**Capstone Project 2**

**Retail Sales Prediction By**

**Contributor Role**

**Suraj Kad (**[**suraj.kad.90@gmail.com**](mailto:suraj.kad.90@gmail.com)**)**

**Shubham Sawant (**[**shubhamsawant248@gmail.com**](mailto:shubhamsawant248@gmail.com)**)**

**Tanvir Patel (**[**tanviripatel1998@gmail.com**](mailto:tanviripatel1998@gmail.com)

**DATA SCIENCE TRAINEES,**

**TEAM ALMABETTER**



**Abstract:**

Sales forecasting refers to the process of estimating demand for or sales of a particular product over a specific period of time. Businesses use sales forecasts to determine what revenue they will be generating in a particular timespan to empower themselves with powerful and strategic business plans. Important decisions such as budgets, hiring, incentives, goals, acquisitions and various other growth plans are affected by the revenue the company is going to make in the coming months and for these plans to be as effective as they are planned to be it is important for these forecasts to also be as good. The sales forecasts are also different from the sales-goals a company has. Sales-goals is what a company wants to happen to execute their future plans for the business. On the other hand, sales forecasts are what is going to happen on the basis of past records, data, trends and various improvement measures taken. The work here predicts the sales for a drug store chain in the European market for a time period of six weeks and compares the results of machine learning algorithms. Keywords: EDA, Correlation, Decision Tree Random Forest, Regression, Forecasting

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:

The interest for a product continues to change occasionally. No business can work on its monetary growth without assessing client interest and future demand of items precisely. Sales forecasting refers to the process of estimating demand for or sales of a particular product over a specific period of time. For a good sales forecast, it is extremely important to get a good dataset as well. Forecasts heavily depend on the past records, trends and patterns observed for sales of a particular store. The variations could be due to a number of reasons. Talking from a business’s point of view, these sales forecasts are done consistently to improve their sales forecasting models as they directly impact their decision-making process, goals, plans and growth strategies. In this Retail Sales Prediction, machine learning models are created that predict sales of these 1115 drug stores across the European market and compare the results of these models. In addition to this, an effort has been made to analyze and find all the features that are contributing to higher sales and the features which are leading to lower sales, so that improvement plans can be worked upon.

Approach:

The approach followed here is to first check the sanctity of the data and then understand the features involved. The events followed were in our approach:

● Understanding the business problem and the datasets

● Data cleaning and preprocessing finding null values and imputing them with appropriate values. Converting categorical values into appropriate data types and merging the datasets provided to get a final dataset to work upon.

● Exploratory data analysis- of categorical and continuous variables against our target variable. ● Data manipulation- feature selection and engineering, feature scaling, outlier detection and treatment and encoding categorical features.

● Modeling- The baseline model Decision tree was chosen considering our features were mostly categorical with few having continuous importance.

● Model Performance and Evaluation

● Store wise Sales Predictions

● Conclusion and Recommendations

Understanding the Data:

First step involved is understanding the data and getting answers to some basic questions like; What is the data about? How many rows or observations are there in it? How many features are there in it? What are the data types? Are there any missing values? And anything that could be relevant and useful to our investigation. Let’s just understand the dataset first and the terms involved before proceeding further. Our dataset consists of two csv files, the first consists of historical data with 1017209 rows or observations and 9 columns with no null values. The second dataset was supplementary information about the stores with 1115 rows and 10 columns and a lot of missing values in a few columns. The data types were of integer, float and object in nature.

Let’s define the features involved:

● Id - an Id that represents a (Store, Date) duple within the set

● Store - a unique Id for each store

● Sales - the turnover for any given day (Dependent Variable)

● 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. An assortment strategy in retailing involves the number and type of products that stores display for purchase by consumers.

● Competition Distance - 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.

Data Cleaning and Preprocessing:

Handling missing values is an important skill in the data analysis process. If there are very few missing values compared to the size of the dataset, we may choose to drop rows that have missing values. Otherwise, it is better to replace them with appropriate values. It is necessary to check and handle these values before feeding it to the models, so as to obtain good insights on what the data is trying to say and make great characterization and predictions which will in turn help improve the business's growth. The historical records dataset had no null values.

Exploratory Data Analysis:

Exploratory data analysis is a crucial part of data analysis. It involves exploring and analyzing the dataset given to find out patterns, trends and conclusions to make better decisions related to the data, often using statistical graphics and other data visualization tools to summarize the results. The visualization tools involved in the investigation are python libraries- matplotlib and seaborn. The goal here is to explore the relationships of different variables with ‘Sales’ to see what factors might be contributing to the high and low sales numbers.

Data Manipulation:

Data manipulation involves manipulating and changing our dataset before feeding it to various regression machine learning models. This involves keeping important features, outlier treatment, feature scaling and creating dummy variables if necessary.

Feature Engineering:

● Some stores were closed due to refurbishment and some on account of week off or holidays. Those stores on those dates generated zero sales and hence removing the rows was important to avoid confusion by the algorithms and then removing the feature altogether because it wasn’t providing any value in prediction of the sales.

● There were features that like Competition Open since Month and Year. It was combined to count the total months since the nearest competition was opened.

● Promo2SinceWeek, Promo2SinceYear indicated promotion 2 opened since week and year. These features were combined to count the total months since promotion 2 is run.

● PromoInterval indicated the months for promotion 2 renewal. Hence, the sale month was compared against the interval and a new feature was created to determine whether the promo2 was renewed in that month.

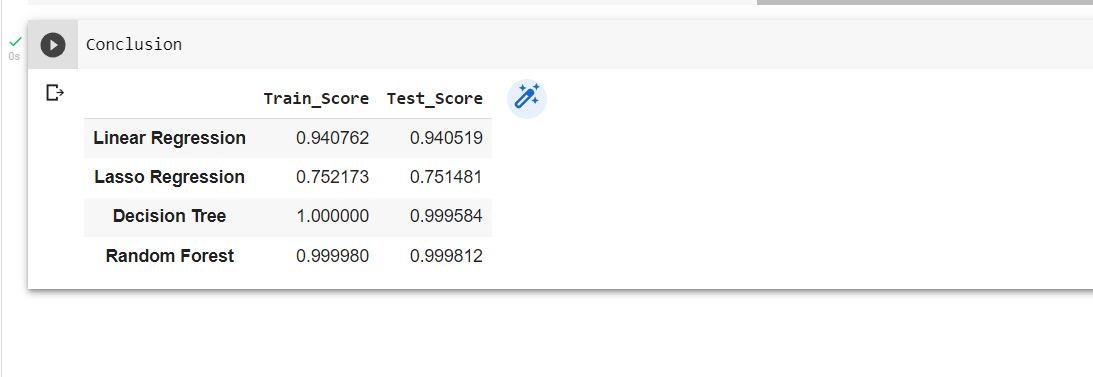
Modeling:

Factors affecting in choosing the model: Determining which algorithm to use depends on many factors like the problem statement and the kind of output you want, type and size of the data, the available computational time, number of features, and observations in the data, to name a few. The dataset used in this analysis has:

● A multivariate time series relation with sales and hence a linear relationship cannot be assumed in this analysis. This kind of dataset has patterns such as peak days, festive seasons etc which would most likely be considered as outliers in simple linear regression.

● Having X columns with 30% continuous and 70% categorical features. Businesses prefer the model to be interpretable in nature and decision-based algorithms work better with categorical data. Train-Test Split: In machine learning, train/test split splits the data randomly, as there’s no dependence from one observation to the other. That’s not the case with time series data. Here, it’s important to use values at the rear of the dataset for testing and everything else for training. The latest six weeks were kept as a testing set and the rest of the historical data was used in the training set.

Model Performance and Evaluation:



Conclusion:

The main objective of sales forecasting is to paint an accurate picture of expected sales. Sales teams aim to either hit their expected target or exceed it. When the sales forecast is accurate, operations go smoothly and future planning for the company's growth is done efficiently. Upon having this analysis, it can be established that given the dataset, the model developed is able to explain 99.9878 % of the variations and is able to predict the sales values in a good range.

Challenges:

● The major challenge would be the computational time and RAM needed to work upon such a dataset in a cloud environment.

References:

● Machine Learning Mastery

● GeeksforGeeks

● Analytics Vidhya Blogs

● Towards Data Science Blogs

● Built in Data Science Blogs

● Scikit- Learn Org

● Investopedia