Task and Data Description

The task involves the analysis of historical sales data collected from a large drug store chain in Europe, Rossmann. This mainly concerns data pre-processing for a sales forecasting problem, the completion of some exploratory data analysis (EDA) on the pre-processed datasets to provide insight into the stores and their operations, and then the application of some machine learning techniques in the aim of sales forecasting.

Accurately forecasting sales is one of the most difficult challenges faced by retailers worldwide, as sales are influenced by many factors, such as promotions, competition, holidays, seasonality, and locality. In this project, the overall business objective is to predict 6 weeks of daily sales for 1,115 drug stores located across Germany, as reliable sales forecasts enable store managers to increase the overall productivity and profitability of the retail business and improve their customer satisfaction.

The datasets in use are briefly explained below.

stores.csv

This Excel comma separated values file contains the supplementary information for the 1,115 drug stores.

Column	Description
Store	The anonymised store number.
StoreType	4 different store models: a, b, c, d.
Assortment	An assortment level: a = basic, b = extra, c = extended.
CompetitionDistance	Distance in meters to the nearest competitor store.
CompetitionOpenSinceMonth	The approximate month when the nearest competitor was opened.
CompetitionOpenSinceYear	The approximate year when the nearest competitor was opened.
Promo2	A continuing and consecutive promotion, 0 = the store is not participating, 1 = the store is participating.
Promo2SinceWeek	The week of the year when the store started participating in Promo2.
Promo2SinceYear	The year when the store started participating in Promo2.
PromoInterval	The consecutive intervals in which Promo2 is restarted, naming the which restart occurs.

train.csv

This Excel comma separated values file contains the historical sales data, which covers sales from 01.01.2013 to 31.07.2015.

Column	Description
Store	The anonymised store number.
DayOfWeek	The day of the week: 1 = Monday, 2 = Tuesday…
Date	The given date.
Sales	The turnover on a given day.
Customers	The number of customers on a given day.
Open	An indicator for whether the store is open on a given day: 0 = closed, 1 = open.
Promo	An indicator for whether a store is running a store-specific promo on a given day.
StateHoliday	An indicator of a state holiday: a = public holiday, b = Easter holiday, c = Christmas, 0 = none.
SchoolHoliday	An indicator for whether the store was affected by the closure of public schools.

Test.csv

This Excel comma separated values file is identical to train.csv, except that Sales and Customers are unknown for the period of 01.08.2015 to 17.09.2015.