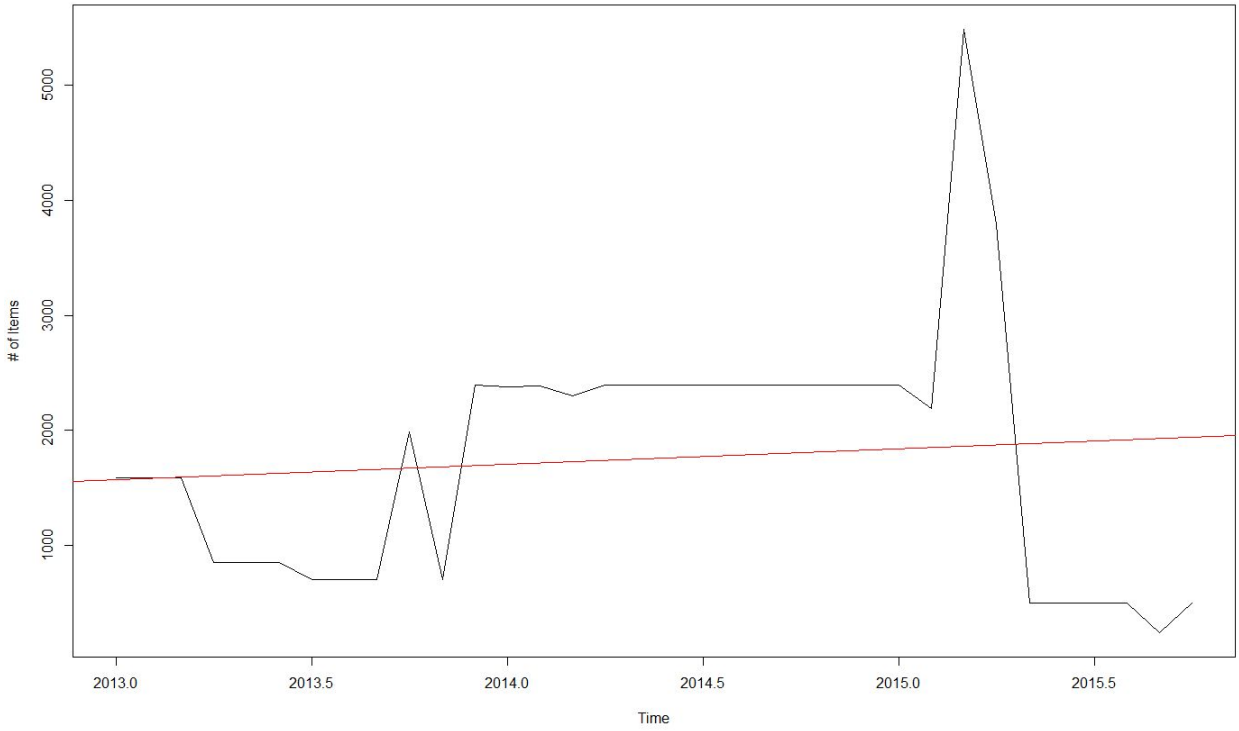
Before any EDA, data preparation is key. I checked for any missing data and there does not appear to be any.

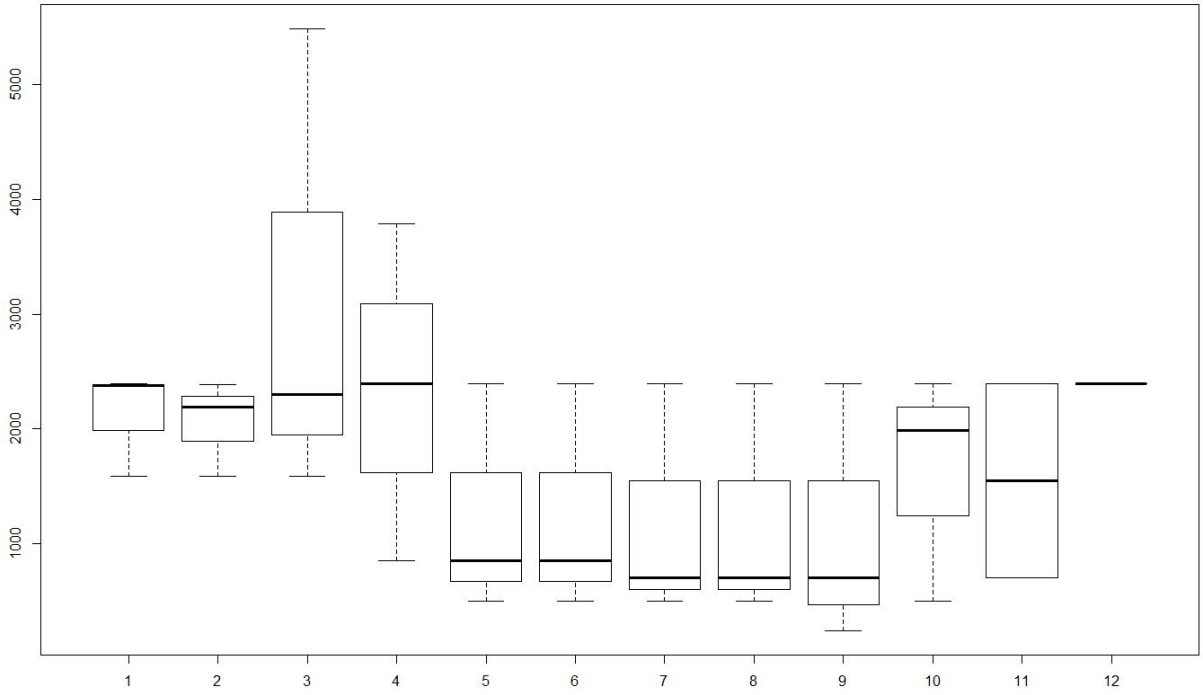
```
item_category_id            item_id                date
               0                  0                   0
  date_block_num            shop_id          item_price
               0                  0                   0
    item_cnt_day          item_name  item_category_name
               0                  0                   0
```

After this I started merging the training set with the other supplemental data sets included: “Items” and “Item_categories”. Then I made sure that the date in my training set was changed from a character vector to a numeric vector.

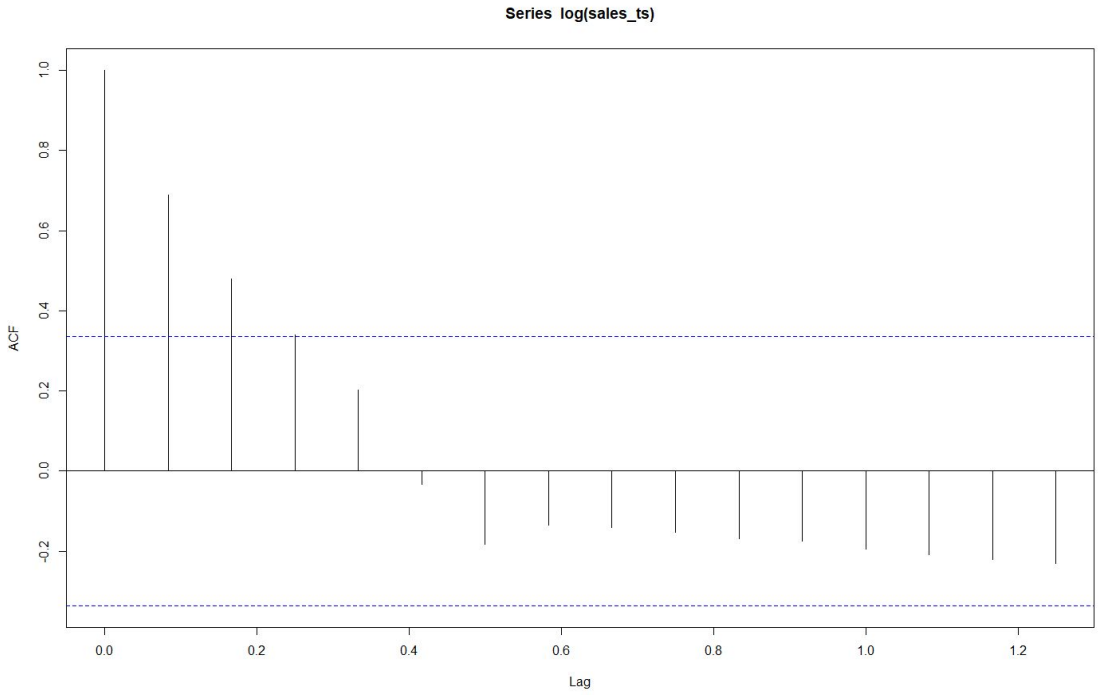
Since we are projecting sales for the next month: November 2015, we want to compute the total sum of the sales for each month. I did this by first grouping “sales_data” by “date”, ”item_category_id”, “item_id”, “shop_id”, “item_cnt_day”. Then I took the sum of “item_price” and called this new feature “total”. With this new feature, I wanted to convert the data to a time series. I took the “total” feature and created a time series

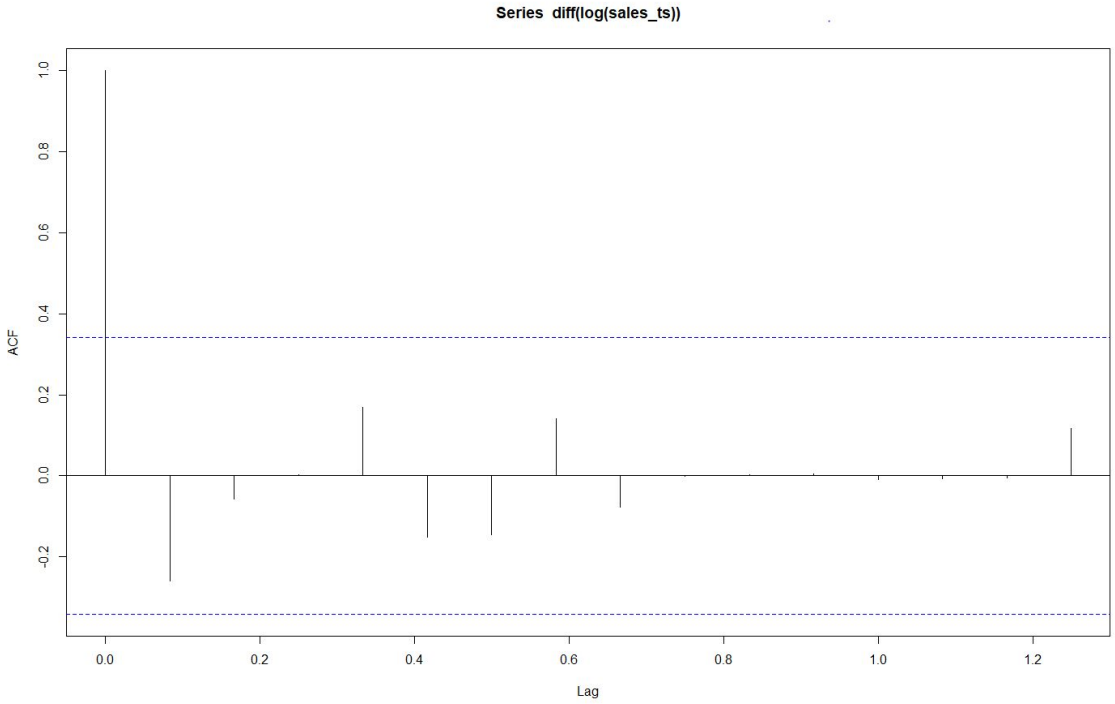
starting in Jan 2013 and ending in Oct 2015 with frequency = 12 (for monthly data). I plotted this data with a fitted regression line and then also made a boxplot





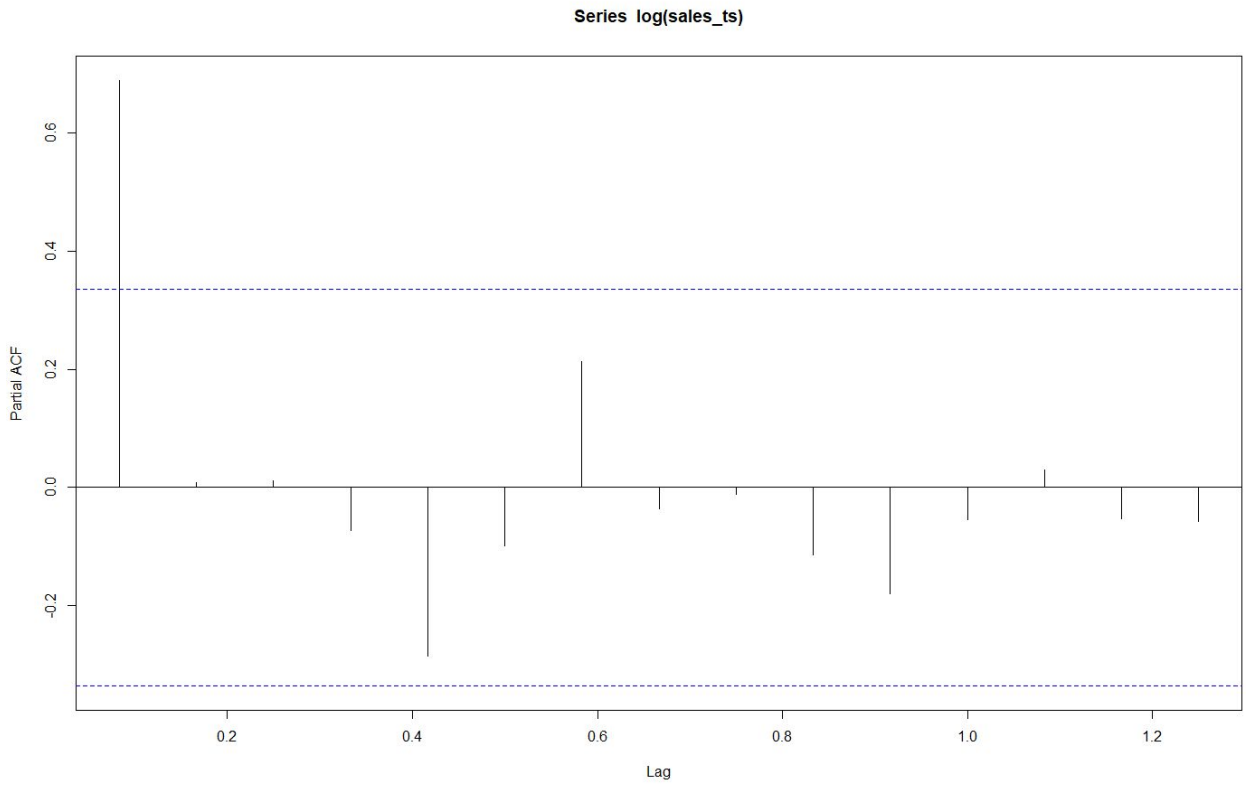
Since there is a seasonal component to the data, patterns in the series can be examined via correlograms. The correlogram (autocorrelogram) displays graphically and numerically the autocorrelation function (*ACF*), that is, serial correlation coefficients (and their standard errors) for consecutive lags in a specified range of lags. We will compute the ACF on data with log transformation and data with log transformation and lagged differences subtracted from the data.





Another useful method to examine serial dependencies is to examine the partial autocorrelation function (*PACF*) - an extension of autocorrelation, where the dependence on the intermediate elements (those *within* the lag) is removed. In other words the partial autocorrelation is similar to autocorrelation, except that when calculating it, the (auto) correlations with all the elements within the lag are partialled

out. Here is the partial ACF for the log transformed data.

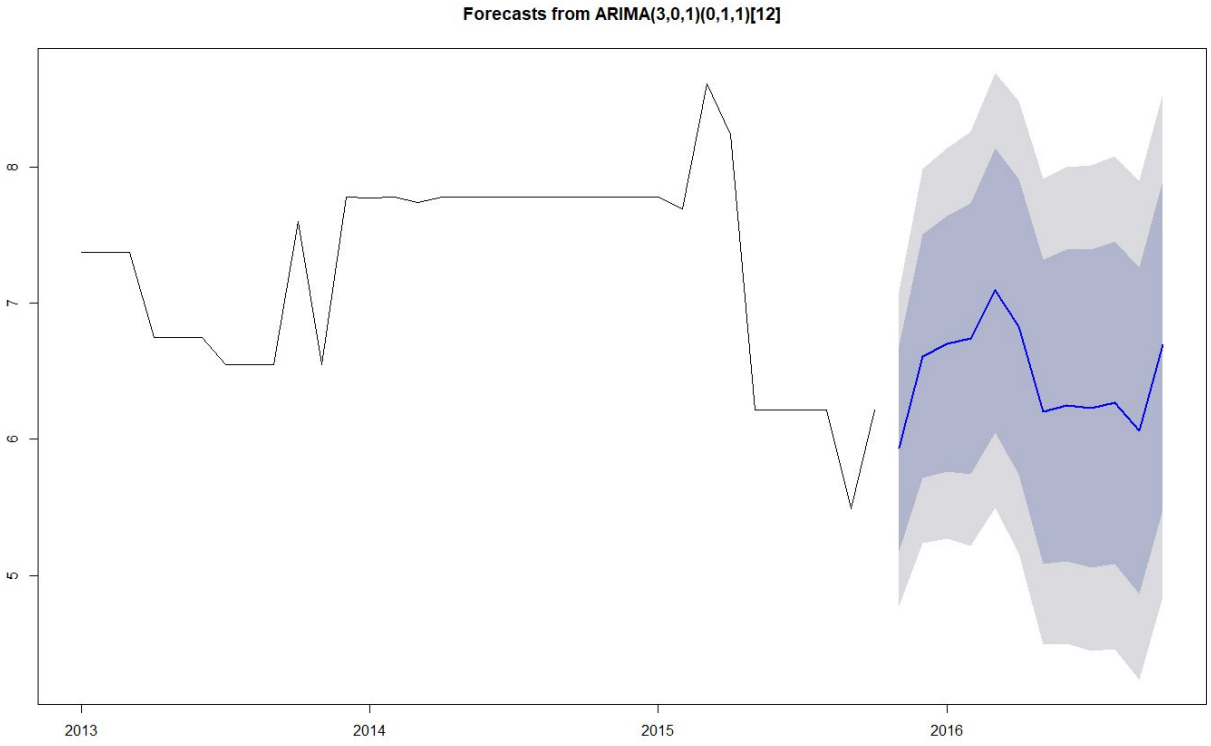


Models Built:

Type of Models: The type of models I built were ARIMA (3,0,1)(0,1,1), ARIMA (1,0,0) with non-zero mean using auto.arima() function, ETS (M,N,N) and a Holt Winters' model. I chose these models because these are the models used in the class and we are working with time series data. I tested these models on two sets of data. One set of data was log transformed and the other set was seasonally adjusted. Seasonally adjusted data (also called deseasonalizing) shows insight about the seasonal pattern in the time series and helps to model the data without the seasonal effects. In my R code, fit and fit4 are the same model just tested on the different datasets. This is the same for fit1 and fit5, fit2 and fit6 and fit3 and fit7. Since fit, fit1, fit2 and fit3 are tested on log transformed data they will get different accuracy measurements so each model should only be tested against those in the same set. Below are plots of the 12 month forecasts predicted using each model and used on the log transformed data. Following the plots

are accuracy metrics for each model. The metrics computed are ME (mean error), RMSE (root mean squared error), MAE (mean absolute error), MPE (mean percentage error), MAPE(mean absolute percentage error) and ACF (Autocovariance). The competition calls for RMSE to be the main metric used so this will be kept in mind.

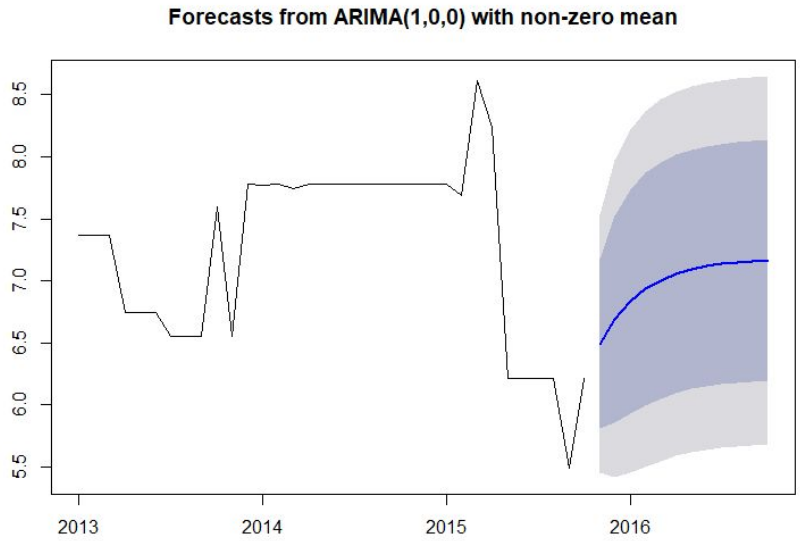
ARIMA Model:



	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.02541698	0.3908773	0.2388269	-0.6890161	3.376306	0.2510798	0.01544

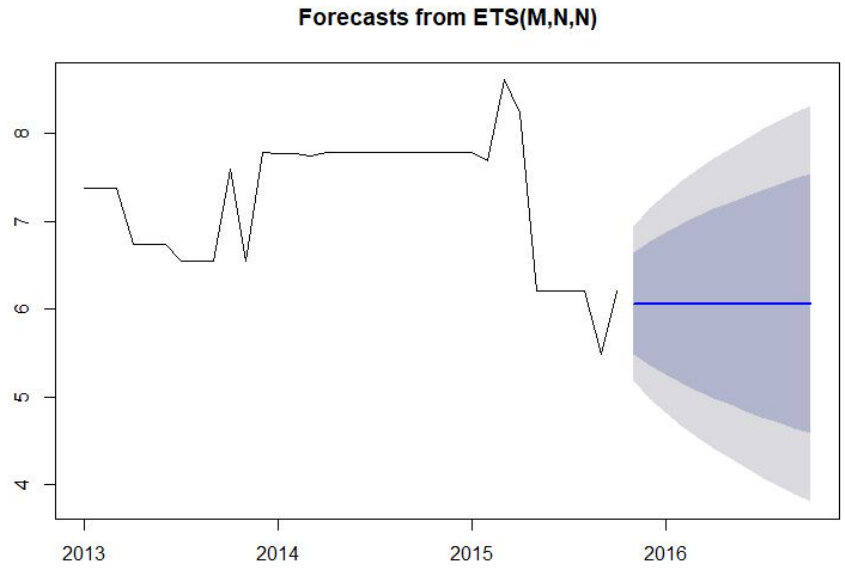
MSDS 413 Midterm Project
John Moderwell

Auto ARIMA Model:



	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.005626775	0.5157667	0.3476581	-0.6412897	5.030376	0.3654944	-0.04817536

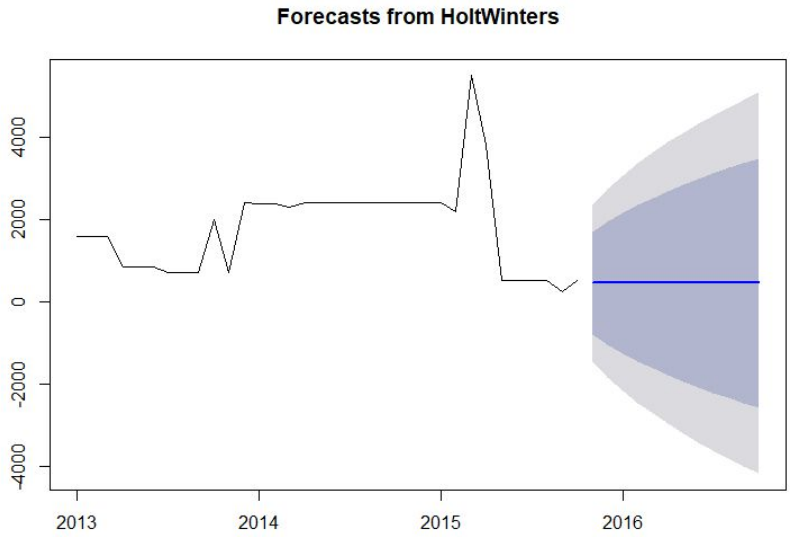
ETS Model:



	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.05185251	0.5283244	0.2854301	-1.153312	4.210169	0.3000739	-0.01250908

MSDS 413 Midterm Project
John Moderwell

Holt Winters Model:



	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-51.50989	953.2086	463.7773	-37.04503	51.86767	0.3347934	0.08217695

Performance / Accuracy: How did your models perform?

By far the best model was the ARIMA model achieving an RMSE score of .391 for its 12 month forecasts. In second was the auto.arima () model with an RMSE of .52, third was the ETS model (RMSE = .53). Last place was the Holt Winters' model. This makes sense because it is a much more simple model.

Literature: FIVE examples from PEER REVIEWED JOURNALS of how the types of models you selected were used in similar situations.

A popular and widely used statistical method for time series forecasting is the ARIMA model. ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a class of model that captures a suite of different standard temporal structures in time series data. The ARIMA model is commonly used to forecast demand, sales or other time series data (stocks etc). In this competition, the goal was to produce forecasts for the total monthly sales of a Russian software company. However, there

are other similar applications. Here is a list of of ARIMA models being used in real-world applications.

- According to authors Patricia Ramos, Nicolau Santos, and Rui Rebelo.at the International Journal of Engineering Business Management, ARIMA models can be used to predict demand for food companies. The results obtained prove that ARIMA models could be utilized to model and forecast the future demand for food manufacturing. These results were provided to managers of the company (Ramos).
- In another example, “ProfARIMA” is a new procedure that selects the lags of a Seasonal ARIMA model according to the profit of a model's forecasts by taking advantage of search heuristics. According to the authors, this procedure is tested on both publicly available datasets and a real-life application with datasets of The Coca-Cola Company in order to assess its performance, both in profit and accuracy (Van Calster).
- Another example was a study done on a South African car company. The goal was to build a model to predict short term seasonal sales. The study included Holt Winters’, exponential smoothing and ARIMA models. The study concluded that ARIMA models produced the most accuracy (Katleho).
- Researchers created a model combining ARIMA with a back-propagation neural network. They concluded that a single ARIMA has limitations to its level of predictive accuracy. Also, it can only deal with small prediction periods in the forecasting process. However, predictive accuracy can be greatly improved by combining it with a neural network. The model they proposed was intended to be used in future prediction researches and industrial data analysis (Ji).
- The idea of a “hybrid model” - combining ARIMA with a neural network has gained lots of support. In another example, it was also used to better predict article popularity for businesses looking to improve their position in online search results (Omar).

MSDS 413 Midterm Project
John Moderwell

Limitations: What are the limitations of the model you designed?

The limitations of this model is that it was not compared to to other methods such as the use of machine learning, neural networks and other ensemble methods. This makes it hard to know if an ARIMA model is the best model for the company and if the results from the metrics used for evaluation actually mean it is an accurate model.

Future Work: What would you do in the future to improve your models?

In think that in the future I would compare these scores to regression models,other machine learning models and models built by ensemble learning methods such as random forests. I would also like to prepare the data in different ways. Because there are 4 different data sets that could be combined, there are many ways to create features and visualize different parts of the data. I think this would give a more complete picture of the reality of this software company and improve the predictive accuracy of the models built.

Learning: What did you learn from building these models?

The main thing that I learned from building these models is that feature creation and defining the data that is to test and train on is a very important step in the model building/validation processes. It is very rare that there is enough features in the data to build a model right away. Before any EDA can happen, data preparation needs to be done. This includes cleaning the data and making sure that there are the necessary features to be used by the models.

Works Cited

1. Katleho Daniel Makatjane, and Ntebogang Dinah Moroke. "Comparative Study of Holt-winters Triples Exponential Smoothing and Seasonal Arima: Forecasting Short Term Seasonal Car Sales in South Africa." *Risk Governance & Control: Financial Markets & Institutions* 6.1 (2016): 71-82. Web.
2. Ji, Shenjia, Hongyan Yu, Yinan Guo, and Zongrun Zhang. "Research on Sales Forecasting Based on ARIMA and BP Neural Network Combined Model." *Proceedings of the 2016 International Conference on Intelligent Information Processing*(2016): 1-6. Web.
3. Omar, Hani, Van Hoang, and Duen-Ren Liu. "A Hybrid Neural Network Model for Sales Forecasting Based on ARIMA and Search Popularity of Article Titles." *Computational Intelligence and Neuroscience : CIN* 2016 (2016): Computational Intelligence and Neuroscience : CIN, 2016, Vol.2016. Web.
4. Ramos, Patricia, Nicolau Santos, and Rui Rebelo. "Performance of State Space and ARIMA Models for Consumer Retail Sales Forecasting." *Robotics and Computer Integrated Manufacturing* 34 (2015): 151. Web.
5. Van Calster, Baesens, and Lemahieu. "ProfARIMA: A Profit-driven Order Identification Algorithm for ARIMA Models in Sales Forecasting." *Applied Soft Computing Journal* 60.C (2017): 775-85. Web.

MSDS 413 Midterm Project
John Moderwell