Rossman Sales Prediction - Project Report

1. Introduction

The goal of this project is to predict the sales for Rossman retail stores using machine learning techniques. Accurate sales forecasting is crucial for retail companies to optimize inventory, staffing, and marketing efforts. By analyzing historical sales data and store-level features, we can develop a model that helps predict future sales, supporting data-driven decision-making in business operations.

The project is divided into the following steps:

- Data Collection and Preprocessing
- Exploratory Data Analysis (EDA)
- Model Building and Evaluation
- Recommendations for Improving Sales Prediction Accuracy

2. Data Collection and Quality Enhancement

The dataset consists of two files:

- train.csv: Contains historical sales data, including store number, date, sales, and other relevant features.
- store.csv: Contains store-level features, such as store type, location, and size.

To ensure the quality of the data, the following preprocessing steps were undertaken:

- Missing Values: Handled by imputing or removing the relevant entries.
- Categorical Variables: Encoded using label encoding.
- Outlier Detection: Outliers were detected using z-scores and IQR techniques to ensure data integrity.
- Box-Cox Transformation: Applied to the target sales variable to stabilize variance and improve model performance.

3. Data Analysis and Insights

Exploratory Data Analysis (EDA) was performed to uncover key insights from the data. Some of the key findings include:

- Sales Patterns: Sales show a weekly pattern, with a noticeable increase on weekends. Holidays also result in significant sales spikes.
- Store Characteristics: Larger stores and stores located in urban areas tend to have higher sales, which aligns with business expectations.

- Seasonality: Certain months of the year, particularly around Christmas and New Year's, see a marked increase in sales.

The following visualizations were used to support the analysis:

- Time Series Plots: Sales trends over time.
- Correlation Heatmaps: Relationships between numerical variables.
- Box Plots: Sales distribution across different stores and categories.

4. Modeling and Recommendations

We used machine learning models to predict sales based on the features available. The following steps were followed:

1. Model Selection:

- We initially used a Decision Tree Regressor; however, it showed high RMSE values.
- We applied a Box-Cox transformation on the sales data, which improved the data distribution and reduced RMSE.
- Finally, a Random Forest Regressor was chosen for its ability to handle non-linear relationships and provide feature importance.

2. Model Evaluation:

- The Root Mean Squared Error (RMSE) after applying Box-Cox transformation was 15.034.
- The R^2 Score was 98%, indicating excellent model fit and predictive accuracy.

3. Recommendations:

- Model Refinement: Further fine-tuning of hyperparameters (using techniques like Grid Search) could improve model accuracy.
- Feature Engineering: Incorporating additional features like promotions, weather data, and economic indicators could enhance prediction quality.

5. GitHub Repository

All project files, including the Jupyter notebook (ross-sales-prediction.ipynb), dataset files (train.csv and store.csv), and necessary documentation, are available on **https://github.com/dscharan97/Rossman-Sales-Prediction.** The repository also includes a README file with instructions on how to run the notebook and use the model.