Forecasting Retail Sales

Steven Miller

Retail Forecasting

- Benefits of better forecasting:
  - Better Planning
    - Staffing can be optimized
    - Forward projections are more accurate
    - Better information for internal and external analysts
    - Fewer surprises
  - Identify drivers of revenue
    - Able to focus on drivers to improve financial performance

The Data

- Retail Data Analytics
  - Retrieved from Kaggle
  - Historical sales data for 45 stores within a chain
  - Weekly sales broken down by department
  - Three tables of data
    - Sales
    - Features
    - Stores

Sales Table

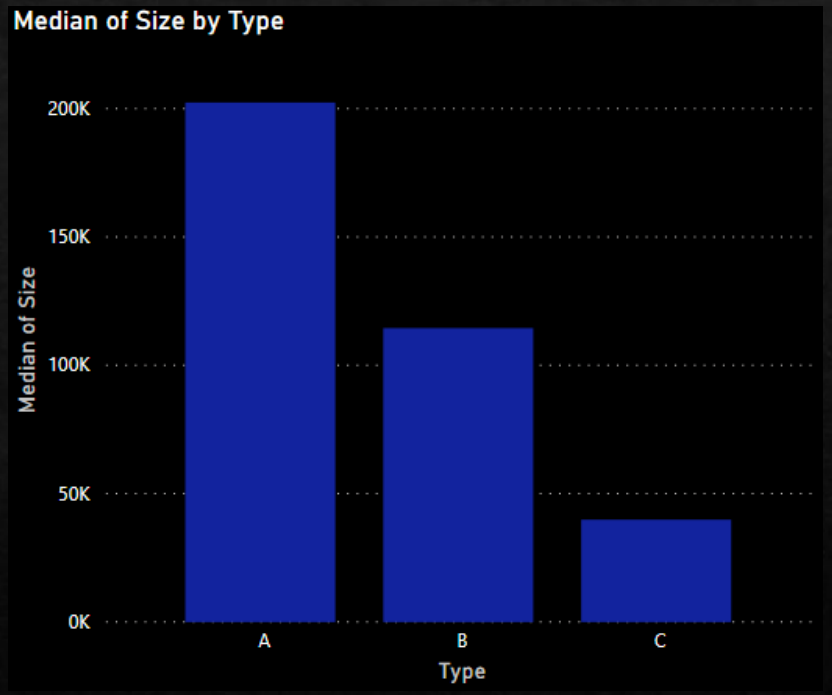
- 421,570 Rows
- 5 Features
  - Store – integer for store id number
  - Department – integer representing the department within the store
  - Date – date representing the week corresponding to the data in the row
  - Weekly_Sales – floating point value representing sales in dollars for the department for the week
  - IsHoliday - Boolean value representing whether a holiday was occurring during the week. Four weeks of each year are considered holiday weeks
    - Super Bowl
    - Labor Day
    - Thanksgiving
    - Christmas

Features Table

- 8,190 rows
- 12 Features
  - Store – integer for store id number
  - Date – date representing week for the features
  - Temperature – Average temperature for store for week in Fahrenheit
  - Fuel Price – price of fuel in dollars per gallon in store's region
  - MarkDown1, 2, 3, 4, 5 – Anonymized data related to the presence or absence of promotional markdowns for the store for the week
  - CPI – Consumer Price Index in store's region
  - UnemploymentRate – unemployment rate in store's region
  - IsHoliday – Boolean value indicating whether there was a holiday during this week

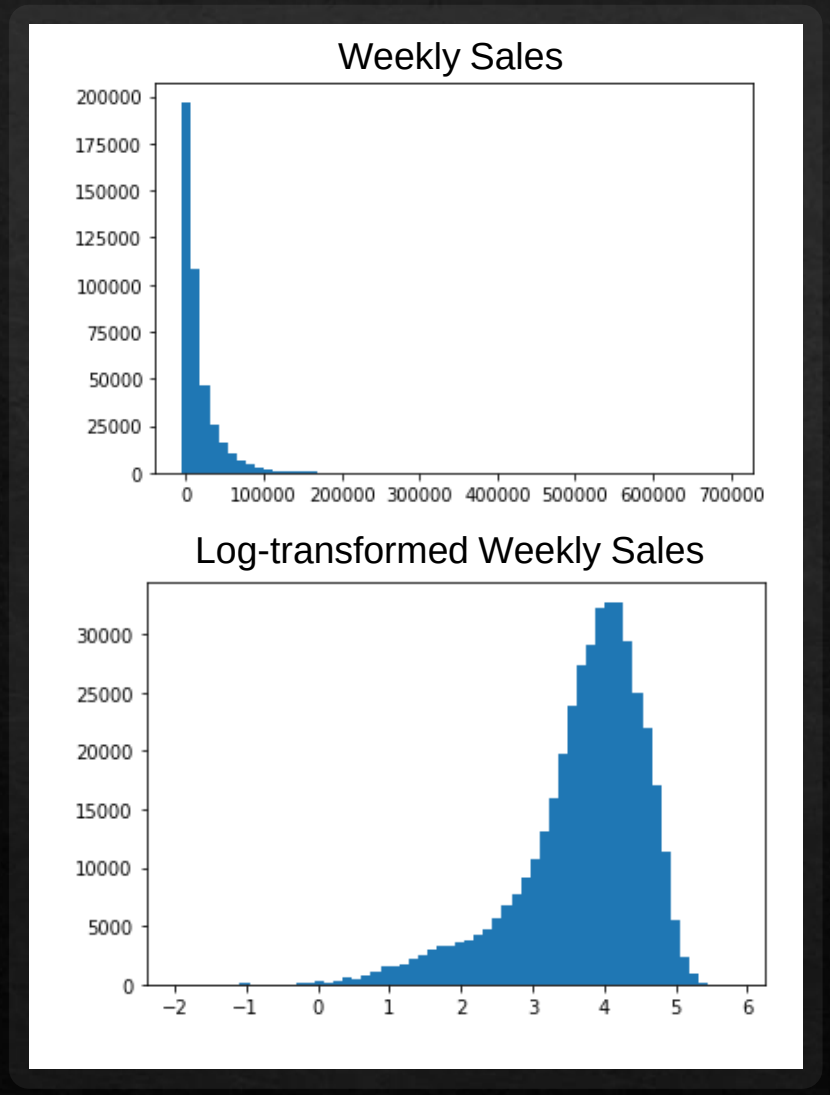
Stores Table

- 45 Rows
- 3 Features
  - Store – integer for store id number
  - Type - a categorical variable denoting A, B, or C indicating type of store. Given size of store types in square feet, seems in line with large retail chain such as Target or Wal*Mart
    - A – "Super" store
    - B – Standard store
    - C – Small store, possibly in more urban area
  - Size – size of the store in square feet



Model Goals

- Predict future weekly sales based on available data
- Answer initial research questions:
  - Does a relationship exist between CPI and weekly sales?
  - Unemployment rate and weekly sales?
  - Fuel prices and weekly sales?
  - How do promotions impact sales?



Data Exploration

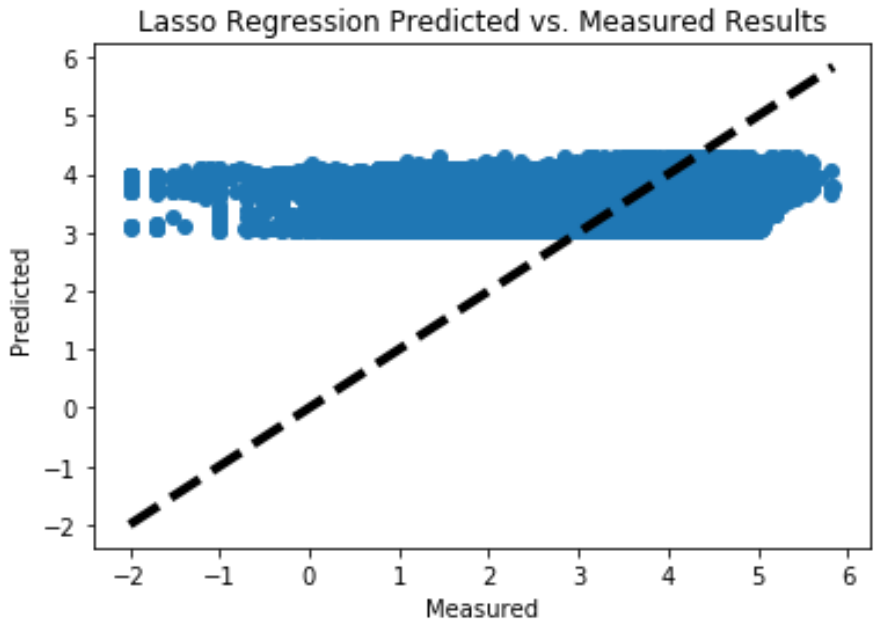
- First concern: Economic indicators may be too correlated to be individually useful.
  - Correlation of -0.19 between CPI and Fuel Price.
  - -0.30 between Unemployment and CPI
  - -0.03 between Fuel Price and Unemployment
- Sales figures were significantly left-skewed
  - Log-transformation of sales moving forward to better normalize sales distribution

First Model

- OLS Regression
- Included 10 features
  - CPI, Temperature, Fuel Price, Unemployment, MarkDown1-5, IsHoliday
- Result: Adj. R^2 of only 0.01
- Not a good start!

Second Model

- Added additional features from the stores table.
  - Store Type, and Store Size added to the model as features
- Adjusted $R^{2:}$ 0.112
- Better, but still not great!
- Additional models were attempted with the same data
  - Ridge Regression with similar results
  - Random Forest with 80 estimators: Adj R^2: 0.01
- Current data isn't capturing variances outside of $100-$1,000 in weekly sales.

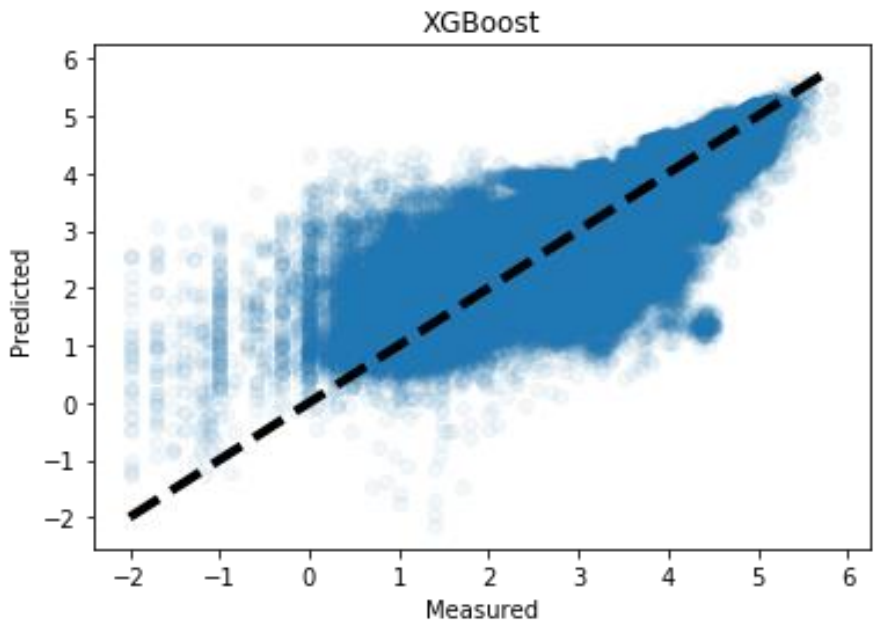


Third Model

- Added dummy variables indicating store and department ID numbers
- Large growth in model size, from 13 features to 134.
- Adjusted R-Squared 0.679!
- Additional models tested:
  - Random Forest Regressor: Adj R-Squared 0.76 with 80 estimators

XGBoost and Parameter Tuning

- Optimizing the model now that we're on the right track
- Using XGBoost regression algorithm to minimize RMSE
- Default parameter results: RMSE 0.505
- After parameter tuning:
  - RMSE: 0.176440
  - Adjusted R^2: .795
- Additional features and tuning have better captured the range of possible sales figures



Research Questions

- Does a relationship exist between CPI and weekly sales?
  - Yes, but it's small. A 1-point increase in CPI results in a 0.001 **decrease** in log-transformed sales.
- Does a relationship exist between the unemployment rate and weekly sales?
  - Yes, but it's a positive relationship. A 1% increase in unemployment corresponds to a 0.006 increase in log-transformed sales.
- Does a relationship exist between fuel prices and weekly sales?
  - Yes, a $1 increase in fuel prices would correspond with a 0.0228 decrease in log-transformed sales.
- How much of an impact do promotions play in sales?
  - Tough to say due to anonymized data.
  - Coefficients for MarkDown features immeasurably small.
  - Holiday weeks have a 0.011 increase in log-transformed sales.