Project: Predicting Rossman Store Sales

BANA 6910, Spring 2016

January 20, 2016

This project is based on a Kaggle competition that ran during the fall of 2015. The data you are provided are ***slightly different*** from the dataset provided by Kaggle, so please use the dataset provided for this class. The descriptions below are taken from the Kaggle description.

In the first few weeks of the course we will focus on exploring the understanding the data. Afterwards, we will move to building models and making predictions. Each group will submit several (hopefully improving) predictions, and the groups will compete to create the best model, just like in a Kaggle competition.

Objective: Forecast sales using store, promotion, and competitor data

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.

In this project, Rossmann is challenging you to predict 6 weeks of daily sales for 1,000 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 whats most important to them: their customers and their teams!

The Data

You are provided with historical sales data for 1,000 Rossmann stores. The data can be found on Canvas under rossman_data.zip. The task is to forecast the "Sales" column for the test set. Note that some stores in the dataset were temporarily closed for refurbishment.

Files

train.csv - historical data including Sales test.csv - historical data excluding Sales

store.csv - supplemental information about the stores

Data fields

Most of the fields are self-explanatory. The following are descriptions for those that aren't.

Id - an Id that represents a (Store, Date) duple within the test set

Store - a unique Id for each store

Sales - the turnover for any given day (this is what you are predicting)

Customers - the number of customers on a given day

Open - an indicator for whether the store was open: 0 = closed, 1 = open

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

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

CompetitionOpenSince[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened

Promo - indicates whether a store is running a promo on that day

Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating

Promo2Since[Year/Week] - describes the year and calendar week when the store started participating in Promo2

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

Evaluation

Submissions are evaluated on the Root Mean Square Percentage Error (RMSPE). The RMSPE is calculated as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

where y_i denotes the sales of a single store on a single day and $\hat{y}_i$ denotes the corresponding prediction. Any day and store with 0 sales is ignored in scoring.