Rossmann Sales Data Analysis

Predicting sales performance is one of the key challenges every business face. It is important for firms to predict customer demands to offer the right product at the right time and at the right place. The importance of this issue is underlined by the fact that figuratively a bazillion consulting firms are on the market trying to offer sales forecasting services to businesses of all sizes. Some of these firms rely on advanced data analytics techniques, the kind of which we will be covering in CS-109 classes.

Rossmann is the largest drugstore in Germany. Moreover, it operates over 3,000 drug stores in 7 European countries. In 2015, 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 their first Kaggle competition, Rossmann challenged Kagglers to predict 6 weeks of daily sales for 1,115 stores located across Germany. Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity and motivation.

Milestones:
1. Project Selection: Form teams of 2 or 3 and select a project from the provided list.

2. Literature Study: By running through the data science process you will be able to answer the following research questions necessary for this project:

- To what extend is sales performance influenced by factors like: promotions, competition, school and state holidays, seasonality, and locality.
- What is an appropriate model to predict sales?

Go through the following resources for background on the project and write a 1 page summary for all three of them:

- "Schlecker drugstores to close for good" (Deutsche Welle, June 4th, 2012) "Attempts to rescue the bankrupt drugstore chain Schlecker, once Europe's largest, have failed. The remaining 3200 stores will close, and the last 13,200 employees will lose their jobs.", URL: http://www.dw.com/en/schlecker-drugstores-to-close-for-good/a-15996229

- Brian Knott, Hanbin Liu, Andrew Simpson, “Predicting Sales for Rossmann Drug Stores”, CS229 Final Paper, Stanford University, URL: http://cs229.stanford.edu/proj2015/218_report.pdf
- Rossmann Kaggle competition, URL: https://www.kaggle.com/c/rossmann-store-sales
- A previous project URL: http://zonakostic.github.io/CS_109_EUROPE/

3. Data Cleaning and Loading: The primary sources of data for this project are publicly available at Kaggle’s website (checkout the “Files” section): https://www.kaggle.com/c/rossmann-store-sales/data

Datasets contain historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set. Note that some stores in the dataset were temporarily closed for refurbishment! In order to deal with this problem, check the Filling Gaps in the Training Set article.

4. Exploratory Data Analysis: Exploring the given data, will help in:

- Reviewing the raw data
- Exploring relationships
- Dealing with NA/missing values

(Note: Keep “Test-set” and features presented there in mind!)

In the following URL, you can observe few interesting EDA examples: *https://www.kaggle.com/amhchiu/rossmann-store-sales/more-exploratory-data-analysis*. Also, you might find attractive interactive EDA examples at the previous project’s webpage: http://zonakostic.github.io/CS_109_EUROPE/

5. Propose a Model: Propose methodologies and ideas to be implemented, tested and interpreted for your final project. Pay a specific attention to:

- Correlation between time and sales
- Seasonality
- Autocorrelation

Apply different statistical models and compare them in order to choose the best performing one. You can start with Decision Tree Regression or Linear regression, and then use some of the Ensemble methods. This part of the project is all up to you!

6. Rossmann Store Sales, Winner's Interview: Gert Jacobusse, a professional sales forecast consultant, finished in first place using an ensemble of over 20 XGBoost models, In his blog, Gert shares some of the tricks he's learned for sales forecasting, as well as wisdom on the why and how of using hold out sets when competing.

- Winning Model Documentation describing the solution for the Kaggle competition “Rossmann Store Sales”, pdf URL: goo.gl/Kg7fOZ

Do you understand why this model is so good in predicting Rossmann Stores Sales? Can you do a better job ?