Capstone 2: Retail Data Analysis

Jarom Hatch

Introduction

Problem Statement

- Retail data analytics can be challenging given:
  - Supply chain delays
  - Global shortages
  - Sudden market shifts
- Given a set of feature variables, can weekly sales be predicted accurately using a machine learning algorithm?

Target Audience

Major grocery store chains and retailers:

Data Overview

Initial Datasets

- **3** main datasets provided:
  - Stores, Features, Sales
- **Stores**:
  - Store number, Store type, Store size
- **Features**:
  - Store number, Date, Temperature, Fuel price, Markdowns (1 through 5), Consumer price index (CPI), Unemployment rate, Is the date a holiday
- **Sales**:
  - Store number, Department, Date, Weekly sales values, Is the date a holiday

- Features data: February 5th, 2010 to July 26th, 2013
- Sales data: February 5th, 2010 to October 26th, 2012

Data Wrangling

- Data from all 3 datasets needed to be combined for machine learning
- The features dataset had NaN values in a few columns
  - Missing markdown values before November 2011
  - Last rows of CPI and unemployment values
- Markdown NaN values replaced with 0
- CPI NaN values replaced using polynomial fitting
- Unemployment NaN values replaced with forward filling method

Data Wrangling (cont.)

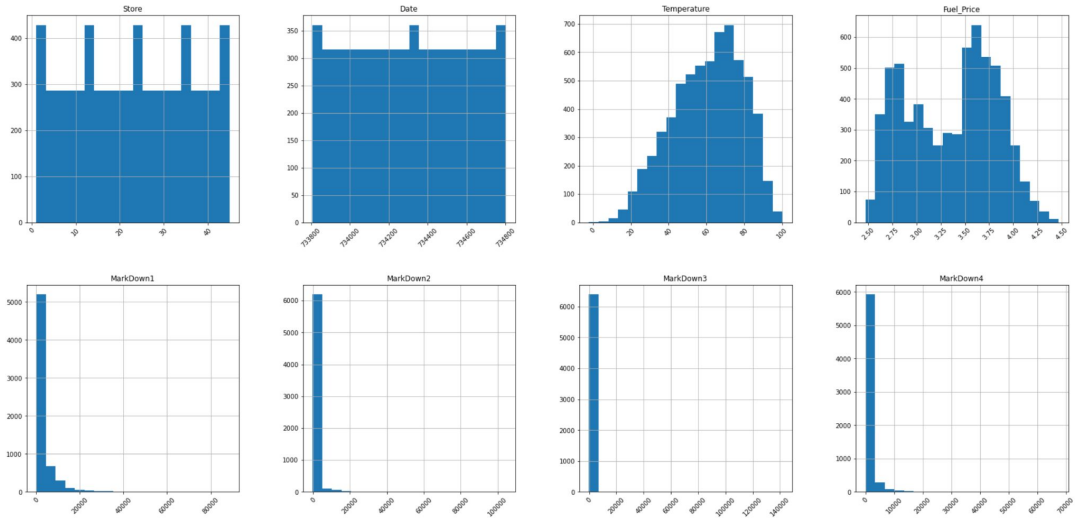
- Datasets had both categorical and continuous variables
- Categorical variables were converted to work with machine learning
  - Store type changed using encoder
  - Dates changed to ordinal values
- Separate weekly sales values for each department caused issues
- All weekly sales for every store and date were added together to make machine learning easier to set up
- All datasets were combined based on date and store number to create one main dataset
- The extra features data without weekly sales we split off for future prediction testing

Exploratory Data Analysis

- Values from the complete dataset were plotted as histograms to determine any initial trends in the data

Normal

Bimodal



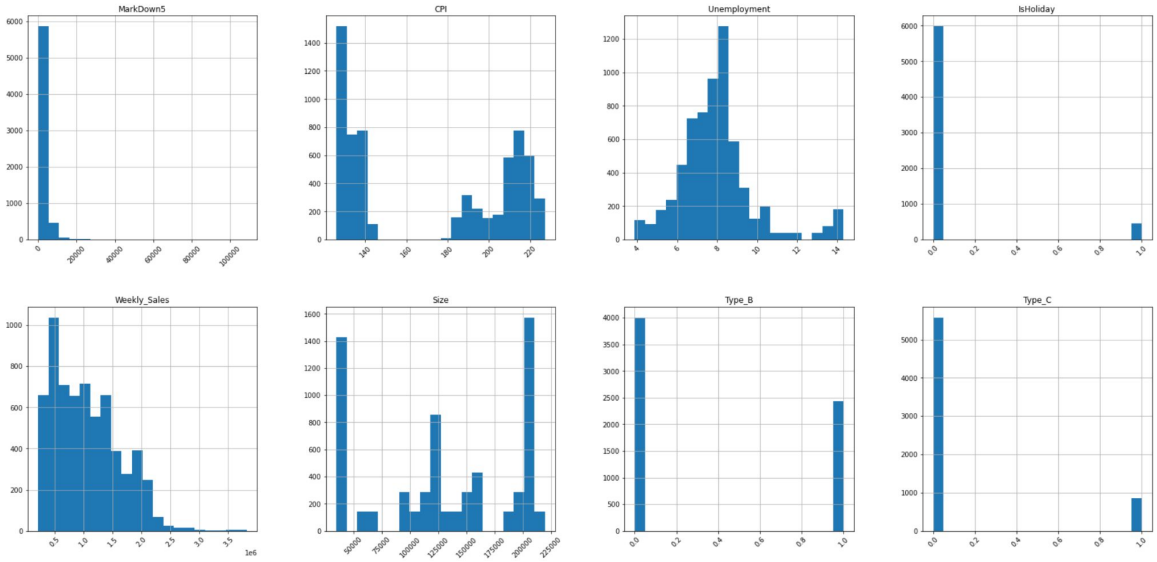
Right skewed

Exploratory Data Analysis (cont.)

- The second set of values is shown below

Bimodal

Normal



Right Skewed

Exploratory Data Analysis (cont.)

- The continuous features showed interesting trends
- Temperature and unemployment seemed normally distributed
- CPI and fuel price seemed to have bimodal distributions
- Weekly sales and most of the markdown columns had right skewed distributions
- Nothing in the visualizations showed problems with the data or any abnormally distributed data

Exploratory Data Analysis (cont.)

- Correlation matrix showing correlations with weekly sales
  - Store size had largest positive
  - Unemployment had largest negative

Machine Learning

Machine Learning Setup

- Dataset split with training and testing sections
- Features were first scaled to prepare for machine learning algorithms
  - Standard Scaler
  - Min/Max Scaler
  - Unscaled
- Machine learning models chosen from typically used model for continuous target variables
  - Linear regression
  - Lasso
  - Ridge
  - ElasticNet/ElasticNetCV
  - Random Forest Regressor
- Models judged on R^2 value based on testing data

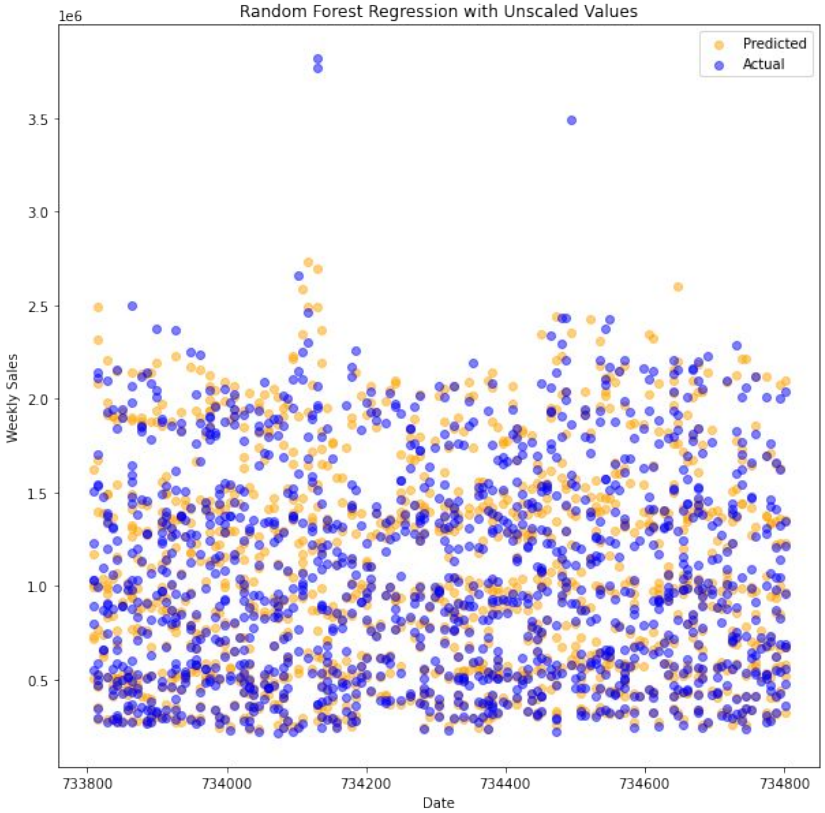
Machine Learning Metrics

Metrics from best models:

Model	MAE	RMSE	R^2
Linear	221404	301197	0.718
Lasso	221403	301197	0.718
Ridge	226313	306939	0.708
ElasticNet	221412	301177	0.718
Random Forest	71262	134777	0.944

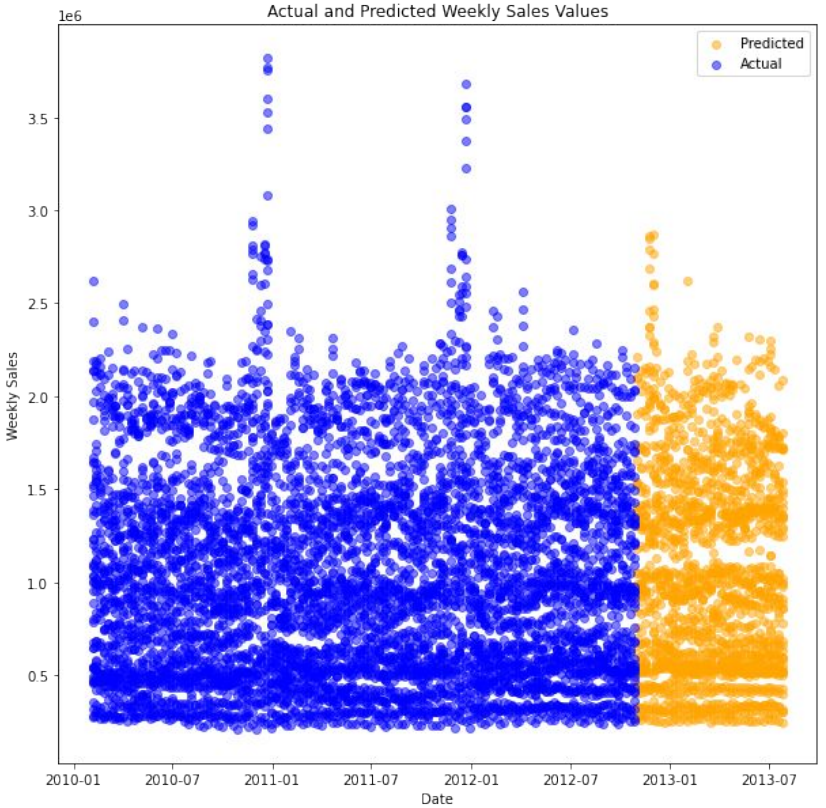
Random Forest Regressor

- The random forest regressor using the unscaled version of the training and testing data produced the best results
- On the right is the plot of predicted versus actual values using this regressor



Random Forest Regressor

- On the right is the actual values from the original dataset plotted with the predicted values using the random forest model on the separate dataset with no weekly sales
- Apart from some deviation towards the middle of the data, the trends seem to follow those of the previous dataset



Conclusion

Conclusions

- Machine learning was able to accurately predict future weekly sales values per store
- Best model was default random forest regressor using unscaled training and test data
- This model can be used continuously and easily by the customer for future predictions and can be retrained using new data for even better results