Retail Data Analytics

Sunday, February 25, 2018

10:22 PM

Overview:

Every retail business wants to attract more customers, sell more to each customer, and have them come back to shop again the next time they need something. To do this, a store needs to offer customers the products they want at the right prices and great customer experience etc.  These decisions can be hard to make, therefore, data analytics has gained increasing popularities among top retailers.

The goal for this project is to use the limited historical sales data across 45 stores at different locations for the year of 2010 - 2012 to answer the following questions:

1. Can we predict the weekly sale by store ID for the next year.
2. Which store and department is more sensitive to markdowns, weather, cost of fuel, and holiday indicator?

Expected clients for this project will be store managers. The results based off this project can help them make better business plan for the coming year. For example, by providing them with predicted weekly sales, they can leverage these information to re-evaluate their cost and benefit analysis. Knowing which department/store is more sensitive to price change can help store managers to plan promotions for holidays etc.

Data:

I will used the datasets from Kaggle website. The dataset contains historical sales data for 45 stores located in different regions from 2010-02-05 to 2012-11-01. It also provides some basic economic stats at the time.

Dataset: <https://www.kaggle.com/manjeetsingh/retaildataset/data>

* Store - the store number
* Date - the week
* Temperature - average temperature in the region
* Fuel\_Price - cost of fuel in the region
* MarkDown1-5 - anonymized data related to promotional markdowns. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA
* CPI - the consumer price index
* Unemployment - the unemployment rate
* IsHoliday - whether the week is a special holiday weekStore - the store number
* Weekly\_Sales -  sales for the given department in the given store

Approach:

* Step 1 : Data wrangling
  + Merge retail, feature and sales dataset with store and date as primary key to form a clean, easy to analyze datasets. The observation will be each store on each day with specific department code
  + Treat Missing values, duplicate entries, wrong column types etc.
* Step 2: Data visualization
  + Diagnose trends, relationships and outliers etc.
* Step 3: Statistic Inference
  + Remove trends, seasonality
  + ARIMA
  + Fir the model with variable of interest, and controlling variables

Hypothesis:

1. CPI and unemployment rate will affect the sales
2. Certain stores will be more sensitive to markdowns.
3. I assume weather and holiday would have an impact on the sales for all stores, and cost of fuel would have an impact on certain stores only.

Deliverables:

1. Python code posted in Github repository
2. Final report and slide deck