**REGex PROJECT TEAM – 5**

**TOPIC – Store Sales Prediction Using Machine Learning**

**Team Members:-**

1. Avinash K

2. Raj Sankar GS

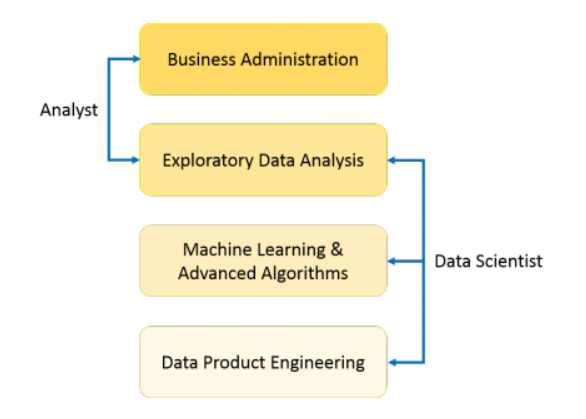
3. Misba Khatija

4. Manasvi Garg

**What is Data Science?**

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. But how is this different from what statisticians have been doing for years?

The answer lies in the difference between explaining and predicting.



As you can see from the above image, a Data Analyst usually explains what is going on by processing history of the data. On the other hand, Data Scientist not only does the exploratory analysis to discover insights from it, but also uses various advanced machine learning algorithms to identify the occurrence of a particular event in the future. A Data Scientist will look at the data from many angles, sometimes angles not known earlier.

So, Data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning.

**Introduction:-**

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.

**Data Description:-**

- 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.

**Approach:-**

- Data Exploration: load and check missing value,

- Data Visualization: EDA

- Data Manipulation: change data types, impute data, create new features,

- Data Modelling: multiple traditional ML algorithms regressions

**Libraries Used:-**

**1. Data Processing**

numpy

pandas

datetime (from pandas)

**2. Data visualization**

matplotlib.pyplot

seaborn

matplotlib.gridspec

LinearRegression, Lasso (from sklearn.linear\_model)

GradientBoostingRegressor (from sklearn.ensemble)

XGBRegressor (from xgboost)

xgboost

LGBMRegressor (from lightgbm)

lightgbm

**3. Data Evaluation**

mean\_squared\_error (from sklearn.metrics)

**4. Statistics**

ECDF (from statsmodels.distributions.empirical\_)

**5. Warning ignore**

import warnings

warnings.filterwarnings("ignore")

**OBSERVATIONS**

* On average customers spend about 9.50$ per day.
* About 20% of data has zero amount of sales/customers probably due to the fact that the store is closed for state holidays or school holidays
* 172817 closed stores with 0 sales.
* 52 store opened but without sales.
* We just keep opened stores with sales for analysis.
* StoreType B has the highest average of Sales among all others, however we have much less data for it
* Clearly, StoreType A come as first, StoreType D goes on the second place in both Sale and Customers
* Clearly any store with promotion attract more customers, leading to higher sales.
* Interestingly, sale escalates toward the end of the year, especially Christmas eve.
* SalePerCustomer is observed at StoreType D with the highest amount, $10 without Promo and 11 with Promo.
* StoreType B has the longest running period of promotion.
* It is noticed, StoreType B doesn't generate huge sale amount, but got the largest average sales.
* There is high relationship between Customers and Sales and Promo, but Promo2
* In the case of no promotion either Promo1 or Promo2, sale peaks on sunday.
* On the contrary, store running Promo1 make the most of sale on monday.
* Promo2 seems irrelevant to overall
* At the first glance, store 1 show stationary pattern. Unfortunately, the trend experience downward.

Table

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

* Seeing both feature importance of XGBoost and LightGBM show similar patterns.
* Day, DayOfWeek, WeekOfYear, PromoOpen, Promo primarily account for sale amount

**Regression Models Used:-**

**Linear Regression:-**

Linear Regression is a machine learning algorithm based on supervised learning. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression.

**Lasso Regression:-**

Lasso regression is another regularization technique to reduce the complexity of the model. It stands for Least Absolute and Selection Operator. It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights. Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0. It is also called as L1 regularization. Hence, the Lasso regression can help us to reduce the over fitting in the model as well as the feature selection.

**Gradient Boosting Regressor:-**

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor’s error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels. There is a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees).

**XGB Regressor:-**

XGBoost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners. The objective function contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e. how far the model results are from the real values. The most common loss functions in XGBoost for regression problems are reg: linear, and that for binary classification is reg: logistics.

**LGBM Regressor:-**

LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.

It uses two novel techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB) which fulfils the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks. The two techniques of GOSS and EFB described below form the characteristics of LightGBM Algorithm. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT frameworks