Methodology Summary of Big Mart Sales Prediction

Related Work

Ranjitha and Spandana (2021) [1], Manika (2019) [8], and Bhavana & Lakshmi (2022) [9] built and evaluated XGBoost, Linear Regression (LR), Polynomial Regression (PR), and Ridge Regression (RR) by accuracy, RMSE, MAE and MSE, with RR and XGBoost giving the better prediction. Rao et al (2023) [6] found that Random Forest (RF) outperforms RR, LR and Regression Tree (RT), using evaluation metric of RMSE. To improve the performance of single model, Kumari et al (2018) [2] applied the ensemble method via 2-layer stacking, with the bottom layer of LR, SVR & Cubist, and the top layer of Cubist, reducing MAE from 0.41-0.45 down to 0.39. Based on 6 single models, Murthy et al (2022) [5] presented Grid Search Optimization (GSO) approach for parameter optimization and hyperparameters tuning, as well as an ensemble of XGBoost, bringing R-square from 0.52 to 0.59. Ilyas et al (2023) [7] compared regression results of SMO, Linear, Additive, MLP and M5P, M5P is the best one based on MAE, RMSE and RRSE. Daniyal (2022) [11] proposed TOR AI approach and compared MSE and variance with LR, AdaBoost, XGBoost, RF and SVM. TOR AI shows lower MSE and variance.

Methodology

Exploratory Data Analysis (EDA)

1. Missing Values: Mean/Median/Category – See Distribution
2. Outliers & Correlation: See Distribution
3. Feature Engineering: New Feature, Feature Selection

Baseline Models

1. Linear Models, with assumption validation and fixing (Lasso, Ridge, Best Subset)
2. Regression Tree
3. SVM

Improvements

1. Ensemble Learning: Bagging + Boosting + Stacking
2. Deep Learning – Neural Network