Forecasting sales using the store, promotion, and competitor data.

Introduction:

Predicting sales performance is one of the main challenges faced by every business. Sales forecasting which refers to the process of estimating demand of products over specific set of time in future is important, as the demand for products keeps changing from time to time and it is crucial for business firms to predict customer demands to offer the right products at the right time. Sales forecasting also helps to maintain adequate inventory of products and thus, improving their financial performance.

In this project case study, we will use machine learning techniques to predict sales of Rossmann stores using dataset taken from Kaggle.

Problem Statement :

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.

Challenges:

There are too many factors involved in predicting sales, and while forecasting sales, we will have to consider the stores product mix, store locations, store competitors, changes in promotional strategies and calendars, changes in seasonality etc.

A shift in any or all these widely impacts sales forecast, and forecasting sales, just based on prior history will often lead to loss of sales.

Traditional approaches:

Intuitive method :

This method depends on gut feeling of sales representatives and sales managers about the forecast of sales, and this method heavily depends on the past experience and smartness of the sales managers and representatives.

This method is not reliable, and the forecast can vary unrealistically from one manager to another.

Forecasting sales using the store, promotion, and competitor data.

Statistical method:

In this approach, forecasting sales depends on analysts relying on historical data and statistical models to manually organize and compute the data, learn past trends and use them as a baseline to predict future sales.

This is a very time-consuming and expensive process, and prone to errors, as it is very difficult to forecast on huge data using these methods.

Error Evaluation Metric:

Metric to be used for this project is Root Mean Square Percentage Error (**RMSPE**).

Formula for the metric :

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2},$$

Source of this image : https://www.kaggle.com/competitions/rossmann-store-sales/overview/evaluation

where :

- y_i is the actual sales of a single store on a given day
- yhat_i is the predicted sales value
- n is the total number of points

Lower the value of RMSPE error metric, better is the prediction.

Dataset :

Data is taken from Kaggle from the following link:

https://www.kaggle.com/competitions/rossmann-store-sales/data

Dataset properties :

Dataset is approximately of size 40 MB and consists of the following four files :

1) train.csv - historical data including Sales
2) test.csv - historical data excluding Sales

Forecasting sales using the store, promotion, and competitor data.

3) sample_submission.csv - a sample submission file in the correct format
4) store.csv - supplemental information about the stores

Data fields :

File	Variables	Number of variables
train.csv	store, day of week, date, sales, customers, open, promo, state holiday, school holiday	9
test.csv	id, store, dayofweek, date, open, promo, state holiday, school holiday	8
store.csv	store, storetype, assortment, competition distance, competition open since month, promo2, promo2since week, promo2since year, promo interval	10

Our task here is to predict sales column

- 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

Forecasting sales using the store, promotion, and competitor data.

Latency requirements:

As this is a sales forecasting problem, this would be an offline prediction challenge and predictions are not required in real-time serving. But, it would be good if the model will be fast enough to generate the sales forecast within few seconds or minutes.

References:

- https://www.kaggle.com/competitions/rossmann-store-sales/overview
- https://www.kaggle.com/competitions/rossmann-store-sales/data
- https://www.kaggle.com/competitions/rossmann-store-sales/overview/evaluation
- https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/022.pdf
- http://cs229.stanford.edu/proj2015/192_report.pdf