Store Sales – Time Series Forecasting

Chakshu Pandya
Computrer Science and Engineering
Ahmedabad University
Ahmedabad, India
chakshu.p@ahduni.edu.in

Darshit Pandya
Computer Science and Engineering
Ahmedabad University
Ahmedabad, India
darshit.p@ahduni.edu.in

Dhruvin Kotak
Computer Science and Engineering
Ahmedabad University
Ahmedabad, India
dhruvin.k@ahduni.edu.in

Fena Changela
Computer Science and Engineering
Ahmedabad University
Ahmedabad, India
fena.c@ahduni.edu.in

***Abstract*—We study the usage of machine-learning models for sales predictive analytics. The main goal of this paper is to consider main approaches and case studies of using machine learning for sales forecasting. The effect of machine-learning generalization has been considered. This effect can be used to make sales predictions when there is a small amount of historical data for specific sales time series in the case when a new product or store is launched. A stacking approach for building regression ensemble of single models has been studied. The results show that using stacking techniques, we can improve the performance of predictive models for sales time series forecasting. (*Abstract*)**

Keywords—Machine learning, Forecasting, Time Series, Sales, Holidays events, Oil, Stores

I. Introduction

Sales prediction is an important part of modern business intelligence. It can be a complex problem, especially in the case of lack of data, missing data, and the presence of outliers. Sales can be considered as a time series. At present time, different time series models have been developed, for example, by Holt-Winters, ARIMA, SARIMA, SARIMAX, GARCH, etc. We need to have historical data for a long time period to capture seasonality. However, often we do not have historical data for a target variable, for example in case when a new product is launched. At the same time, we have sales time series for a similar product and we can expect that our new product will have a similar sales pattern. Sales data can have a lot of outliers and missing data. We must clean outliers and interpolate data before using a time series approach. We need to take into account a lot of exogenous factors which have impact on sales. In this case, we study the usage of machine-learning models for sales time series forecasting. We will consider a single model, the effect of machine-learning generalization and stacking of multiple models.

II. Literature review

Modern time series forecasting methods are essentially rooted in the idea that the past tells us something about the future. Of course, the question of how exactly we are to go about interpreting the information encoded in past events, and furthermore, how we are to extrapolate future events based on this information, constitute the main subject matter of time series analysis.

Typically, the approach to forecasting time series is to first specify a model, although this need not be so. This model is a statistical formulation of the dynamic relationships between that which we observe and those variables we believe are related to that which we observe.

The “classical” approach to time series forecasting derives from regression analysis. The standard regression model involves specifying a linear parametric relationship between a set of explanatory variables and the dependent.

III. Implementation

Here, in this project they have provided us some datasets, which are train.csv, test.csv, sample_submission.csv, store.csv, oil.csv, holiday_events.csv.

We will use these datasets to predict the sales using time-serial forecasting.

A. *Fetching dataset*

First of all, we have to fetch or read all the data to perform all the other operations like, training, testing, model building, etc.

B. *Training data*

Training data is an extremely large dataset that is used to teach a machine learning model. Training data is used to teach prediction models that use machine learning algorithms how to extract features that are relevant to specific business goals.

C. *Calender Engineering*

Dates and times are rich sources of information that can be used with machine learning models. However, these datetime variables do require some feature engineering to turn them into numerical data.

Here we have made date in calendar function from beginning of train until last date of test. We also have concatenated with oil price.

D. *Correletions*

Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable.

- We have found the correlations between calendar and average oil price.
- Then We have fetched the holiday dataset and created the future engineering for holiday. To predict the sales on holiday as well as on the regular days.
- We have also found the correlations between things which sale on holidays and working days as well as the effect of oil price on the sales.

Figure 1 Oil price Plot

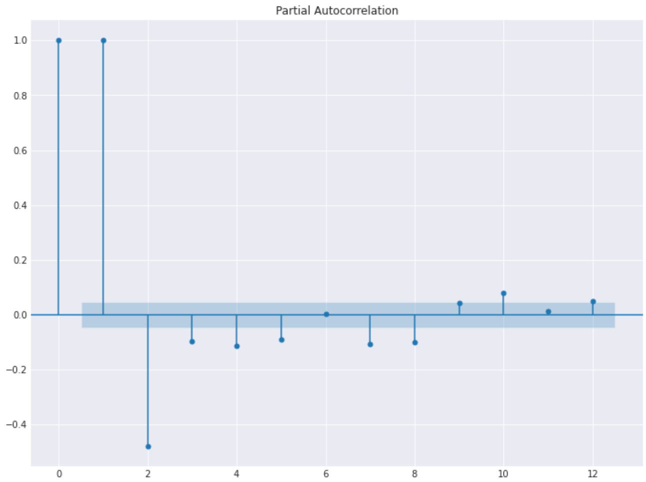

Figure 2 Lag Plot

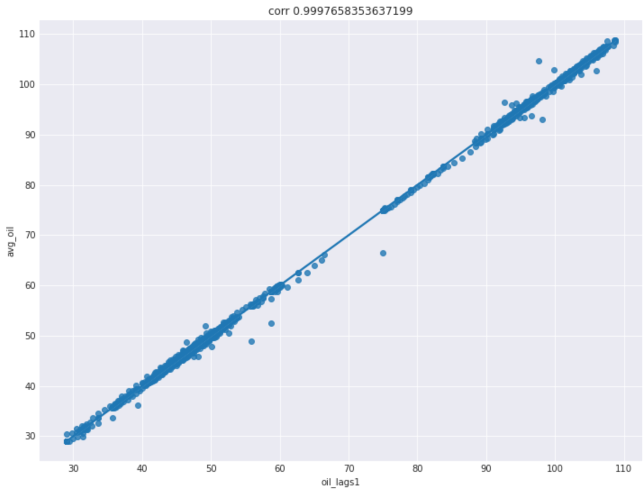

Figure 3 Correlation Plot between Avg vs lag1 oil

Visualization of products

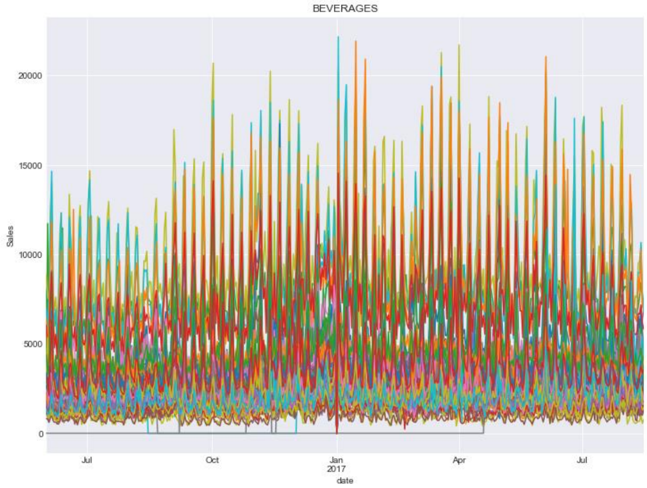

Figure 4 Store Beverages sales

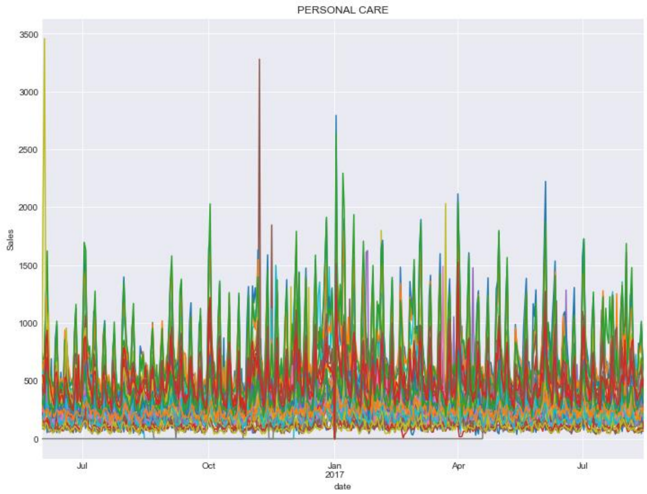

Figure 5 Store Personal Care Sales

E. *Feature Engineering*

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features.

Here, we have fetched the holiday dataset and joined the calendar with the holiday dataset. Then we have transformed our categorical data into the numerical format by implementing the one-hot encoding method. Then the algorithm will check if any of the holidays are overlapping based on the calendar and drop them. Then we have performed the deterministic process by Fourier transform.

F. *Machine learning moddels*

In Multiple Linear Regression, the target variable(Y) is a linear combination of multiple predictor variables x_1, x_2, x_3, ...,x_n.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

Where, Y= Output/Response variable. b0, b1, b2, b3 , bn....= Coefficients of the model. x1, x2, x3, x4,...= Various Independent/feature variable

Support Vector Regression is similar to Linear Regression in that the equation of the line is y= wx+b. In SVR, this straight line is referred to as hyperplane.

G. *Root mean squared error*

The root-mean-square deviation or root-mean-square error is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed.

IV. RESULTS

Figure 6 linear and SVR Regression plot for Book sales

Table 1 Sales prediction

Store No.	Categories	Sales
1	AUTOMOTIVE	4.101209915
1	BABY CARE	0
1	BEAUTY	4.555176899
1	BEVERAGES	2262.699439
1	BOOKS	0.083691179
1	BREAD/BAKERY	376.8196113
1	CELEBRATION	13.12506081
1	CLEANING	735.4486687
1	DAIRY	753.1388008
1	DELI	131.5335143

Table 2 Final output

Id	sales
3000888	4.10121
3000889	0
3000890	4.555177
3000891	2262.699
3000892	0.083691
3000893	376.8196
3000894	13.12506
3000895	735.4487

Table 3 Root mean squared error

Categories	Root Mean Squared Logarithmic Error
BOOKS	0.3455
BABY CARE	0.4857
BEVERAGES	0.4296
DAIRY	0.3935
HOME CARE	0.4530
MEATS	0.4452
PERSONAL CARE	0.4742
PRODUCE	0.4002

RMSLE: 0.5956418338066279

CONCLUSION

In this project, we considered different types of machine-learning approaches for time series forecasting. Sales prediction is rather a regression problem than a time series problem. The use of regression approaches for sales forecasting can often give us better results compared to time series methods. One of the main assumptions of regression methods is that the patterns in the historical data will be repeated in future. Using stacking makes it possible to take into account the differences in the results for multiple models with different sets of parameters and improve accuracy on the validation and on the out-of-sample data sets.

Here, we are also trying to find the difference between predicted values by our model and the observed values. So, we find the root mean squared error in linear regression and support vector regression. RSME for linear regression is 0.3958433835879891, for SVR it is 0.46274117207951615, and the average RSME is 0.3846102422210306.

REFERENCES

[1] Tom M. Mitchell. Machine Learning and Data Mining. COMMUNICATIONS OF THE ACM N. 1999. url: http://www.cs.cmu.edu/~tom/pubs/cacm99_final.pdf.
[2] Matavfall i Sverige – Uppkomst och behandling 2016. Naturvårdsverket. 2018.
url:
https://www.naturvardsverket.se/Documents/publikationer6400/978-91-620-8811-8.pdf?pid=22466.
[3] Coop Värmland. 2020. url:
https://coopvarmland.se/vara-butiker/.
[4] Bohdan M Pavlichenko. "Machine-Learning Models for Sales Time Series Forecasting". In: Data 4.1 (2019), p. 15. doi:10.3390/data4010015.
[5] Kasun Bandara et al. "Sales Demand Forecast in E-commerce Using a Long Short-Term Memory Neural Network Methodology". In: Neural Information Processing Lecture Notes in Computer Science (2019), pp. 462–474. doi: 10.1007/978-3-030-36718-3_39.
[6] Ching-Wu Chu and Guoqiang Peter Zhang. "A comparative study of linear and nonlinear models for aggregate retail sales forecasting". In: International Journal of Production Economics 86.3 (2003), pp. 217–231. doi: 10 . 1016 / s0925 -5273(03)00068-9.
[7] Zeynep Hilal Kilimci et al. "An Improved Demand Forecasting Model Using Deep Learning Approach and Proposed Decision Integration Strategy for Supply Chain". In: Complexity 2019 (2019), pp. 1–15. doi: 10.1155/2019/9067367.
[8] Real Carbonneau, Kevin Laframboise, and Rustam Vahidov. "Application of machine learning techniques for supply chain demand forecasting". In: European Journal of Operational Research 184.3 (2008), pp. 1140–1154. doi: 10.1016/j.ejor.2006.12.004.
[9] Indr˙e Žliobait˙e, Jorn Bakker, and Mykola Pechenizkiy. "Beating the baseline prediction in food sales: How intelligent an intelligent predictor is?" In: Expert Systems with Applications 39.1 (2012), pp. 806–815. doi: 10.1016/j.eswa.2011.07.078.