National College of Ireland

National College of Ireland
Project Submission Sheet – 2020/2021

Student Name:	KARAN VEER SINGH, LAIBA REHMAN, PRIYANKA --
Student ID:	x20146248, x20144032, x20192037
Programme:	MSc in Data Analytics **Year:** 2021
Module:	Domain Application of Predictive Analytics
Lecturer:	Vikas Sahni
Submission Due Date:	02/07/2021
Project Title:	Sales Forecasting for Rossmann Stores Using Machine Learning.
Word Count:	2879

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	KARAN VEER SINGH, LAIBA REHMAN, PRIYANKA --
Date:	02/07/2021

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sales Forecasting for Rossmann Stores Using Machine Learning.

Karan veer Singh
MSc In Data Analytic
National College of Ireland
Dubin, Ireland
x20146248@student.ncirl.ie

Laiba Rehman
MSc In Data Analytic
National College of Ireland
Dubin, Ireland
x20144032@student.ncirl.ie

Priyanka –
MSc In Data Analytic
National College of Ireland
Dubin, Ireland
x20192037@student.ncirl.ie

***Abstract*—Rossmann functions around the world with more than 3,000 medical stores in 7 countries of Europe. The sales need to be monitored for future requirement predictions of the drugs by the managers in the store. Like other business problems it also focuses on the market requirement, Store sales are influenced by many factors, such as marketing strategies, area, timings, holidays, season, location, competitor's data, and promo codes. Extracting the features responsible for contribution in sales, transforming them as per the required objective of the researchers and sales managers, loading them to get the desired result using various models of machine learning such as linear regression, stochastic gradient regressor, random forest, and decision tree.**

***Index Terms*—**Predictive Analysis, Sales, Visualization, Linear regression, random forest, assortment.

I. INTRODUCTION

A. Background of the domain

Retail Pharmaceutical is a major pillar in the healthcare sector for supplying medical equipment and drugs. It has various channels that involve production, retailers, wholesalers, branding, and licensing of the product and customer support. Unlike a regular pharmacy retail pharmacy deals with drugs that do not require a prescription. The authorization of these drug stores is different than other stores. Predicting the proper demand at each retail terminal is essential for performance because it assists product record and catalog management. Reduces overstocking and understocking of inventory across stores at each location, reducing failures, increasing client satisfaction, and sales. This problem of demand forecasting is complex as high stakes are involved in external and internal factors to forecast. It has become a critical problem for any retailer to solve [1]. Further, external factors, including rival competition, climate, periodic generated trends [2], and internal factors sales, store type, assortment, promotion, planning are just a few examples [3]. Retail pharmacies can be standalone or a chain of stores in the same area or different. It is also referred to as Community pharmacy.

B. An Overview of the Data

This dataset is available publicly on Kaggle website [4] Some of the attributes are categorical and continuous in the dataset shown below:

Columns	Datatype	Descriptions
Id	Integer	An Id that represents a (Store, Date) duple within the test set
Store	Integer	A unique Id for each store
Sales	Integer	The turnover on a given day (our target variable)
Customers	Integer	The number of customers on a given day
Open	Integer	open: 0 = the store is closed , 1 = the store is open
StateHoliday	Integer	Indicates a state holiday. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
SchoolHoliday	String	Store on this Date was affected or not by the closure of public schools
StoreType	String	4 different stores:a,b,c,d
Assortment	integer	Assortment level: a = basic, b = extra, c = extended
CompetitionDistance	Integer	Distance in meters to the nearest competitor store
CompetitionOpenSince[MonthYear]	Integer	gives the approximate year and month of the time the nearest competitor was opened
Promo	Integer	Promo or not on that day
Promo2	Integer	Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
Promo2Since[YearWeek]	Integer	describes the year and calendar week when the store started participating in Promo2
PromoInterval	String	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

Fig. 1. Data set description

Fig. 1 displays the detailed information of columns of the dataset with their datatypes. The column 'sales' is an integer type and depicts 'Amount of revenue generated per day'. It is also considered as the independent attribute in this project.

C. Scope

Predicting Sales improves the efficiency among stores using data related to customers. Data can be mined to highlight trends and improve planning, in addition to merely estimating the anticipated sales. Advertisement, recommendation, demand forecasting, consumer-based pricing, holiday/extended sale planning, and product classification are just a few examples. The internal factors such as sales, store type, assortment can be useful for the prediction of proper demand at each retail outlet and sales. However, the attributes like promotion code are not included in the dataset, they are not that important for the prediction as they are duplication of promo interval and reduces the evaluation result of prediction. The external factors that are included in the dataset such as CompetitionDistance (Distance to the nearest adversary store) and CompetitionOpenSince (estimated time the nearest adversary was opened) are consolidated with internal attributes be capable of increasing

the accuracy of the forecast. Using these factors all collectively, it is probable to predict sales for an outlet.

II. LITERATURE REVIEW

Predicting the requirement of the drugs and medical aids plays a vital role in the pharma industry. It involves an in-depth analysis of the existing data available in the store. Lack of historic data availability makes it compulsory for the researcher to use the existing machine learning models that are built on large dataset analysis for predicting the need of such store [5]. Sales prediction in retail pharma uses Support Vector Regression machine learning model but adding the marketing factors to it such as promotion, discounts, loyalty card for customers improve the model accuracy (91 percent) [6]. This paper illustrates that the XGBoost algorithm [7] gave the best outcome in predicting the target variable "sales". It suggested that shops enhance their productiveness and increase income by the analysis of the information, and they also highlighted some key areas which have tendencies to efficiently increase sales such as holiday sales, advertisement, consumer-based totally pricing, and product classification. In this paper [8] how stacking approach can be utilized to enhance the performance of machine learning models in the data which is highly imbalanced has been discussed. The main aim of this approach is that the outcomes from previous level of the machine learning models are utilized as input variable in the other models. They concluded that utilizing multi-level stacking benefits in getting the better results in predicting accuracy when compared with single models.

III. GOAL

The project's principal goal is to determine an algorithm that can accurately forecast sales of medicines 6 weeks in advance for Rossmann Store using both fundamental and extraneous knowledge to help managers maintain inventory for future sales using the machine learning algorithms with highest accuracies to achieve the objective of sales prediction. The secondary goal for the project is to apply exploratory data analysis (EDA) tools, visualizations, and time series analysis for checking the seasonality and trend in patterns of sales to perceive enterprise insights. These new insights will ultimately help the Rossmann Store increase sales and income.

IV. ETHICAL CONCERNS

Experimental techniques are derived by the systematic application data results by making use of the observed evidence. The outcomes are supported by data that is retrieved from various sources. It is difficult to create datasets from scratch in which sensitive information such as real-time customer data are publicly accessible and reuse is allowed. Numerous ethical concerns are involved when we work on such datasets. It becomes important to think about each ethical concern and accordingly proceed further. Pharmacist business ethics are mainly defined by the principles and values of the (retailer).

A. Identify the stakeholders.

To study about the research on possible benefits and damages, the identification of primary, secondary and stakeholders is an essential task. The people who are directly associated with the data are the primary stakeholders such as business owners; The intermediaries are the people who are service provides; comes in the category of secondary stakeholders and the people who are directly involved in the data collection and research come in the category of key stakeholders. Shareholders are made aware by providing optimum information about the company's financial position and the business that will respond to the needs of the shareholder's effectively will be successful.

B. Customer Trust and Credibility

Personal development is vital for any interaction in sales whether it is for the buyer or for the business to build a decent or mutual understanding between the two parties. Moreover, endeavoring to acquire the sales using all means will prompt untrustworthy activities for arriving at the daily limit. When the business representatives engage with the customers, they should first target their mutual confidence and then focus on the sales. The patient-pharmacist relationship is efficient only when the best medical-care services are provided, which is given when the prescription accuracy and the documentation of every employee working in the pharmacy is confirmed.

C. Honesty

The pharmacist's main duty is to take care that the patient is not harmed in any way. The drugs dispensed in the pharmacy and drugs quality are the responsibility of the pharmacists. It is highly important that the pharmacist attempt to provide efficient and sufficient information about the drugs to the patients who are seeking information and make sure that they know about the drugs used and how it can help cure them regardless of if it is benefitting them financially or not. If the pharmacist tries to sell the drugs with false commitments or just oversell it can even lead to the death of the patient due to which the license of the company might also get ceased.

D. Safeguard Confidential Information

The confidential data has been erased as the data made accessible for the investigation safeguard policies and post screening are thought of. It will become an issue of credibility if the information of the patient is leaked. Then again, concerning drugs, share the information about the drug usage and how it helps in curing the patient, but assuring unrealistic results or over-selling the drugs can lead to severe negative outcomes.

E. Justice

The objective of this research paper is not to aim for any demographic, social or cultural group's advantage or disadvantage therefore not causing any injustice to these groups.

F. Public interest

The public domain from where this dataset has been taken from to do analysis has already taken all the necessary precautions which needed to be taken before publishing the data therefore this demonstrates there are no ethical issues in using this data and no violation of the rules have been practiced.

V. BUSINESS STRATEGIES

Old and present data can be used to obtain business insights using appropriate high-level technical instruments. The outcomes of predictive analysis can be used to reveal a variety of business approaches to boost the customer footfall and sales for the Rossmann medical stores, The hidden aspects can be uncovered using the historic data. Various strategies, tools, and algorithms are designed for predicting as given below:

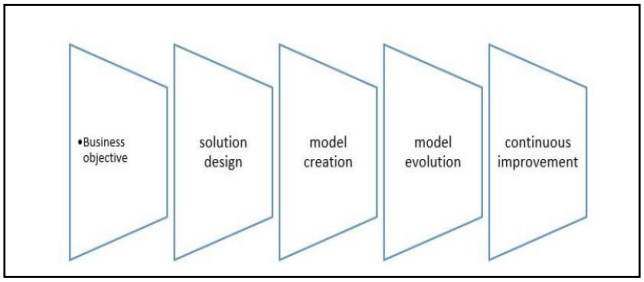

Fig. 2. Business Strategy for retail pharma

A. Product assortment

This strategy work by broadening the product lines from basic to the extended level. The products get to mix with the low-sale product in this plan.

B. Sunday Sales

The weekdays are the busiest days of the week and customers prefer to relax and enjoy leisure than buying products as they can get them on any other day while they are traveling for work. So, the drop in sales dropped on the 7th day of the week which is Sunday. To minimize that offers can be generated on bulk purchases on Sundays.

C. Online delivery

The state holidays are the time when the sales dropped in the drug retails to overcome it, we can strategize our moves by not charging any delivery fee to customers for that day with a predefined minimum purchase amount to avail the fee waiver.

D. Referral and loyalty points for old customers

A particular month can be entitled to loyal customers for using promo codes during purchase. It will be during the month when the sales are going to drop, and the impact will be fewer the customers will purchase in bulk.

E. Online surveys for customer feedback

Customer product and process satisfaction is a must to enhance the interaction with customer digital surveys can play a major role. Analysis of these forums will improve customer support and product selection for the future. Adding assured reward points for every complete forum will be a plus point of authentic information exchange as it will motivate the customer to take out time from their schedule.

VI. PRELIMINARY VISUALISATIONS

Data visualization is the introductory step to obtain the insights of data. It assists to get a precise knowledge of data promptly. Simple visualizations like bar graphs, distribution plots, line graphs, scatter plots, and so on can be used to find developing trends, relationships between dependent and independents values of the dataset, as seen in the plots below.

A. Customer trend with respect to day of week and its impact on sales

The combination of bar plot which represents the average no. of customers on the day of the week and line graph which shows the average sales is shown in fig 3. For the day of the week axis 1 denotes Monday, 2 denotes Tuesday, and so on till 7 which denotes Sunday. It can be observed from the plot that the customer footfall is more on Monday, Tuesday, and Friday in comparison to Wednesday, Thursday, Saturday, and Sunday, because most stores are not open on Sunday, so the sales volume is so less. The average sales are directly correlated to the customers, as the average no. of customers shrinks in Sunday sales gets affected as well.

Fig. 3. Customer trend with respect today of week and its impact on sales

B. Sales with respect to school holidays.

Fig. 4 visualizes the sale concerning the school holidays. As we can observe the sale is more when there is a public holiday in comparison to the non-school holiday. The total sum of sales for the year 2013,2014, and 2015 in the case of school holidays is almost 80 percent which is more the 3/4th of total sales.

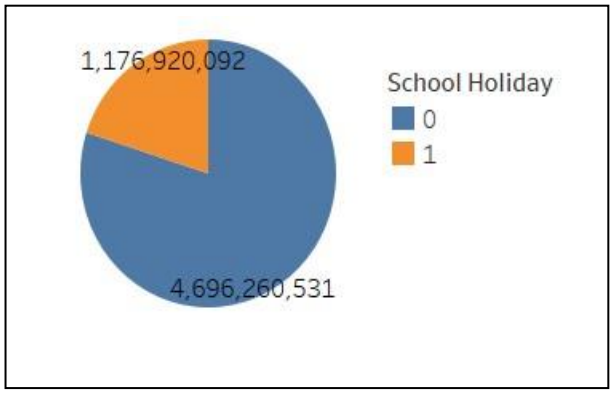

Fig. 4. Sales with respect to school holidays

C. Store type and promo interval impact on sales

The sales of any retail business are highly dependent on the promotions that are offered by them to attract customers. Figure 5 illustrates the store-wise trend of the sales during the different promotional intervals. There are four promotion intervals, divided across the whole year. The sales of the store a, c, and d have been influenced by the promotions in the same manner. But for the b-type stores, there is an exceptional increase in sales for promotional intervals January, April, July, October, and for no promotion period.

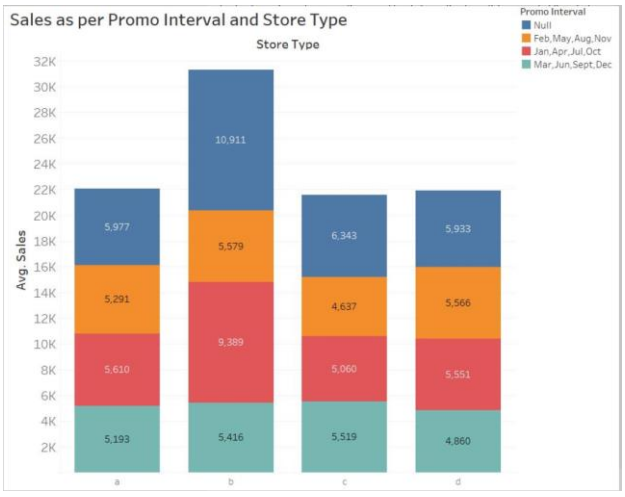

Fig. 5. Store type and promo interval impact on sales

D. Average sales with respect to store type

Figure 6 shows the average sales concerning the store type. It is observed from the plot for store type 'B' average sales are more and for the rest store type average sales are almost constant. The difference between sales of 'B' store type is almost 4k. It means Store type B is the best performing store.

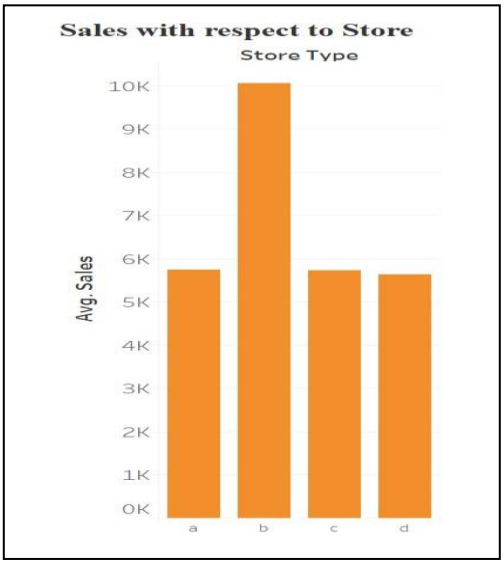

Fig. 6. Average sales with respect to store type

E. Relationship between school holidays and sales

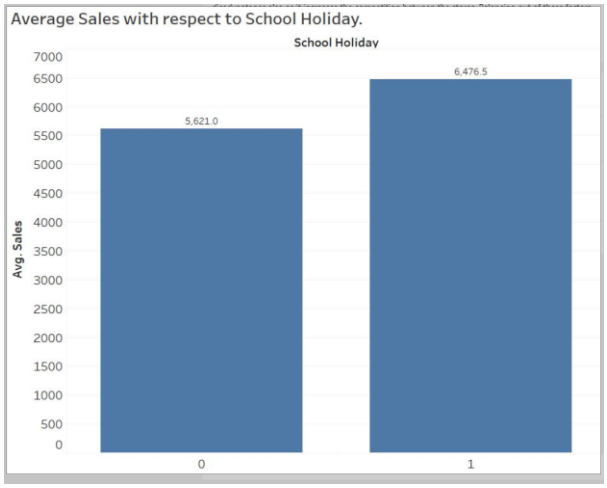

Fig. 7. Relationship between school holidays and sales

On school holidays routine checkups and other healthy lifestyle changes are done for children's and that increases the sales of the drugs in an area for that period. Average sales increases during the school break from 5621 to 6476.5 depicted in figure 7.

F. Relationship between sales and assortment

The assortment is used to increase the product size as it focuses on uplifting the low-sale value product with the high sale value. most retails as well focus on this strategy that just changes the replacement of the products over the counter or add an offer to its combination. From figure 8 it can be observed average sale for assortment type B is more. It indicates assortment B is attracting more customers. Assortment A and C have equal no. of average sales.

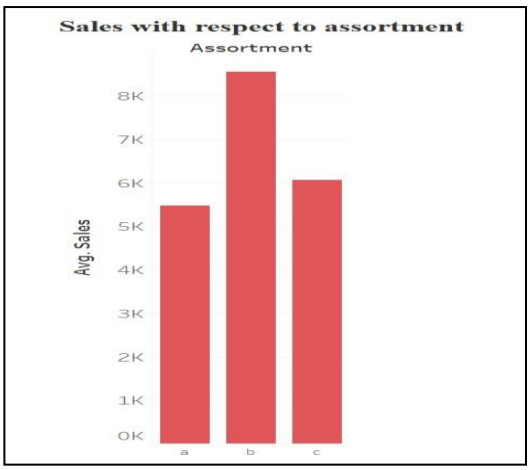

Fig. 8. Relationship between sales and assortment

G. Impact of competition distance on sales

Competitor store distance is an important factor for sales. However, Figure 9 depicts the sales and distance in meters to the nearest competitor store are not correlated. Having proximity to the adversary store is advantageous as customers will have many alternatives for products and disadvantages also as it increases the competition between the stores. Balancing out of these factors causes no impact of competitor store distance to sales.

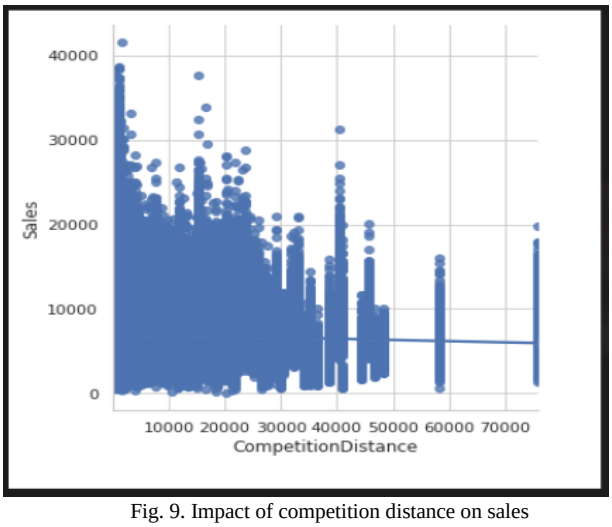

Fig. 9. Impact of competition distance on sales

VII. APPLICABLE TECHNIQUES

A. ***Linear Regression***.

In Linear Regression the outcomes are predicted based on the known values and the outcomes generated from those values.

B. ***SGD Regressor.***

Stochastic Gradient Regressor is easy in implementation, yet it is an efficient optimization approach which is applied to find the parameters of functions for cost function reduction.

C. ***Random Forest Regressor.***

Random Forest is a supervised learning approach for regression that uses the ensemble learning method. The ensemble learning method consolidates predictions from several machine learning approaches to get better evaluation results (accurate) for forecasts than a single model.

D. ***Decision Tree Regression.***

Decision Tree are also called as CART (Classification and Regression Trees). Classification tree is used when the target variable is a categorical or quantitative discrete variable but as in our dataset the response variable (Sales) is a continuous variable thus we will be using Regression tree model.

REFERENCES

[1] S. Kohli, G. T. Godwin, and S. Urolagin, "Sales prediction using linear and knn regression," in *Advances in Machine Learning and Computational Intelligence*. Springer, 2021, pp. 321–329.

[2] A. G. Parsons, "The association between daily weather and daily shopping patterns," *Australasian Marketing Journal (AMJ)*, vol. 9, no. 2, pp. 78–84, 2001.

[3] A. Jain, M. N. Menon, and S. Chandra, "Sales forecasting for retail chains," *San Diego, California: UC San Diego Jacobs School of Engineering*, 2015.

[4] https://www.kaggle.com/pratyushakar/rossmann-store-sales, 2021.

[5] X. Zhu, A. Ninh, H. Zhao, and Z. Liu, "Demand forecasting with supplychain information and machine learning: Evidence in the pharmaceutical industry," *Production and Operations Management*, 2021.

[6] Y. Liu, X. Yang, C. Zhu, and J. Meng, "Drugs sale forecasting based on svr integrated promotion factors," in *Journal of Physics: Conference Series*, vol. 1910, no. 1. IOP Publishing, 2021, p. 012056.

[7] B. S. S. Ramya and K. Vedavathi, "An advanced sales forecasting using machine learning algorithm."

[8] B. Pavlyshenko, "Using stacking approaches for machine learning models," in *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, 2018, pp. 255–258.