Rossman Sales Prediction

Name: Sai Charan Desireddy

Date:2/12/2024

Problem Statement

The Challenge:
Rossman stores need accurate sales forecasts to:

- Optimize inventory levels.
- Allocate staff effectively.
- Plan promotions and marketing campaigns.

The Goal:

- Build a predictive model to forecast daily sales using historical and store-level data.

Approach

1. Data Collection and Understanding:
   - Analyzed historical sales (train.csv).
   - Incorporated store-level features (store.csv).
2. Data Preprocessing:
   - Cleaned and enhanced data quality.
3. Exploratory Data Analysis (EDA):
   - Identified trends, patterns, and anomalies.
4. Modeling:
   - Applied machine learning for sales prediction.
5. Evaluation and Recommendations:
   - Assessed model performance and suggested improvements.

Data Overview

train.csv:

- ~1,000,000 records, spanning multiple years.
- Features: Date, store ID, sales, customers, and promotions.

store.csv:

- Store-specific attributes like size, type, and competition.

Key Challenges:

- Missing values in store.csv for competition and promotion data.
- High variance in sales patterns across stores.

Actions Taken:

- Imputed missing values.
- Removed irrelevant columns.
- Applied Transformations on columns.

Exploratory Data Analysis

1. Seasonality:
   - Significant sales increase during holiday seasons.
   - Weekly sales spikes observed on Saturdays.
2. Store Performance:
   - Larger stores with Type A show consistently higher sales.
3. Customer Behavior:
   - Positive correlation between customer count and sales.
4. Visuals include:
   - Time series plot (sales vs. time).
   - Bar chart (average sales by store type).
   - Correlation heatmap (relationships between features).

Modeling Strategy

Initial Model: Decision Tree Regressor

- Challenge: High RMSE, despite good accuracy.

Model Enhancement: Box-Cox Transformation

- Effect: Improved data distribution, reducing RMSE.

Final Model: Random Forest Regressor

1. Performance:

- RMSE: 15.034 (improved after Box-Cox).
- R^2 Score: 98% (indicating excellent model fit).

2. Why Random Forest?

Handles non-linearity, overfitting, and provides feature importance.

Model Evaluation

1. RMSE:
   - Before Box-Cox: High RMSE.
   - After Box-Cox: RMSE reduced to 15.034, improving predictions.

2. R^2 Score:
   - 98%, indicating the model explains most of the variance in sales.
   - Feature Importance (Random Forest):

Conclusion

Key Insights:

- Random Forest with Box-Cox transformation significantly improved prediction accuracy.
- The model explains 98% of the variance in sales data (R^2 = 98%).

Future Steps:

- Explore further feature engineering and model optimization.
- Implement the model in real-time systems for operational use at Rossman.

GitHub Repository

GitHub Repository: https://github.com/dscharan97/Rossman-Sales-Prediction

Contents:

- Jupyter notebook with the full analysis and model code.
- Data files and preprocessing steps.
- Model evaluation and recommendations.