Drugs Store Sales Forecast via Machine Learning

Hongyu Xiong, Xi Wu, Jingying Yue   Advised by Sam James Corbett-Davies

Problem

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. We join the Kaggle competition to predict the everyday drug sale for each store based on the store, promotion and competitor data. We apply different machine learning techniques to tune the model and make predictions on drug sales using time series analysis.

Dataset

We use the store, promotion and competitor data from the Kaggle competition for Rossmann Store Sales. It contains 1,017,209 daily sales records for 1,115 Rossmann stores. The competitor distance and promotion information of each store is also provided. There's a total of twelve parameters, three of which are excluded from our model because the sale is zero whenever it is a holiday or it is not open. We only analyze Mondays through Saturdays since the stores are closed on Sundays. We take 70% of the training data to train our models and use the rest 30% to estimate the test errors.

Data Preprocessing

In the raw data provided, five parameters have missing data points. We first combine *competitionsincedate* and *competitionsinceyear to a new variable competitionexistedmonth,* scaled by 100. Similarly, we create a new variable *promoexistedmonth* from four promotion variables. We then apply mean imputation for the missing values.

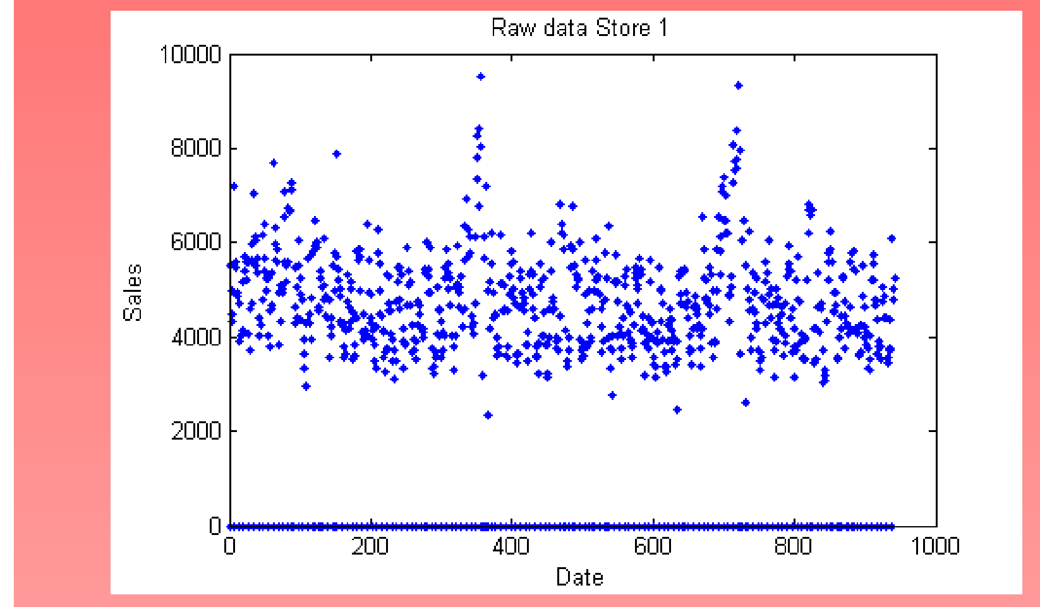
We notice that daily sales evolve as time-series data across the year, using daily sales as time unit would be too noisy, so we decide to construct weekly sales as time unit, for each Rossmann drug store. The value of each time unit is the total of drug sales (open days only) in that week.

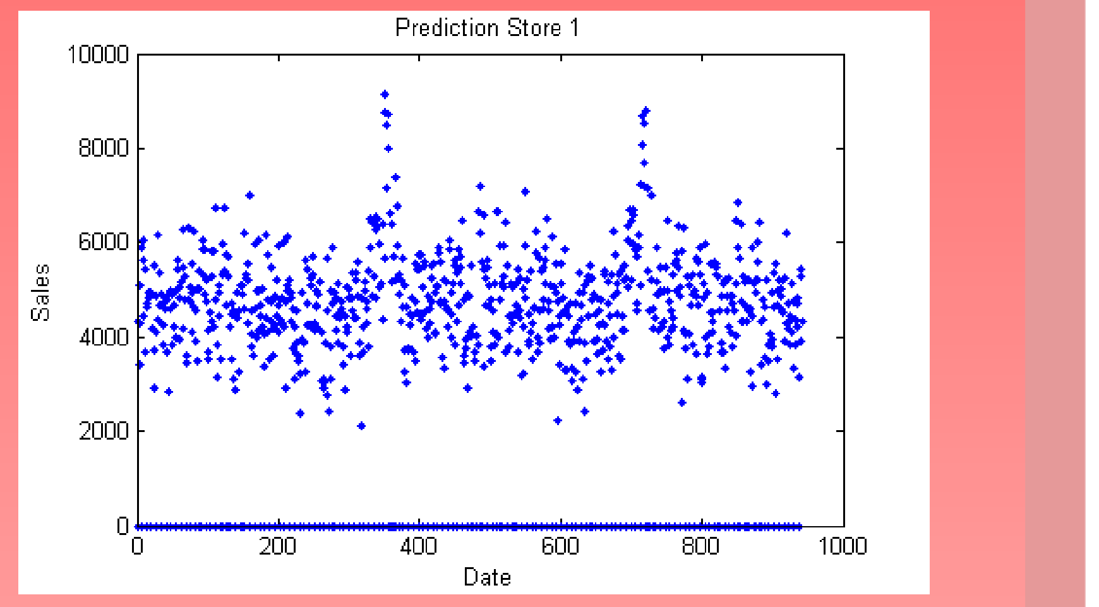
Models

Since the prediction for the daily sale for each store from past data seems to be a time series problem, we manage to solve this problem by building a time series analysis model. The basic method we are using is auto-regression (AR) model:

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

To use this model, we pre-process the data to make weekly sales as a time unit. We compare order p = 1, 2, 3 to see which order number fit best. For each order p, we train parameters φ and ε with 70% of the data and cross validation with the remain 30% to get the test error. Below comparison between prediction (p = 2) and raw data.





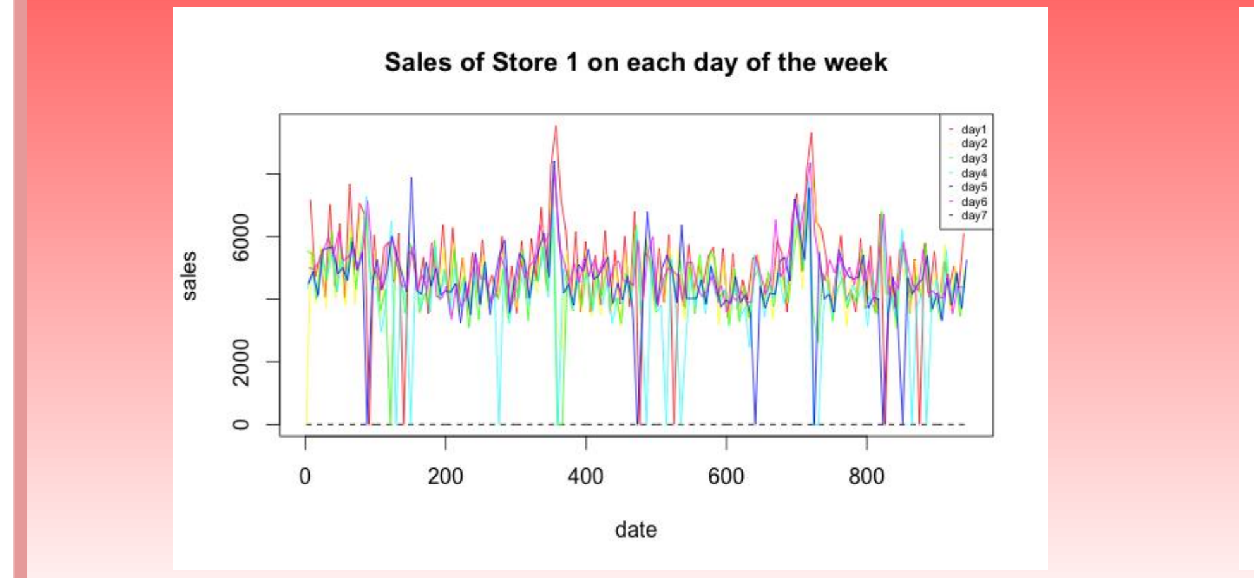
We use Support Vector Regression (SVR) method to find relation between each store's mean sales and different kind of features. SVR is a variation of Support Vector Machine. The essence of SVM is to find a classification boundary with maximum margins to the feature points; by similar token, SVR is to find a regression curve with minimum margins to the feature points. We used the standard liblinear2.1 package in MATLAB to implement SVR.

We compare linear kernel and two other kernels; the test errors are from cross-validation. From the table below we can see that for different types of stores and assortments, the best working kernels are also different.

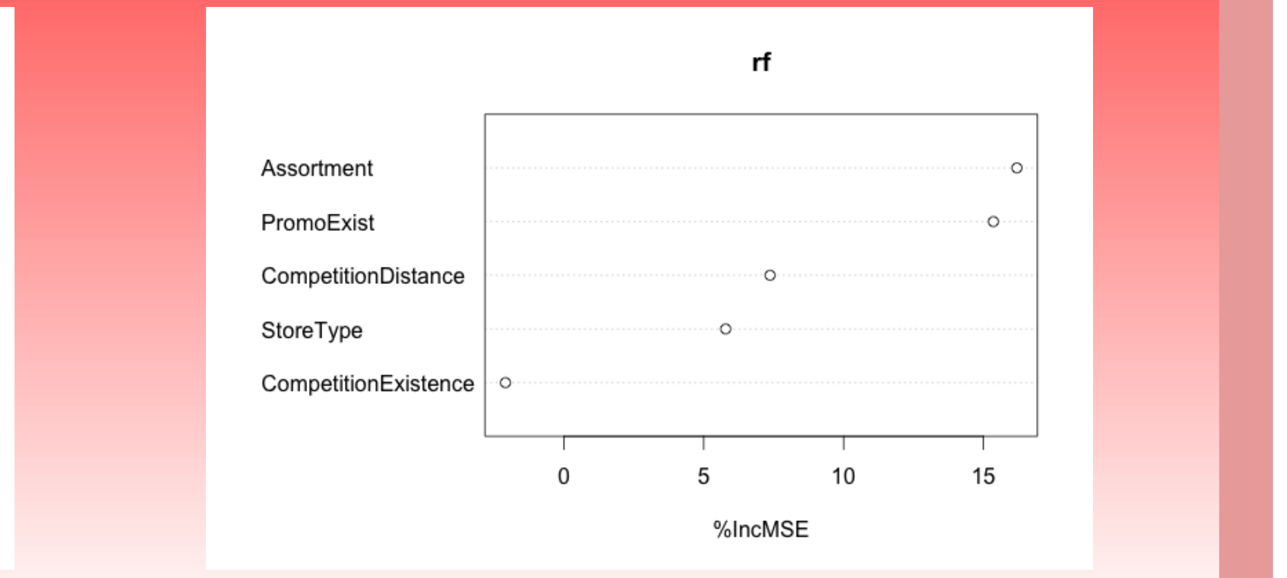
parameters	Cost funtion(a,a)/10^8	Cost function(a,c)/10^8	Cost function(b,a)/10^8	Cost function(b,b)/10^8	Cost function(c,a)/10^8	Cost function(c,c)/10^8	Cost function(d,a)/10^8	Cost function(d,d)/10^8
d,p	2.63	1.60	2.61	0.00412	0.110	0.378	0.263	1.09
d,d^2,p	2.62	1.57	933	0.00497	0.107	0.385	0.305	1.09
d,d^2,d^0.5,p	2.49	1.57	1.15*10^7	0.00938	0.106	0.378	0.228	1.09

Miscellany

G1: Sales of store 1 on each day of the week
G2: Variable importance plot (Random Forest method).



G1



G2

Results

Our model using time series performs better than random forest model. For sales on Tuesday, Wednesday and Thursday, our predictions are within 10% of the real daily sales. However, it gives conservative predictions for Mondays and Fridays, when sales are apparently higher compared to other days in a week.

Future Works

There are certainly rooms for improvements. Firstly, we can make further predictions on daily sales using SVM. Even though we found the relations between the features, and make fairly good predictions on average daily sales for each store, we have not come up with good next steps for applying it to sales in each day. Secondly, we need to automate the process of making predictions on daily sales for all the stores in the time series model.

Reference

[1]. Wensen Dai et al. "A Clustering-based Sales Forecasting Scheme Using Support Vector Regression for Computer Server." Procedia Manufacturing 2 (2015) 82 – 86.
[2]. Marco Hulsmann et al. "General Sales Forecast Models for Automobile Markets and their Analysis." Transactions on Machine Learning and Data Mining Vol. 5, No. 2 (2012) 65-86.