Rossman Store Sales Prediction

Zhuyuan Liu, Tian Yang

CS229-2015 project

Introduction and Motivation

Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity and motivation. In this project, we apply machine learning techniques to a real-world problem of predicting stores sales.

We use popular open source statistical programming language R. We use feature selection, model selection and overfit/underfit concept and methodology to improve our prediction result.

Challenges

- Input features are category features, but output prediction is continuous real number
- Many features have date format. Data processing is hard

Data collection

- Rossmann 1115 Germany stores' sales data from Kaggle.com
- Data statistic:

Number of stores	Number of days	Number of data points
1115	942	1017209

- Feature list related to historical daily sales of each store from 01/01/2013 to 07/31/2015

Field Name	Description
Store	a unique Id for each store: integer number
DayofWeek	the date in a week: 1-7
Date	in format YYYY-MM-DD
Sales	the turnover for any given day: integer number (this is what to be predict)
Customers*	the number of customers on a given day: integer number (this is not a feature. Based on the test data from Kaggle, this feature is not included in test data)
Open	an indicator for whether the store was open: 0 = closed, 1 = open
Promo	indicates whether a store is running a promo on that day: 0 = no promo, 1 = promo
StateHoliday	indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
SchoolHoliday	indicates if the (Store, Date) was affected by the closure of public schools: 1 = school holiday, 0 = not school holiday

- Feature list related to the properties of the stores

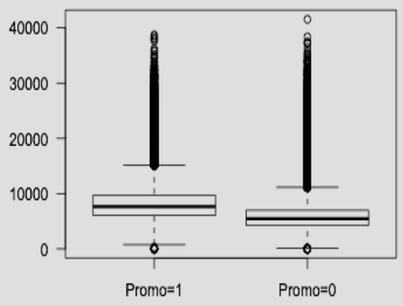
Field Name	Description
Store	a unique Id for each store: integer number
StoreType	differentiates between 4 different store models: a, b, c, d
Assortment	describes an assortment level: a = basic, b = extra, c = extended
CompetitionDistance	distance in meters to the nearest competitor store
CompetitionOpenSinceMonth	gives the approximate year and month of the time the nearest competitor was opened
CompetitionOpenSinceYear	
Promo2	Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
Promo2SinceWeek	describes the year and calendar week when the store started participating in Promo2
Promo2SinceYear	
Promointerval	describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

Feature Analysis

Open

- Sale is 0 when store is closed, so we pre-process the date to removing Open==0 entries

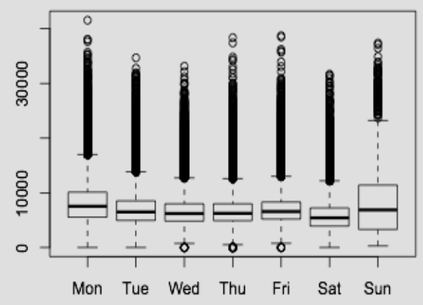
Promo



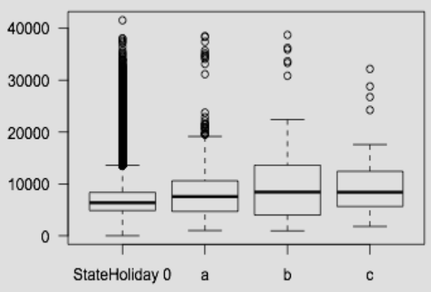
- Sales data distribution on Promo==0/1 shows obvious relation between Promo and Sales

Day of Week

- DayofWeek is represented as numeric number 1-7
- Tue-Fri shows similar distribution

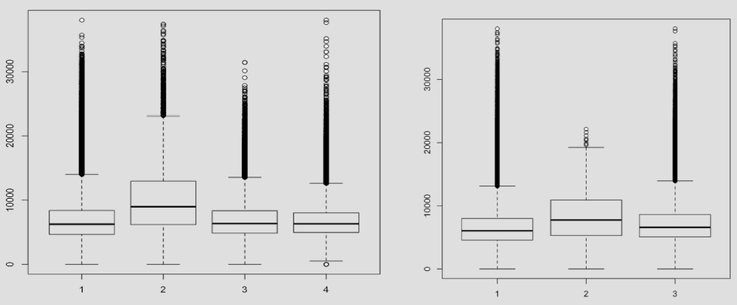


Holiday



- Holiday b and c have little data point and also show similar distribution

Store type and assortment



- Store type and assortment have clear impact on sales

Modeling

Learning

- GLM model with different function was used
- Started with important features and explore more or modified features
- SVM regression was explored using e1071 package
- Modeling across stores vs Modeling with one store

Error Analysis

$$\epsilon = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{|PredSales^{(i)} - Sales^{(i)}|}{Sales^{(i)}}\right)^2}$$

Cross Validation

- N-folder
  - 80% of the training data was used for learning and 20% used for testing
  - Switching testing data to cross validate the model

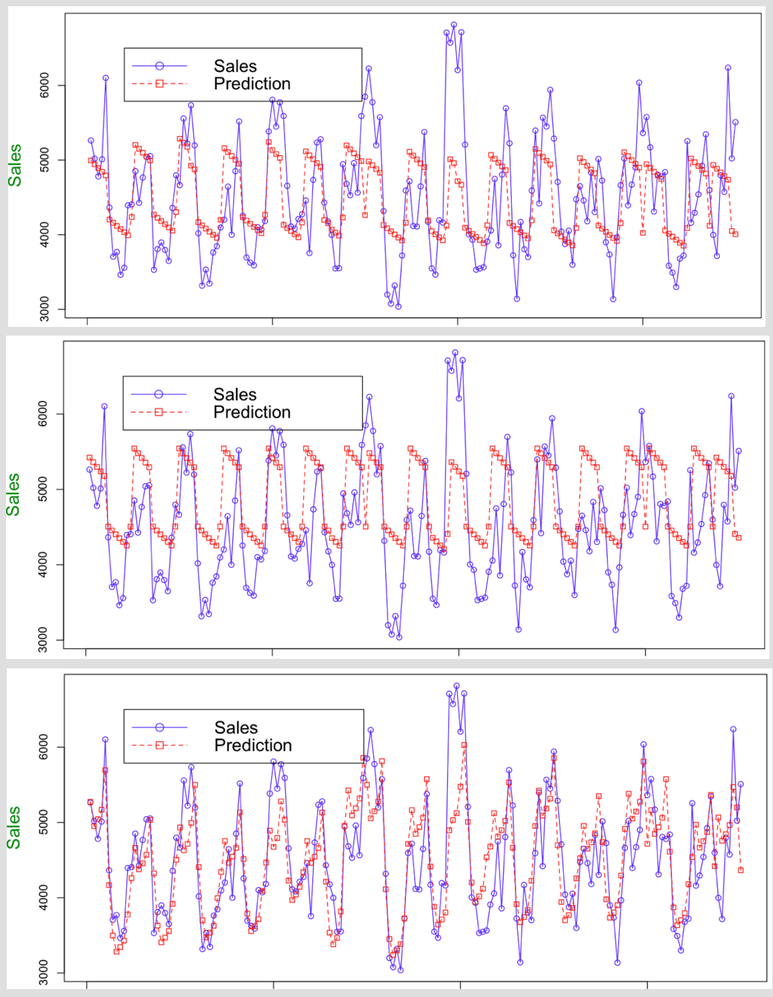
Results and Analysis

Feature Optimization

- Training error reduced when Year and WeekOfYear features are added
- Training model per store has much better result than training with all stores
- Both training and testing error reduced after treating DayofWeek, WeekOfYear as factor

	Training Error	Testing Error
a	0.17891	0.18315
b	0.17747	0.14393
c	0.12670	0.11033

a. Basic Features
b. Adding Year and WeekOfYear
c. Treating DayofWeek, WeekOfYear as factors



Exploration of different GLM models

- We explored different GLM models on one shop
- Results didn't show much different between different models

	Training Error	Testing Error
Poisson	0.1266959	0.110328
Gaussian	0.1291852	0.1137747

Result from SVM regression

- Trained training data using e1071 package
- Current training error we got is 0.2118055

Next step

- We will continue to explore the svm regression
- Will try to include more store info (competition, Promo2, etc) and try to train model on all store data

References

[1]. http://www.kaggle.com/
[2]. http://www.statmethods.net/advstats/glm.html
[3]. https://cran.r-project.org/web/packages/e1071/index.html