SUPERVISED MACHINE LEARNING

By

Dhanraj
D.N.Raghavendra
Data Science Students, Almabetter

Abstract

This project performs is based upon task given to store managers.Exploratory data analysis on a Rossmann Retail dataset.Here they are given 2 datasets Rossmann stores data and store data By exploratory data analysis we have visualized the data, and removed null values if present, deriving the answers from the dataset we are able to gain insights which might be useful for the management teams of the Rossman store managers mentioned in the dataset.

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.

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.

Introduction

This dataset is a collection of the retail stores around 7 european countries around 3000 stores booking information of two hotels. Our goal is to draw insight by analyzing the data. If any null values remove them we handle outliers also and make a linear regression model. We have done an EDA . We visualize the analysis and look for trends and patterns which will give us some insight from the data. These insights are then good as there are no null values in it.

Challenges Faced

- There are 2 datasets so we have merged these data into one dataset.we have to remove null values,if present w
- We have done year to date,Month to date for representing the sales
- As, to get a good insight of it per year,per weekly sales done.
- The dates were presented in three columns, one each for year, month and date. This was made into one.
- We have done group by with sales and date,sales and year
- If any categorical values are there we turn it into numerical values by using one hot encoding.
- We tried to do machine learning model for a good trend line

The Approach Used to Solve the Problem

- To get meaningful insights from this dataset, an approach of asking exploratory data analysis helped a lot
  .
- The dataset was imported and converted into a Pandas dataframe. After handling null values, the dataset was analyzed using Python and Numpy libraries.
- To get a correct trend line we used lasso and ridge regressions also
- We have used mean squared error and root of mean squared error and r2 score to decrease the error and get a good accuracy
- Graphs were plotted using Matplotlib and Seaborn libraries to visualize the analysis and in turn recognize trends in the data. This helped to get insights which are helpful for the store managers of the said Rossman retail store which could ultimately improve their bottom lines.

Libraries used for analysis

1. Pandas : To load the data into a dataframe object and analyze.
2. Matplotlib : To help visualize the data.
3. Seaborn : For added functionality to matplotlib.
4. Numpy : To use the numpy functions in analysis.
5. Sklearn: To do any machine learning model Scikit learn is very useful to make the correct trend line.

Dataset

This dataset consists of the prediction of the Rossman store prediction for upcoming Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. So that the between the years of 2013 and the 2015 weather, the sales are going to increase or decrease at Rossmann stores, namely ‘Resort Hotel’ and ‘City Hotel’. Since this is real data, all data elements pertaining to stores. 1017209 rows and 18 columns of Rossman data. The definitions of columns and the explanation of the categories contained :

Most of the fields are self-explanatory.

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

---

Dataset preparation before analysis

There are 2 datasets one containing Rossman store and another one is the store set. We merged these both and made it into one dataset that is rossman data no null values are there.we group by the sale and week And sales and years we have done linear regression didn't show good accuracy so we have done the lasso and ridge regression using the grid search CV .The dates in the dataset were spread throughout 3 columns, year , month ,date

Conclusion ;Rossman retail store is one of the famous retail store 3000 drug store in European countries here, we have to predict the sales and competition between the store In this data we have 2 dataset files they are Rossman stores data and a sales data

of Rossmann here, to predict we have to do firstly the Exploratory data analysis and visualizing the 2 datas so, we have to merge up both datasets so,that we have 1017209 rows and 9 columns in rossman dataset and in 1115 rows and 10 columns we have to merge these 2 datasets and we have to see are there any null values. we have to import the libraries which are required.we have to find the dataset is open or closed we can use one hot encoding such that the if store is open = 1,store is closed = 0 and we can drop columns which are open in binary pathWe can see that in "Promo2SinceYear", "Promo2SinceWeek", "PromoInterval" nearly 50 % of values are null. There is no point in keeping those so we are dropping these columns We will replace null values with median of columns "CompetitionDistance", "CompetitionOpenSinceMonth", "CompetitionOpenSinceYear" then next we have to discrete variables to continuous variable and while visualizing we have seen they are positively skewed,we have to drop the duplicate values.we should be aware of **outliers and handle them ,*we can groupby sales and day so, that we can find on which day it has highest sales and lowest sales and same as monthly and sales also we have to group by year and sales. Now, competition between two stores distance the visualization shows the positively skewed curve and we have done training and testing for linear regression and we use ***mean squared error,root of mean squared error,mean absolute error and r2 score** and we get a linear regression model same as regularized linear regression and lasso regression and Ridge regression even elastic net all of the models showed us good accuracy.