**Projecting Retail Sales with SARIMA Modeling Techniques**

**David Culhane**

**DSC 630-T301**

**Prof. Andrew Hua**

**Introduction**

**The modern retail sales environment relies heavily on data-driven approaches to make decisions for stores with respect to inventory, staffing, and more. In my time working for a retail chain, my coworkers and I were regularly able to see daily, weekly, monthly, and annual sales projections for our store as well as broken down by departments. These projections served as the underlying basis for staffing decisions to maximize profit at the store level. The quality of these projections directly affects these decisions and have ramifications for the company as a whole. If a retailer is consistently able to create high quality sales projections, they are likely to have an advantage over their competitors. This analysis uses machine learning to attempt to create models to be used on retail sales data from an unnamed set of stores broken down by departments to attempt to accomplish this feat.**

**Data Selection & Preparation**

**The data used in this exercise comes from Kaggle and its link is listed in the references** (Singh, 2017)**. The raw data came from three CSV files that contained store information, weekly external economic data for each store, and weekly sales broken down by store department. The external economic data included metrics for the consumer price index, temperature, unemployment rate, and more for each store’s area in the given week.**

**The data was merged in a manner that created a dataset where each row represented a single department’s sales data, store information, and external economic information for a single week.**

**Initial data exploration focused on the differences between store types, of which there were three – A, B, and C. Each store type typically had a different range of weekly sales and size of the store itself, so it was decided early that they should be separated from each other with models made for each store type.**

**The next step in preparation was making sure that the datasets, now three of them, were ready for SARIMA modeling. Since SARIMA models require related, sequential data, the models being created needed to be fed data from a single department – each department would require its own model in order to encapsulate their individual behaviors and performances. Since this would require additional segmentation to create models for every single department (more than 3000 combinations), it was decided that one model would be created from a department of each store type. The data from these three departments were then isolated to begin the modeling process.**

**Modeling Methods**

**SARIMA models are seasonal autoregressive integrated moving average models. The regression aspect of the model uses prior values of the model to inform what the model should do next while the moving average portion focusses on the regression’s error. The integration aspect enters when differentiating data to stabilize the data being fed into the model. For this analysis, the SARIMAX class from the statsmodels python library and auto\_arima from the pmdarima python library were used to compare the results of each model for each department** (SARIMAX, 2024 and Justin, 2022)**.**

**The data used for these models spans from February 2010 through October 2012. Given that, the decision was made to split the data for the three departments into training and testing sets with the split taking place on April 1, 2012, allowing for six months of testing data out of the 2.75 years of total data for each department.**

**The SARIMAX class model from statsmodels requires the user to manually input the parameters crucial to the success of the modeling process. This required interpreting ACF and PACF plots** (TeKnowledGeeK, 2021)**. The chosen parameters are shown below. The parameters were selected**

|  |  |
| --- | --- |
| **by seeing which value was the last value of significance before the plot appeared to stabilize. The value selected from the PACF plot informed the first value in the order parameter and the ACF plot informed the third value. The middle value was the order of differentiation, and no differentiation took place. This was therefore left to be 0. These same values were used for the seasonal order parameters, with 52 representing the number of periods in a season. This analysis looks at a whole year being the seasonal phenomena, so the 52 weeks in a year were** | A computer screen shot of a program code  Description automatically generated |

**used for that parameter. Projections were then made to compare them with the actual sales data for each department that underwent modeling. The exog parameters in the models contained the weekly economic data to add additional information to the modeling process to inform model creation and predictions.**

**pmdarima’s auto\_arima function takes away the slightly arbitrary nature of ACF and PACF plot interpretation in favor of a search within a set of parameters to apply to the model making process. As such, this approach had a higher computational cost and took much longer to accomplish for the three models being created. The auto\_arima function also contains its own parameter to incorporate seasonality, making it an auto-SARIMA model in this scenario.**

|  |  |
| --- | --- |
| **The auto\_arima models began their searches for the ideal models using the same p and q (first and third) values and ended with values of 0 and 3 for store type A, 2 and 1 for store type B, and 0 and 1 for store type C. These values were then used with the models created to make projections to compare with the actual sales data, as was done with the models created using SARIMAX.**  **Results**  **The models created by SARIMAX and auto\_arima were evaluated visually by plotting the predictions and actual sales data as well as quantitatively using the root mean squared error (RMSE), mean absolute error (MAE), and R2 values for each model. The standard deviations and** |  |

**median values for the department’s weekly sales were included for additional context since these values are partially related to the sizes of the projected value, weekly sales, which is in the thousands of dollars. Arrangements of this data are shown below.**

|  |
| --- |
| **SARIMAX Metrics** |
|  |
| **auto\_arima Metrics** |
|  |

**For both model creation methods, the model used for store 30 department 1 (type C) were rated poorly by R2 values and the reason for this was due to the predicted values for just before September 2012 sales behaving one way when the actual sales data behaved another way. This was due to the nature of SARIMA modeling, using past behavior to inform modeling and projections. Before September 2010, the department’s sales were moderately stable and then increased while 2011 saw a drop before the start of the September and then increased as it had in 2010. The model predicted a drop in sales before September 2012, like 2011, when the actual sales data was moderately stable, similar to 2010. This is likely could have been avoided if more than 2.75 years of sales data was available.**

**The RMSE and MAE values were each smaller when compared to the median weekly sales values and standard deviations of the department’s weekly sales. The models for departments at store types B and C produced MAEs that were in the single thousands of dollars, an amount of daily variability of around $150 per day (found by dividing by 7), which could be reasonably expected. Type A stores appeared to have higher volume, explaining the larger error values. Even still their MAE translated to a daily variability of around $500 per day. In terms of MAE, the models appear to do well even in situations where the previously mentioned behavior acted in one previous manner when the other was repeated.**

**Conclusion**

**Both modeling methods proved worthwhile for predicting sales data given past sales data and economic figures for the past and predicted timeframes. The departments modeled were not described or labeled outside of an individual number, which makes further interpretation of their sales difficult. A daily variability of $150 in an electronics or grocery department could be a single customer while the same value in a clothing department could span multiple customers.**

**The models likely suffered from a possibly small training set for each model. In a professional setting, a retail chain would be able to provide its modelers with a larger sales history for each department at each store in the chain. These numbers would allow for an even more powerful model to be created than the ones shown in this analysis.**

**When it comes to choice of model creation method, SARIMAX was much more efficient computationally. auto\_arima trades that efficiency for convenience given its parameter search process. The “correct” choice would depend on the situation. An individual modeler on their own may be better off using SARIMAX while a commercial modeler needing to make a large number of models and has no concerns about computational efficiency for model creation could likely choose to use auto\_arima.**

# References

Brownlee, J. (2019, August 21). *A Gentle Introduction to SARIMA for Time Series Forecasting in Python*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/

Justin, L. a. (2022, September 7). *How to build ARIMA models in Python for time series forecasting*. Retrieved from Youtube: https://youtu.be/-aCF0\_wfVwY?si=zB6hYwpk5-N7MPZ4

*SARIMAX*. (2024, October 3). Retrieved from statsmodels: https://www.statsmodels.org/stable/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html#statsmodels.tsa.statespace.sarimax.SARIMAX

Singh, M. (2017). *Retail Data Analytics*. Retrieved from Kaggle: https://www.kaggle.com/datasets/manjeetsingh/retaildataset

TeKnowledGeeK. (2021, October 21). *Time Series Forecasting using SARIMAX and compared with ARIMA*. Retrieved from Youtube: https://youtu.be/JO0gFP\_q4uc?si=HAYi1WxAkZWEM3yp