Project Proposal: Rossmann Store Sales Prediction

Hao Qian, Yuning Yang

September 23, 2016

1 Project Description

In this project, we want to predict the sales of stores given historical sales data.

Why is it valuable to predict the sales of stores? As we know, sales revenue is closely related to the operation and profits of stores. It is especially important for retail chains to understand and predict their sales of distinct spots. By modeling and predicting the sales over time, it would help the specific stores to schedule their staff more efficiently. It would also help the company to find out both the most and the least profitable stores and adjust their business strategies accordingly.

Rossmann, the second largest chain drug store company in Germany, held a competition on Kaggle and released some of its stores' sales data. Our project aims at modeling the sales of these Rossmann stores and predict their 6 weeks of sales to support the stores daily operation. From the analysis, we'll also try to learn in what degree various factors contribute to the sales.

2 Data Description

Our dataset could be found at: *https://www.kaggle.com/c/rossmann-store-sales*

The dataset consists of historical sales data for 1115 Rossmann stores. Each record contains information of store, sales, customer, competitors, promos, which are all closely related to sales results. There's also a supplementary dataset of detailed store information which provides store types, recurring promos and so on. Our train set has over 1,000,000 records and test set has around 40,000 records, which we believe, are enough to provide necessary information to train a model of sales that generalizes. We can also submit our prediction results to kaggle to see the performance on given test set and make adjustments to our model if needed.

3 Feasibility

The accessible large dataset and multiple features with a clear description provided by Rossmann would enable us to explore the effective learning models. Besides, as prediction is a quite heated topic in machine learning, there are many mature algorithms that we could refer to and try out.