STORE SALES PREDICTION

XUECHEN ZHANG

Abstract. The selected topic of this project is Kaggle Tabular Playground Series - Jan 2022, which is a time series problem,this topic gives two stores in three The daily sales volume of three products in different countries from 2015 to 2018 requires us to forecast them Sales volume in 2019.

Contents

Date: (None).
2020 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Machine Learning, Data Mining, ...

1. Introduction

Formula for Introduction

GLi:

A good paper introduction is fairly formulaic. If you follow a simple set of rules, you can write a very good introduction. The following outline can be varied. For example, you can use two paragraphs instead of one, or you can place more emphasis on one aspect of the intro than another. But in all cases, all of the points below need to be covered in an introduction, and in most papers, you don't need to cover anything more in an introduction.

Time series (or dynamic series) refers to a series of numbers that are arranged according to the time sequence of occurrence of the same statistical index. The main purpose of time series analysis is to predict the future based on existing historical data. Most of the economic data are given in the form of time series. Depending on the observation time, the time in the time series can be year, quarter, month or any other time form.

Time series is a group of random variables sorted according to time. It is usually the result of observation of a potential process at a given sampling rate in an equal interval of time. Time series data essentially reflects the trend of one or some random variables changing with time. The core of time series prediction method is to mine this rule from data and use it to estimate future data. [1]

Components: long-term trend, seasonal change, cyclical change, irregular change.

- e.g., First Long-term trend (T) is the general trend of change formed by some fundamental factors in a long period of time.
- e.g., Second Seasonal variation (S) is a regular periodic change that occurs with the change of seasons in a year.
- e.g., Third Cyclic variation (C) is a regular variation of wave undulation pattern with a period of several years.
- e.g., Fourth Irregular change (I) is a kind of irregular change, including strict random change and irregular sudden change with great influence. [2]

Motivation

What is the specific problem considered in this paper?

Time series prediction is mainly based on the principle of continuity. The principle of continuity means that the development of objective things has regular continuity, and the development of things is carried out according to its inherent laws. Under certain conditions, the basic development trend of things will continue in the future as long as the conditions on which the law works do not change qualitatively.
Time series prediction is to use statistical techniques and methods to find out the evolution mode from the time series of prediction indicators, establish mathematical models, and make quantitative estimation of the future development trend of prediction indicators.

2. Topic requirements

The topic gives a training set and a data set.Data set: the title gives a 26298 line × 6-column training set, a 6570 row × 5-column test set and one Submit samples. The training set includes sales data of each date-country-store-commodity combination. date From 2015 to 2018, there are three countries, two stores and three commodities. Test set compared with training set Lack of sales volume. The training set header is as follows:

‘ date	country	store	product	num sold
2015/1/1	Finland	KaggleMart	Kaggle Mug	329
2015/1/1	Finland	KaggleMart	Kaggle Hat	520
2015/1/1	Finland	KaggleMart	Kaggle Sticker	146

3. Analysis and Method

Analysis:Three countries, two stores, three commodities can form 18 combinations. Charting these 18 combinations shows that the sales volume of each combination has a similar fluctuation in the year, indicating that the sales volume is seasonal in each month. And the sales volume on weekends is higher than that on weekdays, which means that the sales volume in a week is also seasonal. Generally speaking, the sales volume of each combination is increasing year by year, but the sales volume in Norway is abnormal. The sales volume in Norway in 2015 is higher than that in 2016, indicating that the sales volume is also related to other factors. Method:Because the structure of the data set is not very complex and there are few influencing factors, the linear regression model is selected for this model, and the method of combining time series with linear regression model is used. Add some elements of time series, such as Fourier characteristics, seasonal indicators, and real world GDP data.

QWu: Qiong Wu has worked up to here.

4. Experiment

After the above processing, there are 26 columns of data in total. We passed the processed data set into the linear regression model in sk-learn for training, and finally obtained the model we need. Then use the model to predict and find that there is a huge deviation between the predicted loss value at the end of each year and at the beginning of each year, which means that the sales volume will also be affected by the new year, but I don't know how to improve it for the time being. I hope that the model can be improved when there is time in the future. Finally, submit the forecast results to the official website of Kaggle, ranking 498 with a score of 6.48662.

5. Conclusions

This is a good time series topic, which combines not only several important elements in the time series, but also factors in the real world. Through this topic, I improved my ability of data analysis. Unfortunately, the score of the final result is not high. The main reason is that the consideration is not comprehensive enough. I hope that I can make progress next time when I do similar questions.

References

[1] Gleb Beliakov and Gang Li. Improving the speed and stability of the k-nearest neighbors method. *Pattern Recognition Letters*, 33(10):1296–1301, 2012.

[2] Gleb Beliakov, Simon James, and Gang Li. Learning choquet-integral-based metrics for semisupervised clustering. *Fuzzy Systems, IEEE Transactions on*, 19(3):562–574, 2011.

List of Todos

(A. 1) School of Computer Science,, Nanjing University of Science and Technology, Jiangsu 210000, China
Email address, A. 1: 1131780712@qq.com