IST: 707
Prof. Ying Lin

# Predict Future Sale with Machine Learning



Shloak Gupta


Henglong So


Singh Abhishek

Syracuse Unverisity
School of information


## Motivation

Data Science plays a huge role in forecasting sales and risks management in Retail Stores Sector. Majority of the leading retail stores implement Data Science to keep a track of thier customer needs and makes better business decision. We believe by using Maching Learning algorithms will be able to predict the total of products sold in every stores.

## Package

numpy, pandas, seaborn, sklearn
matplotlib, xgboost, SARIMA, dash

## Data Source

This data is publicized on Kaggle by Russian software firm - 1C Company.

## Data Description

60 Stores

22,170 Unique Items

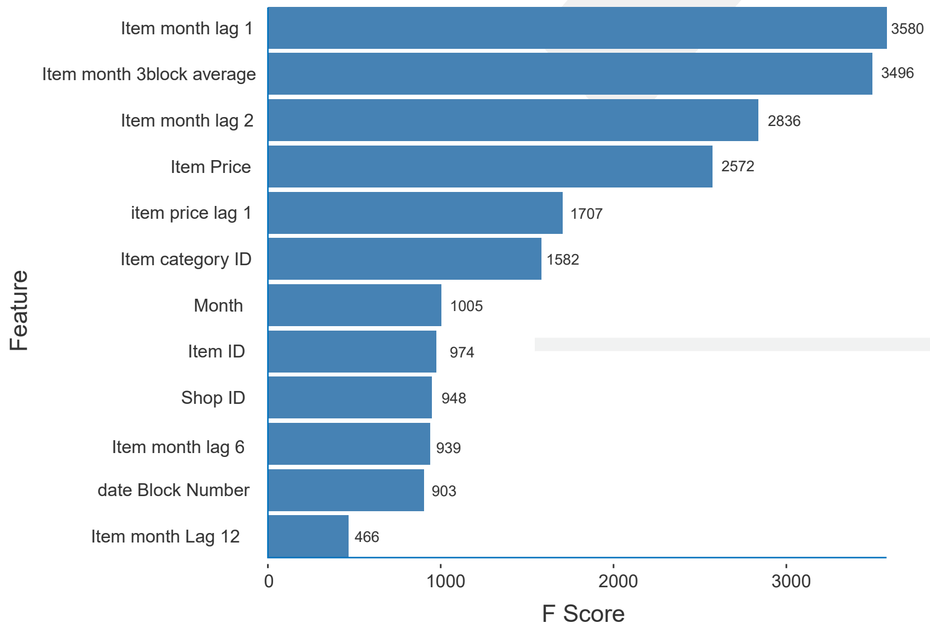
2013 to 2015 (Daily)

Data is retrieved from Kaggle public by one of the largest Russian software firrm - 1C Company. This dataset is challenging with time-series consisting of daily sales data. This data contains 6 variables and 2,935,849 observations.

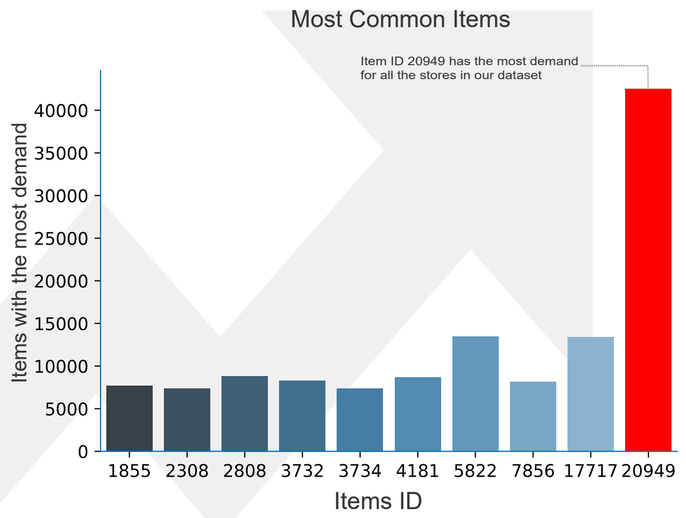
## Feature Engineering

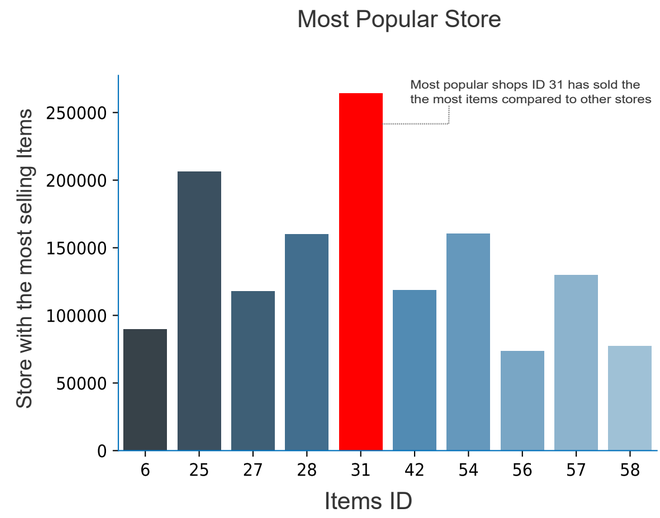
Supply Chain Analytics is a Time Series Analysis problem. To capture the trends and sesonality new features need to be added like lags, month of year, moving averages etc. The other features inculcated are product discontinued or new product added.

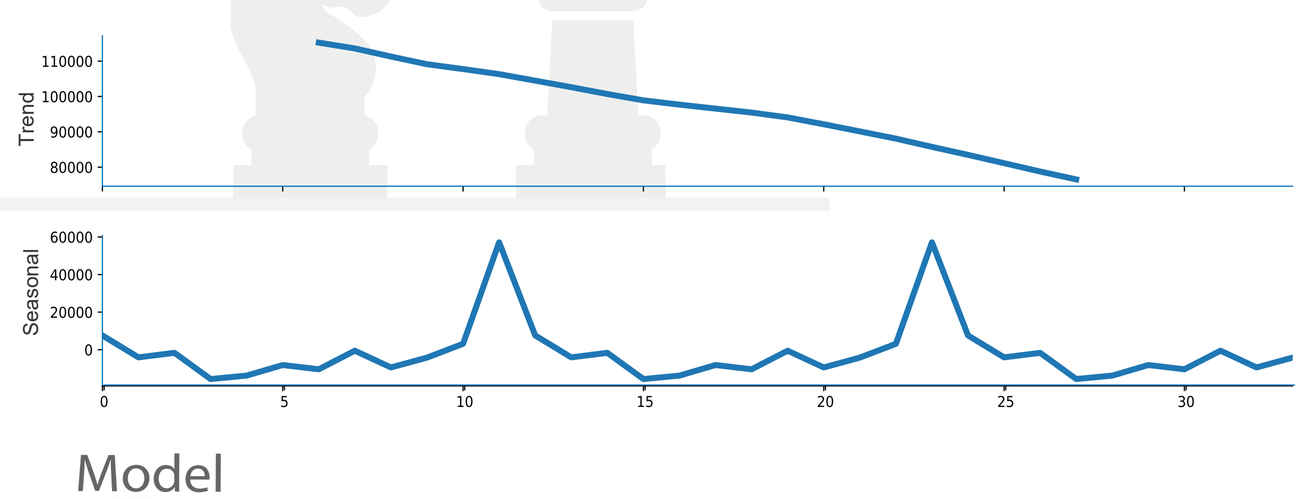
| Feature | Description |
|---|---|
| Time: | |
| Date of transaction | Date of sale |
| Shop features | |
| Shop ID | Unique ID for each Shop |
| Product features: | |
| Item ID | Unique ID |
| Item Price | Price of item |
| Item Category | Category of item |
| Availability | Is Product Discontinued or not? |
| Trend analysis features: | |
| Lag | Lag of 3 months to capture trend |
| Moving Average | To adjust to outlier |
| Seasonality | Month of year to utilize monthly trend |



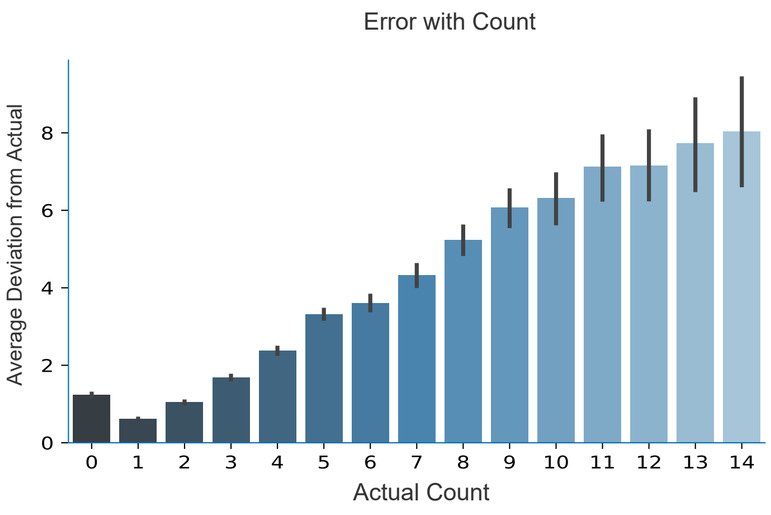

## Key Statistics



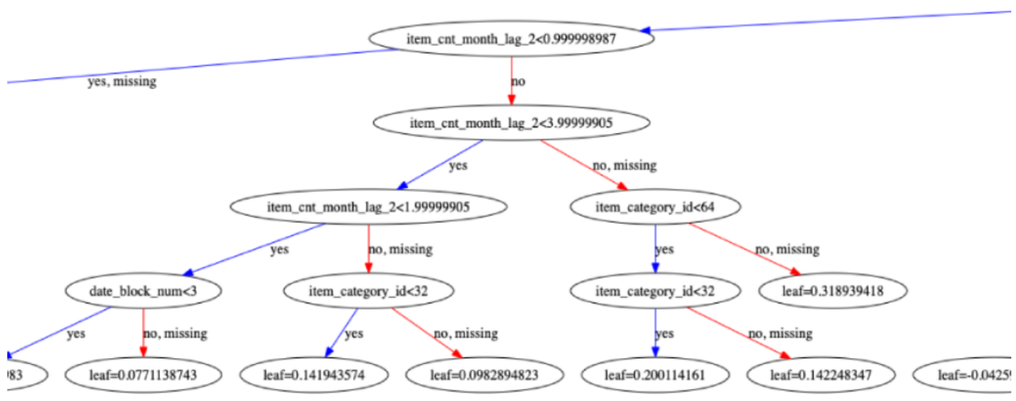




## Result




We used data for the last month for cross-validation, the performance of the model was measured using root mean squred error as the metrics for our task. Following is the performance RMSE : 0.33

## Model




We are using XGboost as our model. This is a sequential technique which work on the principle of an ensemble. It combines a set of weak learner and delivers improved prediction accuracy. This model outcomes are weighed based on the outcomes of previous instant t-1. the outcomes predicted correctly are given a lower weight and the one miss-classified are weighed higher for the next round. XGBoost also performs faster than other algorithms with large datasets and thus made it an easy choice for our task.

## Future Work

1- Drilling down blocks in item category and shop to get trends of specific shops & categories
2- Find ways to deal with items having higher counts and which are very volatile
3- Solutions to deal with saprse data in an efficient way

## Join Us on LinkedIn

Shloak Gupta

Henglong So

Singh Abhishek