CPTS 475/575 Data Science

Project Progress Report

Harsha Maddipati - 11633702 | Nikhil Chakravarthy - 11655295

Introduction

We are given 5 years of store-item sales data to predict 3 months of sales for 50 different items at 10 different stores. Sales of items in the stores change every day of the year, we must analyze at which part of the year the sales increase and decrease when making our prediction. This means that we take an account for seasonality when creating our model.

We take on this challenge as time-series forecasting and apply gradient boosting machine learning algorithms (which is suited for time-series data) to get our prediction. Time-Series forecasting is a demanding area for all stores as it helps them predict what kind of products are being bought at what part of the year. It also lets them prepare the inventory and allot a budget to different types of items.

Dataset Description

The training data set consists of the number of sales of a particular item at a store on a date. While test data consists of itemID, item, storeID and date. We must predict the sales of the items on the dates.

We make use of the date column to analyze how sales differ for each item at different stores. We also check the yearly trends and weekly trends to get an idea of how seasonality affects sales.

Training Data

date	store	item	sales
2013-01-01	1	1	13.0
2013-01-02	1	1	11.0
2013-01-03	1	1	14.0
2013-01-04	1	1	13.0
2013-01-05	1	1	10.0

Test Data

date	id	store	item
2018-01-01	0	1	1
2018-01-02	1	1	1
2018-01-03	2	1	1
2018-01-04	3	1	1
2018-01-05	4	1	1

Approach

Going through resources online about time-series forecasting methods, we saw that gradient boosting techniques are the most popular ones. We decided to go through with it as well. There are lots of ML libraries which make use of this technique, we are going to focus on scikit-learn's Gradient Boosting regressor and XGBoost.

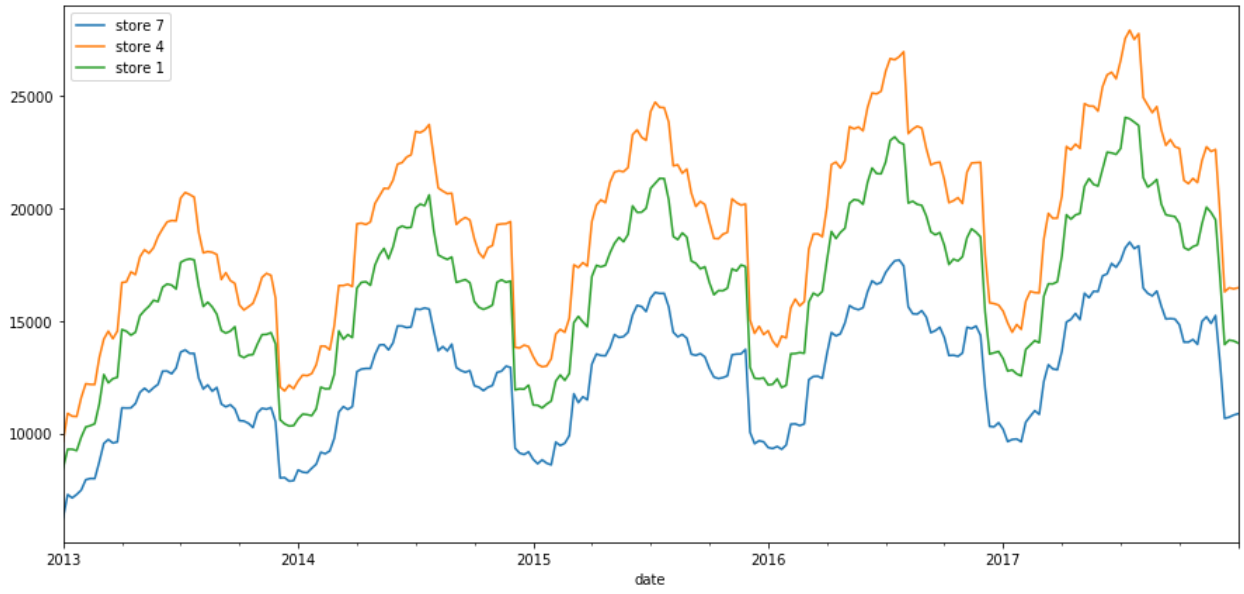
- **Pre-Processing**

  We started off by checking for any missing values or null values in the dataset but have not found any. Date attribute is elaborated to sperate fields to get a better understanding of how the sales vary monthly, yearly and weekly.

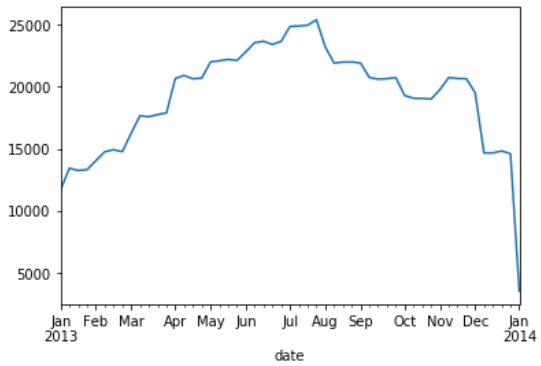
date	store	item	sales	year	day	month	dayofweek
2013-01-01	1	1	13.0	2013	1	1	1
2013-01-02	1	1	11.0	2013	2	1	2
2013-01-03	1	1	14.0	2013	3	1	3
2013-01-04	1	1	13.0	2013	4	1	4
2013-01-05	1	1	10.0	2013	5	1	5

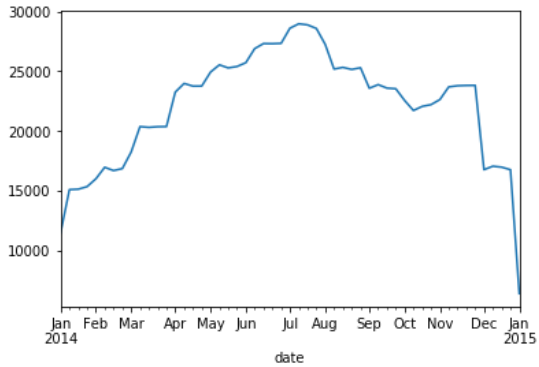
- **Seasonality**

  To understand seasonality, we took 3 stores at random and plotted their total sales of items. We notice similar trends for all the stores.



Let us take a closer look at the sales of store 2 in the year 2013 and 2014





Here we notice the sales go up during summer and winter, just before the holidays. While sales are very low at the start and end of the year.

Splitting the date column into multiple subcolumns helps us extract more information on the data, which can be used for better prediction. We plan to use yearly, weekly and monthly data when making our prediction.

References

- https://hackernoon.com/gradient-boosting-and-xgboost-90862daa6c77
- https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/
- https://towardsdatascience.com/trend-seasonality-moving-average-auto-regressive-model-my-journey-to-time-series-data-with-edc4c0c8284b
- http://www.statsoft.com/textbook/time-series-analysis