TSWD Project

PHASE 01: Dataset Identification & Analysis

Project Name: Rossmann Store Sales

Amitha Nayak
Computer Science Engineering
PES University, ECC
Bangalore
Email: nayak.amit.blr@gmail.com

Introduction

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, 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.

A brief overview of the data fields in the dataset:

- Id - an Id that represents a (Store, Date) duple within the test set
- Store - a unique Id for each store
- Sales - the turnover for any given day (this is what you are predicting)
- Customers - the number of customers on a given day
- Open - an indicator for whether the store was open: 0 = closed, 1 = open
- StateHoliday - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- SchoolHoliday - indicates if the (Store, Date) was affected by the closure of public schools
- StoreType - differentiates between 4 different store models: a, b, c, d
- Assortment - describes an assortment level: a = basic, b = extra, c = extended
- CompetitionDistance - distance in meters to the nearest competitor store
- CompetitionOpenSince[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened
- Promo - indicates whether a store is running a promo on that day
- Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
- Promo2Since[Year/Week] - describes the year and calendar week when the store started participating in Promo2

- PromoInterval - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store.

Problem Statement

Predicting sales performance is one of the key challenges every business faces. 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 billion consulting firms are on the market trying to offer sales forecasting services to businesses of all sizes. Many of these firms rely on advanced data analytics techniques to predict sales which aren't necessarily understandable to non-technical users. One of the best methods to bridge the gap between these two roles of a business company is through a business dashboard. A dashboard can simplify complex data sets and provide users with, at a glance awareness of the current performance.

In this project, I will be creating a business dashboard visualizing the results of ML models, which were used to predict the store sales of Rossmann Stores.

Approach

Data Cleaning & Data Transformations

Things done to handle missing data:

1. The dataset was cleaned by getting rid of columns ('Promo2', 'PromoSince', 'PromoInterval') that have more than 40% missing values.
2. Whereas for the columns which had a lesser amount of data missing, the previous values or the next values or the 'mode' of the very same column (depending on the situation / data column) was used as a substitution.
3. The outliers were efficiently weeded out, after visualizing it, through a boxplot.
4. The columns containing nominal categorical values were integer encoded, for the better performance of the ML models.
5. Columns having a 'data-time' datatype, were split into multiple columns, 'Day', 'Month', 'Year'.
6. Additional columns like 'Week' and 'Season' were inferred from the 'date-time' columns.
7. The column 'CompetitionDistance' was log transformed, as this column had huge values (10^10) whereas the rest of our dataset didn't, and this led to highly skewed visualizations.
8. Multiple additional columns were created ('Avg. number of Customers per Week', 'Avg. number of Customers per Month', 'No. of holidays per Week', 'No. of Promo per Week') for the betterment of the ML model predictions.

Using External Data

The dataset lacked in certain cases - it didn't have any information on the store's location, the type of items sold, the type of items the customers bought, what kind of promotional offer was running, or the weather, traffic, trends, etc. This vastly impaired the ML models and the business dashboard visuals. However, a few of the limitations were removed, by making use of external data to find the states in which the store was based, and the weather history of Germany.

Things done to merge data accurately:

1. The primary .csv files (store.csv, train.csv) were merged together using the Store ID
2. Found the state in which the store is based, by checking on which days the store celebrated a 'State Holiday' or a 'School Holiday'.

3. Merged the weather dataset with the primary dataset, based on store location and dates.

Reference

- *Primary Data Source: https://www.kaggle.com/c/rossmann-store-sales*
- *Website containing information on Germany holidays: https://www.timeanddate.com/holidays/germany/2013*
- *Store location dataset: https://www.kaggle.com/c/rossmann-store-sales/discussion/17048*
- *Weather history dataset: https://www.kaggle.com/c/rossmann-store-sales/discussion/17058#97075*

All references used are part of the public domain.