**Project Summary**

**Retail Sales Prediction for Rossmann Stores**

**Project Overview**

This technical document provides a summary of a project aimed at predicting retail sales for Rossmann Stores. The project involves the utilization of historical data, encompassing information about the stores, competitors, types, holidays, customers, and sales transactions. The workflow includes essential steps such as data collection, cleaning, exploratory data analysis (EDA), feature engineering, and model implementation.

**Data Understanding and Cleaning**

Before diving into model development, the team undertook a comprehensive data understanding phase. This involved collecting and merging datasets, handling missing values, and preparing features for visualization. The EDA phase provided valuable insights that informed subsequent machine learning model decisions.

**Feature Engineering and Preprocessing**

Feature engineering was a crucial aspect of the project. New features like PromoDuration and CompetitionDuration were created, and irrelevant features were removed. The team also detected and treated outliers in the data. Categorical features were encoded numerically using One-Hot Encoding to make them suitable for machine learning algorithms.

**Data Transformation**

To ensure the data met the assumptions of machine learning models, data transformation techniques were applied. These techniques aimed to achieve a normal distribution and uniform scaling.

**Model Implementation**

The dataset was split into training and testing sets, and several machine learning algorithms were explored, including:

1. **Linear Regression:** Achieved 85.25% accuracy with an R2 score of 0.70.
2. **Decision Trees:** Captured 90% of variance with 92.85% accuracy.
3. **Random Forests:** Reached 93.86% accuracy and reduced processing time.
4. **XGBoost:** Selected as the final model with 93.94% accuracy and a 6.06% mean absolute percentage error.

**Model Selection**

The final selection of the XGBoost model was based on its high accuracy and low mean absolute percentage error. The residuals' mean of 1.31 indicated normally distributed residuals, making XGBoost the optimal model for deployment.

**GitHub Repository**

You can find the complete project, including code and documentation, on the GitHub repository: [Retail Sales Prediction - Regression Project](https://github.com/arjundomle07/Retail-Sales-Prediction--Regression-Project)

**Problem Statement**

**Sales Prediction for Rossmann Stores**

**Background**

Rossmann, a prominent pharmacy chain with stores in 7 European countries, faces the challenge of accurately predicting daily sales for their 3,000+ stores. Sales are influenced by numerous factors, including promotions, competition, holidays, seasons, and local conditions. The current approach relies on store managers to make predictions, leading to variations in accuracy due to individual circumstances.

**Project Goal**

The primary goal of this project is to create a predictive model for the "Sales" column in the test dataset. It's essential to note that some stores in the dataset were temporarily closed for refurbishment, adding complexity to the forecasting task.

**Challenge**

In simpler terms, the challenge is to build a model that can accurately predict the daily sales of Rossmann stores, taking into account various factors, and to do so consistently across all stores, even when some are temporarily closed for renovation. This predictive model will help Rossmann optimize their operations and resource allocation.

This project offers a data-driven solution to a real-world problem faced by Rossmann, enabling them to make more accurate and consistent sales predictions, ultimately enhancing their operational efficiency.