Retail Sales Forecasting

Springboard Data Science Career Track

Dylan C Neal

9/1/20

Introduction:

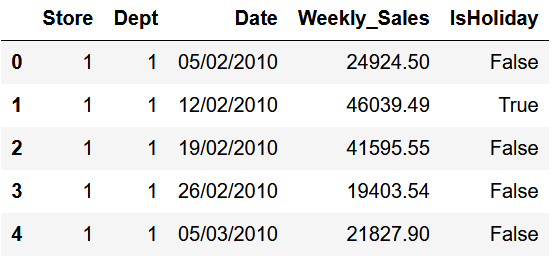
Sales forecasts are essential information for financial planning. Without accurate sales forecasting, it will be difficult to plan effectively for the future. By anticipating future sales, companies are able allocate resources to reduce waste and optimize revenue. Executives will be able to plan for major purchases while having confidence they will make the sales to be able to afford them.

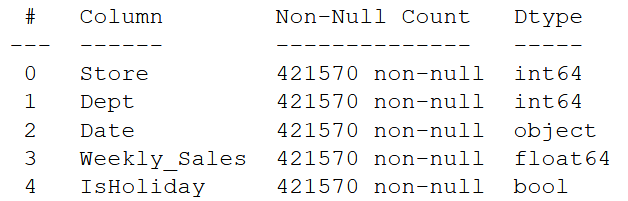
Problem:

A major national retail chain has been keeping weekly sales data on each department in 45 of its stores. Historic sales data is limited, and only goes back ~3 years. Can a model be built to forecast sales data for each department in every store?

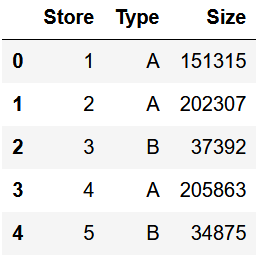
Step 1 – Data Wrangling:

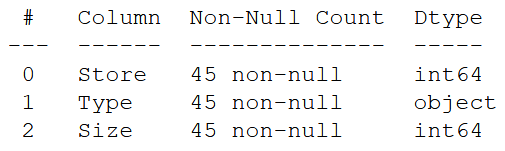
* The data consist of 3 tables:
  1. SalesDF - Weekly sales data on every department in each store. Includes IsHoliday feature which indicates whether there is a major holiday that week. Holidays include the Super Bowl, Labor Day, Thanksgiving, and Christmas.



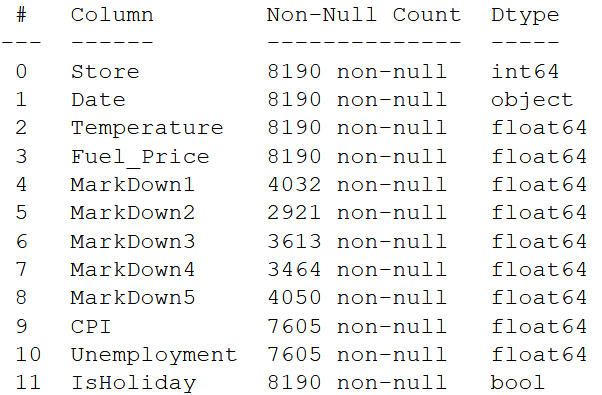
*SalesDF.head()*  
  
  
*SalesDF.info()*

* 1. StoreDF - Details including the size and type of each store. Stores are categorized as type A, B, or C stores.



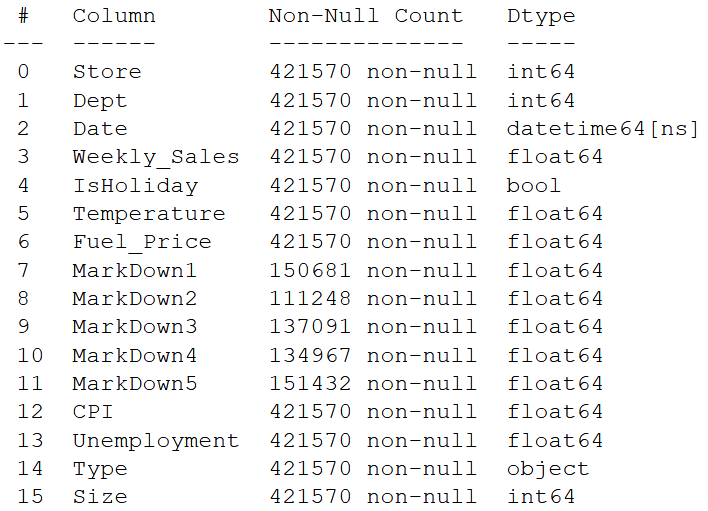
*StoreDF.head()*  
  
  
*StoreDF.info()*

* 1. FeaturesDF - Additional weekly data for each store. Data include the average temperature and fuel price in the region, consumer price index, and unemployment rate. There are also 5 markdown features with numerical data. However, markdown data is not present for every store and is only available after November 2011.



*FeaturesDF.info()*

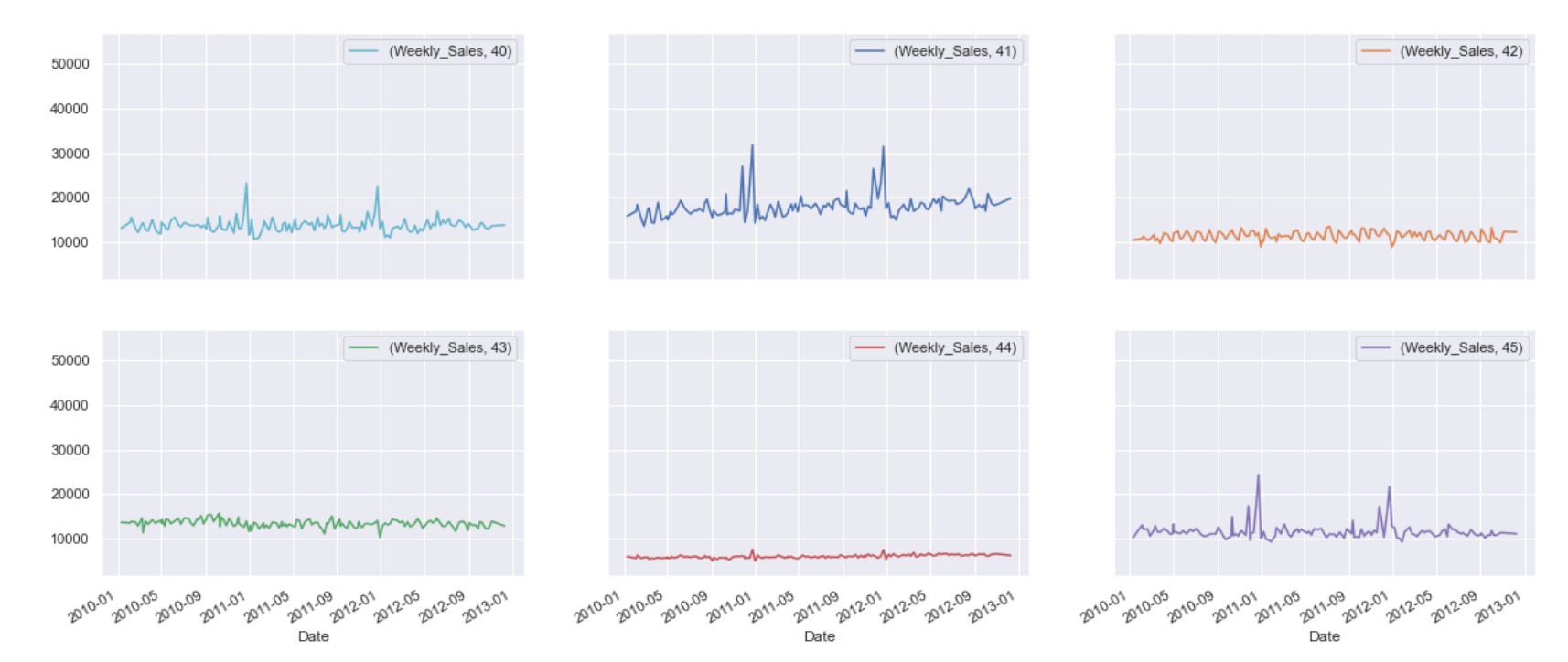
* The date columns were converted to datetime type. FeaturesDF was joined to SalesDF on the Date, Store, and IsHoliday features. StoresDF was joined to the main dataframe on the Store feature. Info on the joined dataframe is displayed below.



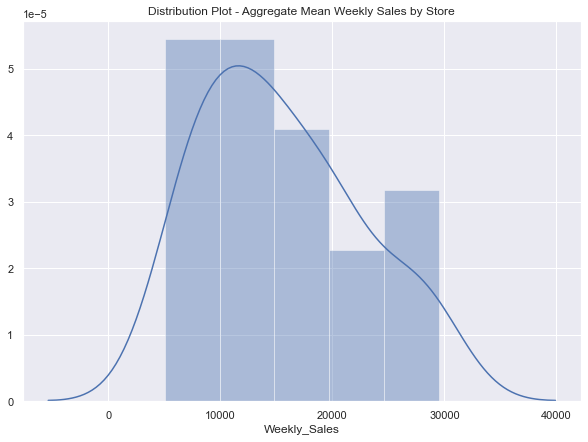
* Note that the only features containing null values are MarkDowns. Rather than removing nearly 75% of the data, it was decided to ignore markdown data and build the model without it.

Step 1 – Exploratory Data Analysis:

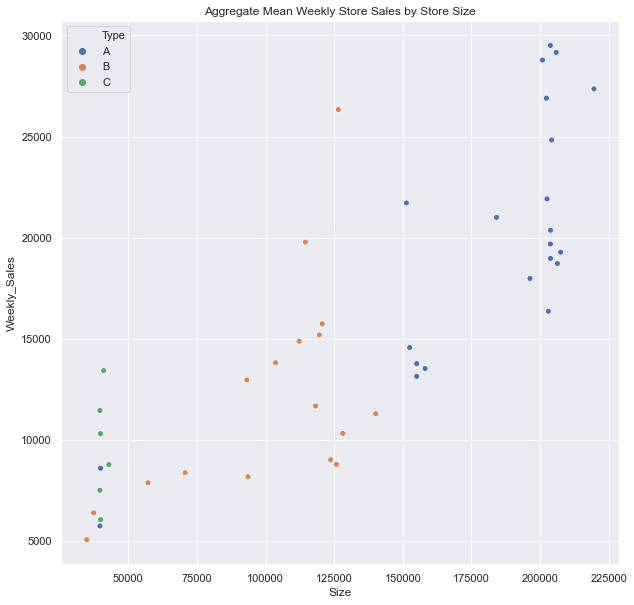
* The data were sorted by date, and then grouped by Store & Date. Weekly sales mean was aggregated and then plotted by store.
* A few of the line plots are seen below. While there is seasonality, it does not look like there is much trend in this data. If the data were over a longer span of time, there would most likely be observable trend.



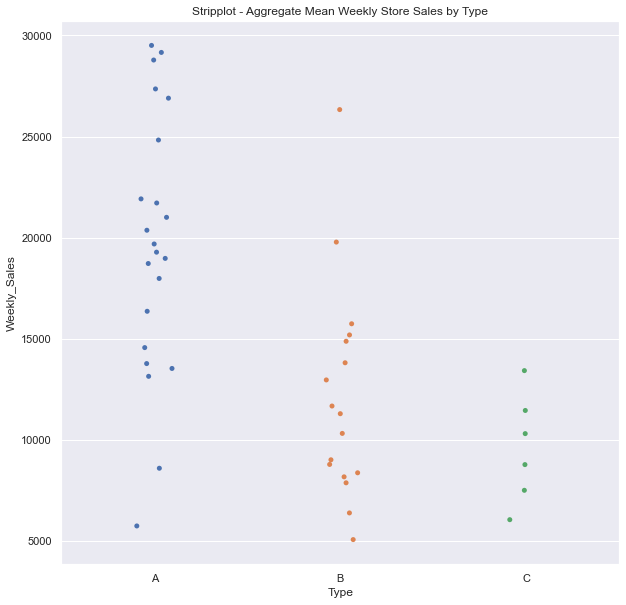
* To get a sense of store sales spread, a distribution plot was made.



* The data were then grouped by Store, Type, and Size, and weekly sales means were aggregated. From this a scatterplot was created to see the relationship between store size and weekly sales. Points were hued by type.

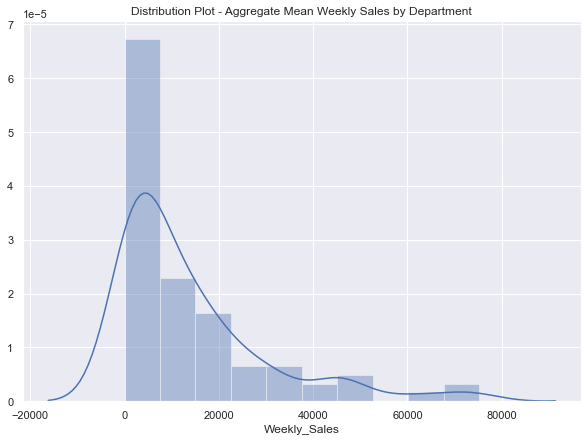


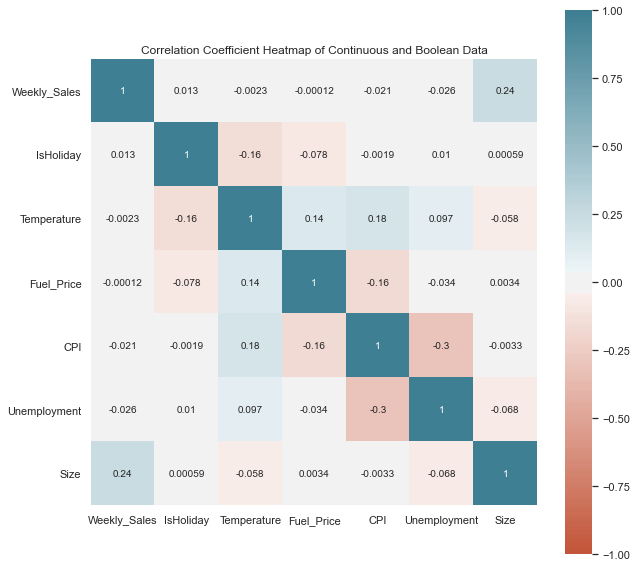
* A strip plot of aggregate mean weekly sales by store type was made.



* There is a relationship between size of store, store type, and mean weekly sales.
* Line plots of aggregate mean weekly sales by store type were plotted by date as well.



* A correlation coefficient heatmap was made
* To get a sense of department sale spread, a distribution plot for departments was made
* 

Most departments company-wide have weekly sales < $20,000. However, there are departments that have upwards to $80,000 in average weekly sales.  


* Store size seemed to have the biggest correlation with weekly sales. Weekly sales must primarily depend on store, department, and date. It does seem that temperature, fuel price, consumer price index, and unemployment all negatively correlate with weekly sales, which could all make sense logically.

Step 3 – Pre-processing

* Store type values were converted from letters to numbers A:2, B:1, C:0
* Month and Week features were extracted from the Date Feature.
* Markdown columns were dropped
* Dummy features were created for Department, Month, and Week.
* Because Type seemed to have a major impact on the weekly sales, it was decided to split the data out by type and create separate models for each store type
* Once the data were split by type, dummy features were also created for Store.
* Before dropping the Date feature, a time-series train test split was performed, setting the training data before January 1st, 2012 and the test data after.
* A MinMaxScaler was used to scale the continuous features. The training data were fit to it and then the test data were transformed by it. This was done separately for each data set.
* 5-Fold cross-validation segments were generated from the training data for each store type. This was done manually via custom function. The cross-validation split was done in an appropriate manner for time-series, in which each training set consists only of observations that occurred prior to the test observation that forms the test set. This is called walk-forward validation.

Step 4 – Modeling

* 3 types of models were trialed for each store type – Linear Regression, Ridge Regression, and Random Forest Regression
* Each model was tried out using a manually defined cross-validation function.
* Hyper-parameter tuning was performed using model specific functions defined to take parameter range inputs and iterate over the values.
* In addition to using the manually defined Hyper-parameter cross-validation tuning functions, scikit-learn’s GridSearchCV and RandomizedSearchCV were also used as comparison.
* Standard Linear Regression performed very poorly. It overfit to the training data and did not translate to modeling test data.
* Ridge Regression performed significantly better, with CV R2 test scores maxing out at 0.9061, 0.6825, and 0.8177 for Store Types 0, 1, and 2, respectively.
* Random Forest Regression performed the best but took significantly more resources to fit the data and perform hyper-parameter tuning.
  + CV R2 test scores maxed out at 0.9809, 0.9009, and 0.9509 for Store Types 0, 1, and 2, respectively.
* In an attempt to improve modeling of Store Type 1 data, trials performing PCA dimensionality reduction prior to Random Forest Regression were performed.
  + The best CV R2 test score achieved was 0.8874, so PCA was not used in the final models.

Who Cares?

* Executives (interest in maximizing profit)
* Supply Chain Management (interest in stocking stores and departments appropriately)
* Hiring (interest in staffing stores and departments effectively to reduce redundancy)