**Analysis and prediction of Rossman Store Sales**

Applied Data Science – Foundation Project

November 13, 2015

Ding Ma, Yuqiao Cen, Yuxiang Zhang, Zeyu Jiang

CUSP 2015

1. **Introduction**

The Rossmann Sales Prediction was one of the ongoing competitions in Kaggle Community recently. Through the competition, we can get cleaned dataset which provide time and spatial information on the commercial sales of stores. Those store-sales are time series which are in good shape and we can base our further analysis on the statistical learning techniques we have learnt. In our project, we would like to find correlations between features. 1) Does store sales have a strong correlation with customer numbers? 2) Whether a competitor within little distances indicating a higher sales or lower sales? And in which level of significance does it imply? 3) What’s the effect of promotion coupon? Does it 4) Is there a seasonal behavior in the sales data?

With answering these questions, we have a better understanding of the datasets, and would mainly use two approaches to predict store sales, one of statistical learning, like linear regression, and the method using machine learning methods like random forest.

As the team member, I will work with Ding Ma and try to write the code for predicting store sales, using both statistical learning and machine learning methods. Yuqiao Cen will analyze the seasonal and custom effect on the sales and Zeyu Jiang will find out the influence of competitors on the shop.

1. **Data Description**

We got datasets with detailed information about Rossmann stores, and train and test dataset for the history of sales. The basic information provided by the data is the everyday sales history with add-on information about number of customres and their distances to competitors.

1. **Methodology**
   1. Time and Spatial Analysis
      1. Time Series Analysis
      2. Spatial Analysis
   2. Linear Regression
   3. Supervised Machine Learning
   4. Model Compare
2. **Conclusion**

**Reference:**