CASE STUDY-RETAIL

STORE SALES FORECASTING

Copyright:

SCO 394, Sector-29
Next to Iffco Metro Station, Gurgaon – 122011
Website: www.analytixlabs.co.in
Email: info@analytixlabsl.co.in

Disclaimer: This material is protected under copyright act AnalytixLabs©, 2011-2018. Unauthorized use and/ or duplication of this material or any part of this material including data, in any form without explicit and written permission from AnalytixLabs is strictly prohibited. Any violation of this copyright will attract legal actions.

Business Context:

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.

Rossmann would like you 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. By helping Rossmann create a robust prediction model, you will help store managers stay focused on what's most important to them: their customers and their teams!

Data Availability & Business Problem:

This dataset is taken from a kaggle competition. You are provided with 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.

Data Files:

train.csv - historical data including Sales

test.csv - historical data excluding Sales

sample_submission.csv - a sample submission file in the correct format

store.csv - supplemental information about the stores

As part of this exercise, you required to build a model to predict sales in various stores and explore and build predictive model.

Most of the fields are self-explanatory. The following are descriptions for those that aren't.

- ✓ 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

Expectations from the Trainees:

1. Understand the data & perform the data preparation before perform all the analysis
2. Provide detailed insights/observations based on the analysis
3. If you build any statistical model,
   a. Understand the output from the software and explain the model fit.
   b. How would you determine what is the best model?
   c. Apply transformations to the given variables and find out the possible best model after transformations
   d. Generate the final equations if applicable
4. What are the key factors that that driving total spend? Do these factors make sense?
5. Data cleaning including missing values, outliers and multi-collinearity. Describe your predictive model. How did you select variables to be included in the model?
6. Apply variable reduction techniques for reduction of variables if applicable.
7. Apply multiple algorithms and compare the results. Choose best algorithm and provide insights on the same.
8. Provide the code with comments and generate outputs(results, plots and insights) in the format of word/pptx/html