Drugs store sales forecast

PROBLEM STATEMENT:

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

Rossmann is challenging you to predict 6 weeks of daily sales for 1,115 stores located across Germany. Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity and motivation. By helping Rossmann create a robust prediction model, you will help store managers stay focused on what's most important to them: their customers and their teams!

---

1) Introduction

Nowadays medical-related sales prediction is of great interest; with reliable sales prediction, medical companies could allocate their resources more wisely and make better profits. You are going to predict the everyday drug sale for each store based on the store, promotion and competitor data. You are insisted to apply different machine learning techniques to tune the model and make predictions on drug sales.

2) Data

We provide you the training data of 1115 Rossmann stores' daily sales dated back to 2013, with 1,017,209 entries in total. The training data includes features of promotion and competitors' information. Since we are not able to access to the real sales amount for testing, you are allowed to use 70% of the given training data as the training set for your model, the rest 30% as test set for cross validation. For now, the cross validation will be your estimated test error.

3) Method and Results

Since the problem involves time series data, we intend you to use time series analysis model to deal with it in the first place. Also you can establish any algorithm model and train it using the data we gave to get the parameters. You can test models with different order numbers, and calculate the test errors for TSA algorithm.

In the time series analysis part, you can predict each store's sale based on just its past data. Now you want to see how stores are different according to their different features. First, you can establish a time-independent model by averaging the daily sales per store and collapsing the time dimension; in this frame you can use some algorithm to select features and then use some other algorithm to fit different features (such as assortment type and competitors) to each store's mean sales.

Finally, you can manage to generalize the time series model or other model to predict several stores at a time, instead of one by one in the first section.

3.1 Time Series Analysis

In order to get a big picture, you can first plot several stores daily sales with respect to time evolution. From the plots, you can recognize that it's a time series data and for a certain store, its daily sales evolve periodically with a period of a year. For example, we can see that there will be sales peaks at certain dates, but in general the sales keep at a constant level.

3.2 For Clear Insight

Since there are so many factors influencing the sales, such as store types, promotions, competitors, and even holidays, you must try to identify the most important features that influence the sales. You can use an algorithm to do that; in order to decrease the amount of data to test the model and the program, we average the daily sales for each store, so the amount of data decrease from a million to a thousand.

4) Conclusion and Prospective

You are insisted to select any algorithm model to predict the sales with small discrepancy to the test data, and you can also use any algorithm to find relations between store mean sales and other features. There are certainly rooms for improvements. We can make further predictions on daily sales. Even though you can find the relations between the features, and make fairly good predictions on average sales for each store. By doing so, you could automate the process of making predictions on daily sales for all the stores in your model.

ALL THE BEST