PREDICTIVE RETAIL ANALYTICS

Nicole Aguilera

MILESTONE 5

DSC 630 T301

June 6, 2021

Executive Summary

This project analyzes the sales that a department, or departments, of a store has made in the past to attempt to predict the sales for that said sector in the next year. It is important for businesses to understand how their sales will do in the future so that they are able to prepare their store with the correct amount of inventory, marketing, etc. to meet and, potentially, exceed those predictions. These preparations are not included or advised in this project.

The results of predicting the sales for a store are good. For a store with only one department or one type of merchandise, I was able to find a method that predicts the sales of the store for the following year within 1.45% of the actual sale. The method that was found to predict the sales of a single department, with the consideration of whether it was a holiday week or not in that location, can predict results with a 3.52% error. When looking at a method that can predict the sales of two departments within the same store, along with the corresponding holiday weeks, the method found was able to predict results with an error of only 2.70%.

Abstract

The purpose of this study is to predict sales for a store department for the following year. This analysis will focus on modeling the sales of one department within a store, the sales of one department and the binary variable of whether it is a holiday or not, and two departments from the same store along with the holiday binary variable. After running all of the models, the model that performed the best was analyzing the single department, with an

error of 1.45%. The model running two departments along with the holiday variable could predict results with a 2.70% error, and the one department with the holiday variable predicted results with a 3.52% error.

Introduction

Retail has become such a big industry, not only in the United States, but all around the world. It can range from an online shop specializing in one type of product to superstores of discount retailer that are considered to be a one stop shop for everything that you need. Each of these establishments have consistent reports on sales, which can be used to find out what the sales are in each department. The model created through this project will be used to predict the sales by store for the following year, separated by department.

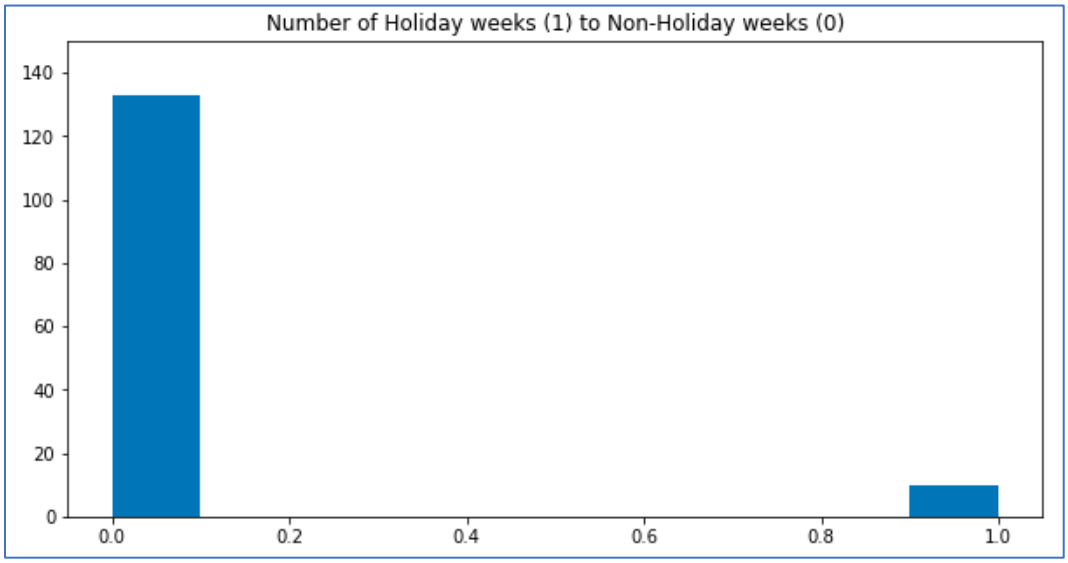
Data with regards to retail stores can be all over the place. This is mainly because these types of stores can potentially cover a variety of different products in many different departments. Discount sales that come with the holidays can vastly affect the profits for the store. Having the tool of predictive analytics will assist us in understanding of when sales take place for each department and how that affects the prediction of sales in the future.

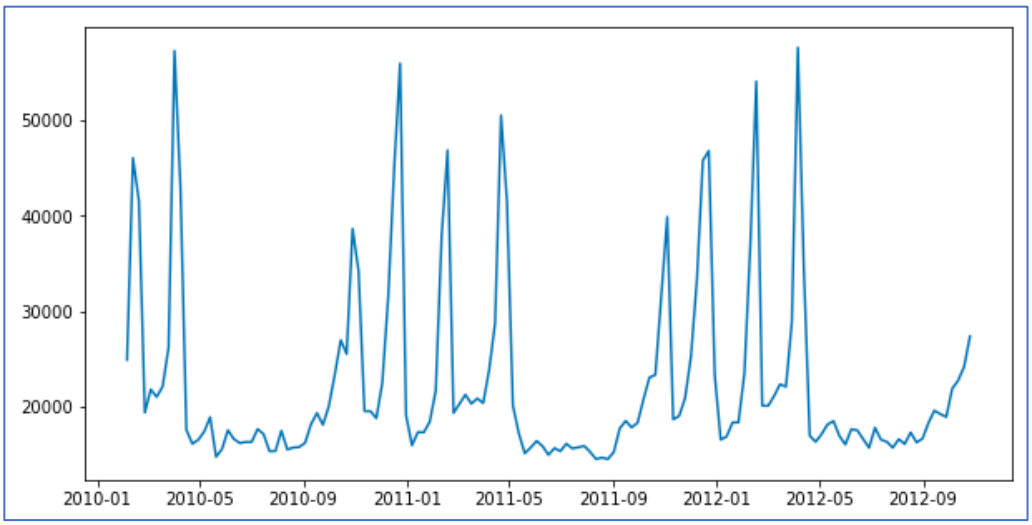
The goal of this project is, as stated before, to predict what the sales will be for the following year after the data from the set. The project should be able to produce a tangible model that, ideally, retailers would be able to use for their own benefits within their store. This will be done by analyzing sets of retail data that include weekly sale numbers and the indication of whether it is a holiday for a store and its departments for that week. Since the store and

department names remain confidential, selecting any store and department is random. The data set that I will be using is found on Kaggle, called “Retail Data Analytics”.

As stated before, the main goal of this project is to build a successful model to aid retail stores in predicting sales for the following year, with “model” being the key word here. Because we are not able to receive actual data from retailers in the area or even online, it is hard to make this model applicable to all stores using the data that we have. Once the model is perfected, data of sales by week from individual stores would ideally be able to be inserted into the model, which would give the results of predictive sales for the following year for that particular retailer. These results would most likely vary from the model but align with the individual store’s projection.

Before getting into the predictive analysis, I was able to do some basis exploratory data analysis on the set. The graphing tools of Python and R programming were used to make these visuals. They were constructed to visually show the data that was being worked with. Below are the histogram of the binary variable indicating if the week is a holiday or not, and a line chart displaying the data, respectively.





There are 143 data points that span over 3 years, which means that we are evaluating the sales of 143 weeks per department. In the histogram shown above, there are about 10 data points or weeks that are considered to be holidays, and the remainder of the weeks are not. The line graph of the data points shows the sales per week. You can see that the spikes are more or less consistent for each year, which could possibly indicate a sales week that has a holiday that would contribute to those increases in sales.

Methods

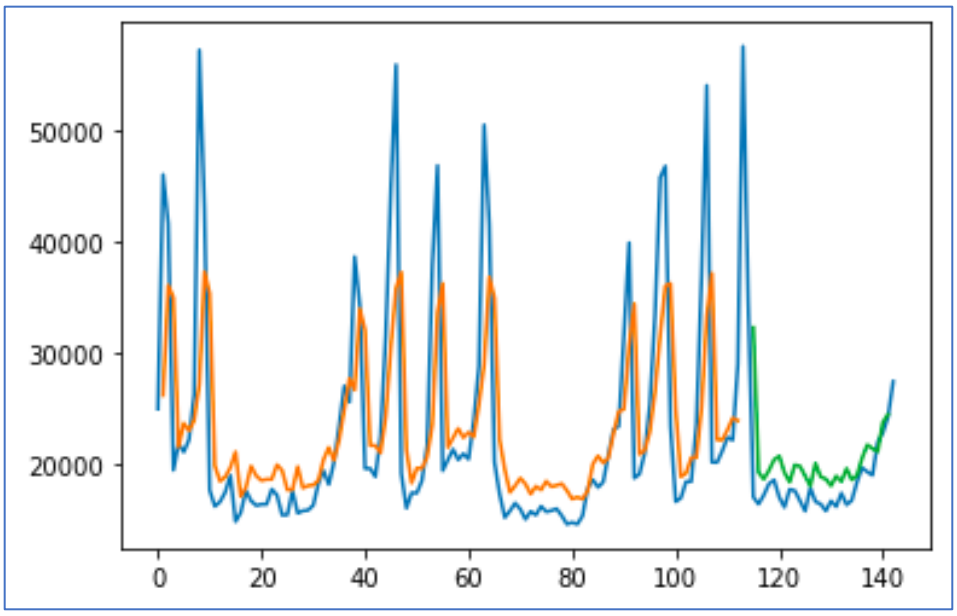
The methods that were used in this project were based around the "Long Short Term Memory" (LSTM) network. This type of network is a "recurrent neural network that is trained using Backpropagation Through Time" (Brownlee, 2016). It can create large recurrent networks that address specific sequential problems in machine learning that are difficult and is connected through layers instead of neurons. This time series prediction model used the Keras deep learning library. The first LSTM model that I used was phrased as a regression-type problem only because the length of time between each give data point was the consistently the same

(one week), so the specific days did not matter. Because of this, the first model only required the sales data. The second model that was used was a variation of the previous model, called LSTM regression with time steps. This time steps model did the same thing; however, the model was slightly refined due to the past observations of data being used as input features rather than separate features that the former model used (Brownlee, 2016). This makes the regression time steps model more accurate than the regression by itself.

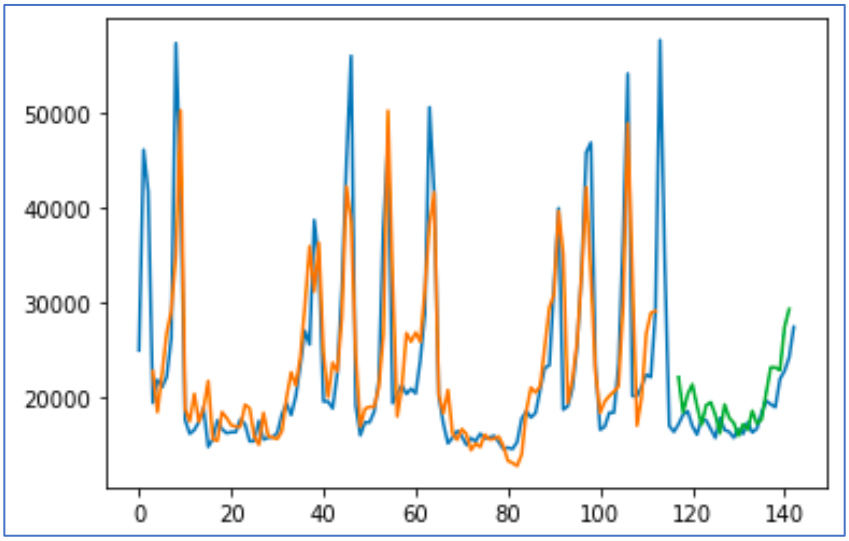
The final type of LSTM model used was the multivariate time series forecasting that also utilized the Keras deep learning library in Python. This was used for the third and fourth models of the project. The LSTM network is able to model multiple input variables, which gives them an edge over classical linear methods of modeling (Brownlee, 2017). This model concatenates the values of the variables or columns because there are more variables to work with compared to the first two models.

Results

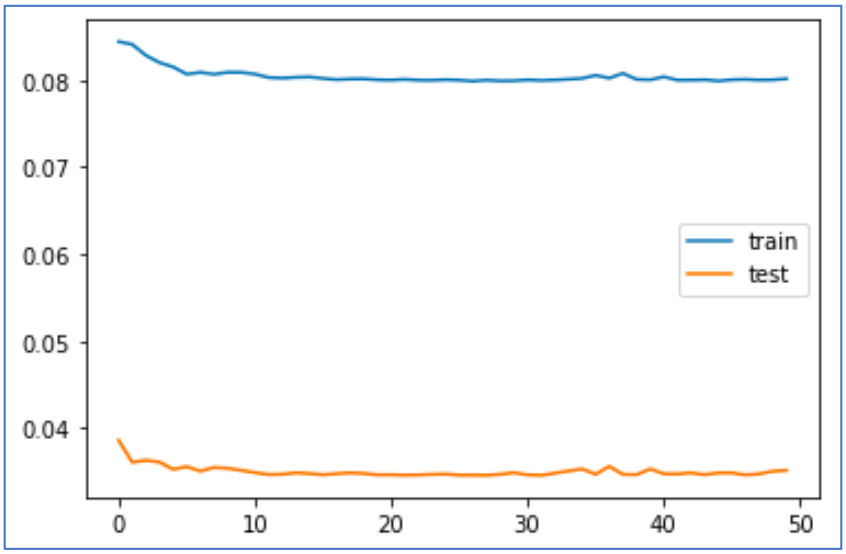
The results of these models were actually fairly well, in my opinion. The error of the models was less than 5% for each and the models themselves ran smoothly. For the first model, which looked at the LSTM regression of the sales variable by itself, the train score of the model was able to predict future prices within $7,846.19, while the test score of the model was able to predict $3,665.01 within the future sales prices. The model had a loss of only 3.36%, which is under the error threshold for 5% that is usually allocated for prediction models, in general. This is a good result because the train and test sets of the data were able to reflect the original data set very closely, as shown in the following figure.



The next model analyzes the sales variable in LSTM regression with time steps model. This proves itself to be more accurate than the previous model because of its results. The train scores of this set were able to predict futures results within $5,090.35, which is over $2,000 of a smaller margin than the preceding model. The test scores were smaller in this model, as well, because the results were able to predict future sales within $2,655.37, compared to the former simple regression model. Visually, these two sets of data were also able to reflect the original data set that can be seen in the following figure.

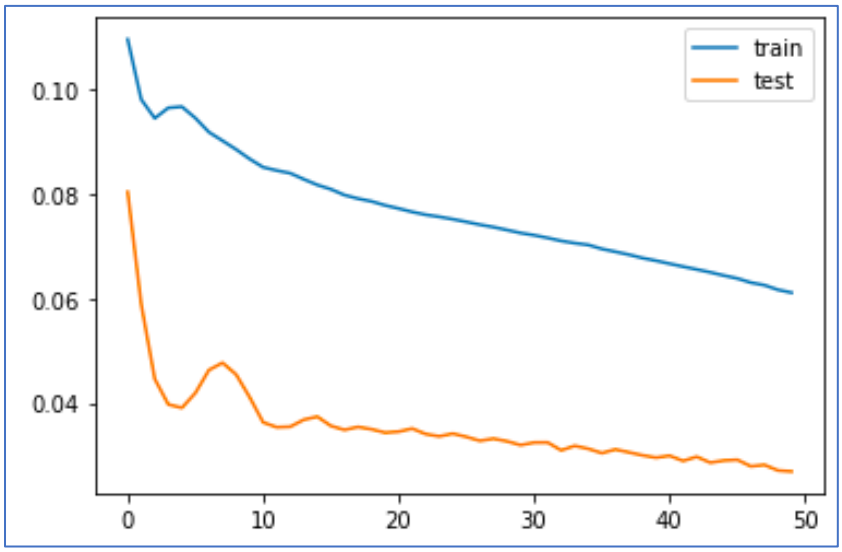


Now, we move into the multivariate time series forecasting by using the LSTM model through Keras. These models were a bit more loosely fit, as far as their results go, than the univariate models that have already been covered in this project. I first began with adding in the holiday variable to see if the model could predict the future sales of that one department by using the indication of whether that week's sales had a holiday or not. This model was able to predict future sales results within 3.52%, along with an 8% loss in the training data set. This is not ideal that the train and test data would produce results that have a difference of 0.045; however, it is still good that the test data set produced lower results than the training data set. Below is the graph of the model for both the training and test data sets.



The final model added in another random department's sales data to the LSTM multivariate time series forecasting model. This model more or less showed the same kind of results as the one before it, with not great but good results. The results of the model were able to predict future sales within 2.70% using past data observations, with a loss in the training data set that converged to around 6.11%. It had a smaller difference of training and test data

results than the previous model, with a difference of 3.41%. Below is the visual of the model's results for both the training and test data sets for this model.



Conclusion

Overall, the LSTM models that displayed time series forecasting in this project were able to predict future sales based on the past observations of sales data provided with relatively low error margins. The results of the univariate LSTM with the simple regression and time step models showed to be very accurate and mimic the data. This could be implemented into the retail industry in a store that does not have multiple departments with their own data that would help to predict future sales data for the next year. The univariate LSTM regression time step model would also be the one that I would feel most comfortable with publicizing, as the margins of error were the lowest here.

As far as the multivariate regressions went, they performed well, seeing that the model applied was self-taught. The results of the training and test sets seemed to be far apart, however, the differences between the two were both were only less than 0.05. Personally, I

think that if this project were to be done again, I would attempt to find a different model to fit for the multivariate models, possibly an ARIMA model. This was not done due to the time constraints of the project and learning a new model outside of the course. The reason that I stayed with the LSTM format for all four models was to keep some form of consistency between the different types of data being run through the models.

Acknowledgements

First off, I would like to personally thank my father, Pete Aguilera, who has supported me in my data science journey throughout this program and has also reviewed my project as a second teammate. I would also like to thank Dr. Werner for answering my countless questions about this project and both assisted and supported me, even when the data proved that it would need more than the models that were provided in this course. My former employer of the Walt Disney Company also deserves some thanks, due to them offering me the opportunity to further my education through their Disney Aspire program for its Cast Members. Finally, I would like to thank Bellevue University for offering this program and allowing me to pursue my dreams that have landed me a career of a Data Analyst.

References

1. Brownlee, J. (2016). *Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras.* Machine Learning Mastery. https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/
2. Brownlee, J. (2017). *Multivariate Time Series Forecasting with LSTMs in Keras.* Machine Learning Mastery. https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/
3. Retail Data Analytics. (2017). Kaggle. https://www.kaggle.com/manjeetsingh/retaildataset