Retail Sales Prediction

BAISHNAVEE MAHATO

Data science trainee,
AlmaBetter, Bangalore

1. INTRODUCTION:

The Rossmann Store Sales dataset is a public dataset, which contains daily historical sales data for 3 Rossmann stores from the 1st January 2013 till the 31st July 2015.

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.

The product range includes up to 21,700 items and can vary depending on the size of the shop and the location. In addition to drugstore goods with a focus on skin, hair, body, baby and health, Rossmann also offers promotional items ("World of Ideas"), pet food, a photo service and a wide range of natural foods and wines. There is also a perfume range with around 200 commercial brands. Rossmann has 29 private brands with 4600 products (as of 2019). In 1997, the first own brands Babydream, Facelle, Sunozon and Winston were introduced. The best-known Rossmann brands are Isana (skin, hair and body care), Alterra (natural cosmetics), domol (cleaning and laundry detergents) alouette (paper tissues etc).

The company logo consists of a red name and the symbol of a centaur integrated in the letter O: a mythical creature made of horse and man from Greek mythology, which symbolically stands for "Rossmann" (English: "Horse man"). The company's own brands have a small centaur symbol above the name.

Rossmann branch in Albania
Since 2018, Rossmann has been publishing a sustainability report for the development of corporate climate protection activities.

In 2021, sales increased by 8.1 percent to 11.1 billion euros. There are a total of 4,361 Rossmann branches, 2,231 of which are in Germany. The current number of foreign branches is: Poland (1580), Hungary (more than 220), Czech Republic (more than 150), Turkey (more than 120), Albania (15), Kosovo (6) and Spain (5).

2. OVERVIEW OF BUSINESS:

To understand the objective of the project, it is necessary to understand the business.
The company was founded in 1972 by Dirk Rossmann with its headquarters in Burgwedel near Hanover in Germany. The Rossmann family owns 60% of the company. The Hong Kong-based A.S. Watson Group owns 40%, which was taken over from the Dutch Kruidvat in 2004.

The product range includes up to 21,700 items and can vary depending on the size of the shop and the location.

3. 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.

I have been 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.

3. Data Description:

Rossmann Stores Data.csv - historical data including Sales

store.csv - supplemental information about the stores

Data fields

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
- State Holiday - 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
- School Holiday - indicates if the (Store, Date) was affected by the closure of public schools
- Store Type - differentiates between 4 different store models: a, b, c, d
- Assortment - describes an assortment level: a = basic, b = extra, c = extended
- Competition Distance - distance in meters to the nearest competitor store

- Competition Open Since[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
- Promo Interval - 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.

4. Null Values Operation:

Here only df2 has missing values so I filled them with appropriate values.

- Filled Promo2SinceWeek, Promo2SinceYear, PromoInterval with 0.
- Filled CompetitionDistance with mean distance.
- Filled CompetitionOpenSinceMonth, CompetitionOpenSinceYear with most occurring month and year respectively.

5. PROCEDURE:

(EDA)EXPLORATORY DATA ANALYSIS:

EDA is the process of trying to understand data in the ways possible in order to derive insights from it.
Used exploratory data analysis to understand important factors or characteristics such as How many open or closed stores are in the data, How many Promo is in the data, Which day has the Promotion, Does Promotion affect the Sales Price, Which store has the more Customers, State Holidays and Sales, Sales Vs State Holidays, Sales Vs School Holidays, and also checked for missing or null values and outliers.
Exploratory data analysis is the process of looking at available data sets to identify patterns and anomalies, test hypotheses, and validate assumptions using statistical means.
Using Python in exploratory data analysis processes and visual comparisons between variables is easy to understand and insightful.

Inferences from EDA:

- All of the weekdays have a promotion and the weekend has no promotion.
- The sales price is highest when there is a promotion.

- All Type ‘b’ stores have comparatively higher sales and it is mostly constant with peaks appearing on weekends.
- Majority of Stores remain closed on state holidays.

- School holidays increase the sales price and customers proportionally.

Feature Engineering:

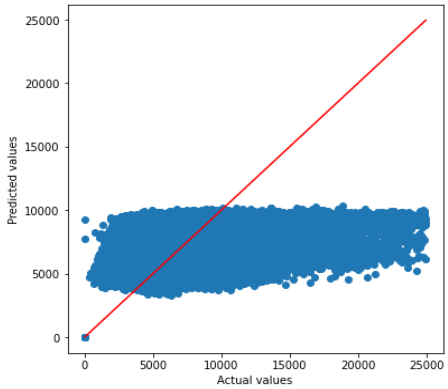
Feature engineering is the pre-processing step of machine learning, which extracts features from raw data. It helps to represent an underlying problem to predictive models in a better way, which as a result, improves the accuracy of the model for unseen data. The predictive model contains predictor variables and an outcome variable, and while the feature engineering process selects the most useful predictor variables for the model.

Here I have done Converting Categorical Variables to Numeric, Looking at the scenario where the Stores are open and yet there is no sales on that day.

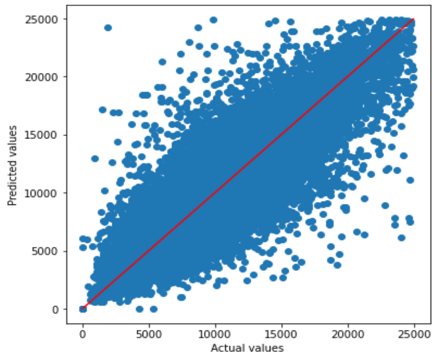
Building a Regression Model:

Created Regression Model is used to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable (x) is associated with a value of the dependent variable (y).Here we want our ML model to predict sales only when they are open and we know that there will be no sales if the store is closed.
Splitted the data into train and test.

From the analysis I found **Linear regression** model is performing badly as it is not making any predictions more than 10000 even for 25000 sales.



The decision tree regressor performed well compared to Linear Regression.



Random Forest regressor had the lowest error as compared to other models, which means it is better at predicting sales than other models.

Conclusion:

Random Forest Regressor is the best for this Dataset Problem.

REFERENCES:

https://en.wikipedia.org/wiki/Rossmann_(company)

https://dbpedia.org/page/Rossmann_(company)

https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10