Rossmann Store Sales Prediction

Group #8
Group Name: Miners
Group Members: Jeet Takwani and Hriya Maharaja

Abstract:

This project aims to predict sales for a pharmaceutical store in order to help them to be more effective in terms of their staff schedules, promotional activities and managing inventory.

Introduction of the background:

Rossmann operates over 3,000 drug stores in 7 European countries.
Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality.

Problem Definition:

Through their Kaggle competition, Rossmann stores are asking to predict their daily sales for up to six weeks in advance for 1,115 of their stores located across Germany.

Data Description:

Rossmann stores have provided their historic sales data for **1,115 stores** across Germany.

The sales data is provided for each of the store starting from 1st January 2013 up to 31st July 2015 for a total of **1017209 records**.

Each record contains:

- **Store** - a unique Id for each store
- **DayOfWeek** – the day of the week on that date
- **Date** – the date for which the sales is recorded
- **Promo** – whether the store was running a promo on that day
- **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 school

Apart from this, various other attributes about the stores are also provided :

- **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

Given this dataset, we are supposed to predict the value of sales for each of the 1,115 stores for a 6 week duration from 1st August 2015 to 17th September 2015.

Preprocessing:

Observations for the given dataset :

- **Days when stores are closed:**

When a store is closed there will be no sale on that day. And so, that record will have no impact on predicting future sales. We therefore proceeded to remove all records where the store was closed. This leaves only the records for days on which the stores were open.

- **Attribute indicating whether store is open:**

The column indicating whether the store was open on that particular day is now redundant and no longer needed. We therefore removed that column while creating the prediction model.

- **State Holidays:**

We also noticed that the stores are always closed on state holidays, and since we removed entries for days when the stores were closed, keeping track of state holidays is no longer required. We thus removed that attribute from consideration.

- **Number of customers visiting the store at a future date:**

For predicting the sales on any future date, we will not have information regarding the number of customers that will visit the store on that day. The test data set thus does not provide the number of customers that will visit the store on any future date. If we include that attribute in the training set while building the model, it will create an inconsistency in prediction since the test dataset will not have that information. The number of customers visiting the store is not included while constructing the model.

After this we have the following attributes , Store ID, DayOfWeek, Date, Promo, School Holiday upon which to construct our model. And additional store information including store type, assortment, distance to competition and so on as noted above. We now proceed to the first attempt on building a prediction model.

Methods description

Since this is a prediction problem for a matrix dataset, our initial approach is using Linear Regression to predict the values. Since this is Linear Regression we removed the date of sale indicated in the record.

Steps:

1.) Preprocess the data as specified above
2.) Get the training dataset from the preprocessed data
3.) Find an optimal beta for each store
4.) Create a test dataset separate from the training dataset
5.) Measure accuracy of calculated model
6.) Run the model on the test dataset provided by Kaggle

Experiments design and Evaluation

Evaluation Metrics:

The evaluation metric used by Rossmann Stores is :

Root Mean Square Percentage Error (RMSPE)

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

where y^i denotes the sales of a single store on a single day and y hat_i denotes the corresponding prediction. Any day and store with 0 sales is ignored in scoring.

Experiment 1:

We ran the training dataset to build a linear regression model using the stochastic gradient descent approach.

Observation:

Beta exploded and we could not find an optimal beta

Evaluation:

For further optimization, we can study line search approach to calculate an optimal beta for each iteration, to keep the beta from exploding.

Experiment 2:

We normalized the training dataset and ran it again to find beta using stochastic gradient descent.

Observation:

Beta no longer explodes but beta fails to converge to a local optimum. Terminated the process after 10,000 iterations and noted the value of beta.

Evaluation:

As before we can probably look into implementing line search method for finding optimal beta. We ran the test set with the noted beta and submitted results to kaggle. The evaluation score was 0.25222

Experiment 3:

Ran the normalized dataset to find beta using closed form approach.

Observation:

Beta does not explode. Noted the model for each store separately.

Evaluation:

Ran the test set provided by kaggle and submitted the result to kaggle. The evaluation score was 0.25143.Improvement on the previous attempt.

Experiment 4:

In order to gather as much information about a store, we proceeded with the assumption that stores that are of the same type and have the same assortment level, may be correlated. And hence we can use the records of the stores having the same type and assortment, collectively to predict their sales.

Observation:

Upon grouping stores having same type and assortment, we observed that are only 9 groups. Leaving a huge number of stores in each group.

Assort | Type	a	b	c	d
a	381	7	77	128

b	0	9	0	0
c	221	1	71	220

Thus the model for stores belonging to the same group will be similar. This leads to under fitting of the data. Combined with the fact that most attributes are binary, and day of the week has only 7 possible unique values, this leads to similar prediction values for stores belonging to the same group for most test records.

Evaluation:

Generalizing the model for multiple stores does not lead to a good prediction model. Keeping a separate model for each store is a better approach.

Progress Discussion and Conclusion:

For further insights and a different approach, we will now look into time series model for prediction, since this data contains the date for each sales record. We will be able to gain more knowledge like seasonal trends and if one store is a good predictor for other stores. we will be able to make more accurate predictions based on the month or time of year for the date for which we are predicting.

References:

1.) http://www.ccs.neu.edu/home/yzsun/classes/2015Fall_CS6220/Slides/03Matrix_Data_Prediction.pdf
2.) http://cs229.stanford.edu/notes/cs229-notes1.pdf
3.) Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, and Jian Pei. 3rd edition, Morgan Kaufmann, 2011
4.) https://www.kaggle.com/c/rossmann-store-sales