TOPIC: *Rossmann Store Sales Prediction*

Concept of the project-

The demand for a product or service keeps changing from time to time. No business can improve its financial performance without estimating customer demand and future sales of products/services accurately. Sales forecasting refers to the process of estimating demand for or sales of a particular product over a specific period of time.

Problem Statement-

Build a predictive model using machine learning for predicting sales of a major store chain Rossmann. Rossmann operates over 3000 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.

The objective of the project-

The objective is to analyze historical sales data from Rossmann stores to identify key factors that influence sales performance and develop a predictive model to forecast future sales.

Data sources used-

Rossmann-store-sales/Data

- **Rossmann Stores Data.csv** - Historical Data Including Sales
- **Store.csv** - Supplemental Information About The Stores

Data Analytics software used-

Python & Google Colab Libraries used:

- **Numpy**- Solve complex mathematical problems
- **Pandas**- Use for Data frame Manipulation
- **Seaborn**-To create data visualization
- **Matplotlib-** To create data visualization
- **Ipywidgets-** Interactive analysis

Machine Learning Algorithms used-

Linear Regression

Data sets probable visualizations-

Bar Graphs and Line Charts will be used using for better visualization

Methodology-

We try to understand the distribution between dependent and independent features.

We plotted the relationships using various plots like bar plot , histogram , line plot and scatter plot etc. we found some insights and relationship like the sales of day 7 (Sunday) are the lowest among them (because of the store being closed on Sunday). The sales of store greatly impact by promo, promo running increases the sales. Also, the number of customers and sales are directly correlated, the correlation coefficient being 89%.

Probable Outcome-

- Day of weeks was found to be the most important feature (using xgboost Regressor), which is contributing the highest in predicting the target variable.
- We found 89% accuracy through linear regression.
- The Regularization techniques of linear regression (Lasso, and Ridge) did not help in improving the accuracy much.
- In decision tree regression, we found 97% accuracy.
- The ensembles of decision tree i.e. Random forest gave us the highest accuracy i.e. 98% in predicting the target variable.
- However through Xgboost Regressor, We got 93% accuracy.
- So we found Random forest to be the best performing algorithm.

Conclusion-

Starting with loading the data so far we have done EDA, null values treatment, encoding of categorical columns, feature selection and then model building.

In all of these models our accuracy revolves in the range of 89% to 98%.

And there is no such improvement in accuracy score even after hyper parameter tuning.

So the accuracy of our best model is 98% which can be said to be good for this large dataset.