Project Proposal for Forecasting Sales for Rossmann

Introduction:

Rossmann is Germany's second-largest drug store chain, with over 3,000 stores in 7 countries in Europe. Sales for Rossmann branches fluctuate in all locations over time. Since retail sales forecasts enable store managers to create effective staff schedules that increase productivity and motivation, estimating sales for a near future is be essential. This was a Kaggle competition case, requiring participants to predict 6-week of daily sales for stores located across Germany. We are going to use historical sales record and other features including promotions, competition, school and state holidays, seasonality, and locality to forecast daily sales for 1,115 stores.

Project Scope and Objective:

The main objective of this project is to predict 6 weeks of daily sales for its all 1115 stores located across Germany according to the past 3-year transaction records of each store and the unique characteristics and attributes of stores. Our goal is to build a robust forecasting model, which will be dramatically beneficial for Rossmann to create effective staff schedules, increase the productivity properly, and manage inventory in an optimized level. This project will be executed through a series of steps:

1. Preprocess data and get insight their relationship
2. Clustering stores based on the available information
3. Develop predictive models for each clusters to forecast sales
4. Visualize the forecasting results

Data Description:

The data provided by Rossmann on Kaggle consists of three time-series datasets, including Store Information dataset, Training dataset and Testing dataset. The Store Information dataset contains store-level information for all 1,115 stores, consisting of variables of each store's type, assortment, distance to its competitor, the time when it's competitor opened, whether a store is running a promotion on a specific date, and the promotion interval. The Training datasets containing historical daily sales record for 1,115 stores from 01 Jan 2013 to 01 July 2015. With 1,017,209 observations, it contains stores' ids, sale amount, number of customer on a given day, whether the store was open that day, whether the day was a state holiday and whether the day was a school holiday. And the Testing dataset contains the same information with training except leaving the sales volume to be predicted.

Potential Techniques:

In this project, we would exert different techniques to make predictions from classical statistical models to data mining algorithms, and explore some machine learning techniques. In our project, MSPE of the validation set would be used as the criteria to measure the accuracy of the forecasting model. In the first two stages, we will explore data and extract useful information. Since the number of stores is over 1000, we will group similar stores together firstly, using K means or Hierarchical Clustering. In the third stage, we will focus on building forecasting model in two main approaches. One direction is time series analysis, because the sales of those Rossmann stores are a series of typical time series with obvious trend and seasonality. In this approach, we will use Moving Average, Exponential Smoothing (Holt-Winters Algorithm) and ARIMA to forecast the sales on the time scale. On the other direction, we will adopt the causal-effect method. This approach is to discover the important factors to influence the sales and use them to make a prediction. The potential models are multiple linear regression, CART model, Random Forest, and Neural Network. In the last stage, the results will be visualized and delivered to the stakeholder. Our team will use R as the main tool in the whole project.