Time-series forecasting of retail products sales using Machine Learning

Vartika Tewari

Abstract

Accurate Time series prediction has various real applications. In this project we tackle a similar problem provided by Kaggle, as part of their ongoing "Predict Future sales" competition [2]. We will handle challenging time series data of daily sales from the largest Russian software firms- 1C Company. Given past data the goal is to predict monthly sales for every possible shop item pair, and make the model robust to the seasonal changes and trends. We will employ and compare three models, ARIMA, xgboost and LSTM for prediction, and evaluate them on Root Mean Squared Error(RMSE).

1 Introduction

Forecasting future sales has been an important problem in all organizations. These help in making decisions on budgeting, prospecting, and other revenue-impacting factors.[4] Having accurate estimation of any organizations growth and changes helps in preparing strategies for short-term demand and supply and identification of areas for investment.

There are a lot of factors which can impact sales forecast like, market changes, product changes as well as seasonality. Robustness of the model to such changes is an essential part of the project.

In this project we'll work with a challenging time-series dataset, from an ongoing Kaggle challenge "Predicting Future Sales" [2] consisting of daily sales data, provided by one of the largest Russian software firms - 1C Company. The data consists of shop and product changes each month. The problem statement is to accurately predict total sales for every product and store in the next month. That is, to predict monthly sales for every possible shop item pair.

2 Proposed Project

2.1 Data

The data-set consists of daily sales data. There are about 3 million such records in the training set, collected over 60 shops selling 20,000 unique items.

Dataset	Time series
Min Date	01.01.2013
Max Date	31.12.2014
#Items	22170
#Shops	60
#datapoints	2935849

Table 1: Dataset Statistics

Training data consists of records with information that a particular item had been sold in a particular shop, in a particular day, in the training period. The training period is about one and a half year. Testing period is the month that falls on training period. The data-set also has missing data as not every item is sold in the above time period.

2.2 Proposed techniques

We will do data cleaning to handle missing values, exploratory data analysis and feature engineering consisting of temporal and non-temporal features to capture the stationarity as well as seasonality and trends of the historical data provided.

We propose to try out three models for our regression problem:

2.2.1 Auto Regressive Model

As AR models are the go to models for time series prediction, we will use ARIMA [3] to form the baseline.

2.2.2 Extreme Gradient Boosting

We will use extreme gradient boosting technique like XGBoost [1], because it has advantages over general gradient boosting in terms of providing regularization through a combination of both the ridge and lasso regressions. It also handles different type of sparsity patterns in the data efficiently.

2.2.3 Deep Learning Models

Finally we'll compare how adding complexity helps in improving accuracy by employing deep learning models [5] like Multilayer Perceptron(MLP) Model and Long Short Term Memory(LSTM) Model. As LSTM's capture context and dependence between items in a sequence, they will help in time series prediction.

References

[1] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[2] *Kaggle-Challenge:Predict Future Sales*. URL: https://www.kaggle.com/c/competitive-data-science-predict-future-sales/overview.

[3] Wes McKinney, Josef Perktold, and Skipper Seabold. "Time series analysis in Python with statsmodels". In: *Jarrodmillman Com* (2011), pp. 96–102.

[4] *Sales Forecasting Blog*. published May 14, 2020, updated May 14 2020. URL: https://blog.hubspot.com/sales/sales-forecasting.

[5] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.