CSE523 Machine Learning

Weekly Project Report

Date: 01-04-2023

Project title: Big Mart Sales Prediction

Group 10

Name	Enrolment no.
Meet Patel	AU2040010
Dev Patel	AU2040056
Kush Patel	AU2040137
Vatsal Shah	AU2040019

1. **Task performed and outcomes of task performed this week**

- We implemented hyperparameter tuning and regularization on the XGBoost model as it was overfitting on the training.
- The R^2 value on training set was 0.86 and on test set was 0.53.
- There are many hyperparameters XGBoost model but we selected the six most important parameters to tune:

   1. **n_estimators (number of trees):** Setting the number of trees informs the algorithm when to stop, to prevent over-fitting.

   2. **max_depth (maximum tree depth):** The larger the tree depth, the higher the probability of over-fitting; therefore, it is prudent to increase it reluctantly and only by units of one and even then, probably never higher than 5

   3. **learning_rate:** The learning rate (α) will be multiplied by the weight in every tree. There is usually an inverse relationship between the learning rate and accuracy. In other words, a lower learning rate improves the final model measured in predictive accuracy (lower cost or error), even though it makes it slower to train.

   4. **min_child_weight:** The minimum child weight hyperparameter is technically defined as the Hessian minimum sum of an instant weight that is necessary in a child. It regularizes by limiting the depth of trees, which helps prevent over-fitting.

   5. **reg_lambda:** This is the lambda value of the L2 regularization. This is a regularization factor to help prevent over-fitting and make models more parsimonious — less complex — when there are many features.

   6. **Gamma:** Gamma specifies the minimum loss reduction required to make a split. It makes the algorithm conservative. The values can vary depending on the loss function and should be tuned. The larger gamma is, the more conservative the algorithm will be.

- We tried different combinations of the values for the above mentioned parameters to prevent the model from overfitting.
- We got the best result by using the following values for the parameters:

   1. **max_depth=2,**

   2. **n_estimators=100,**

   3. **learning_rate=0.2,**

   4. **min_child_weight = 200,**

     5. **reg_lambda=200,**

     6. **gamma=700**

- The training set R^2 value is 0.611 and test set R^2 value is 0.604.
- We prevented the model from overfitting, but the accuracy decreased.

2. Tasks to be performed in the upcoming week

- We will perform hyperparameter tuning to improve Random Forest model accuracy.
- We will improve XGBoost model to increase its accuracy.