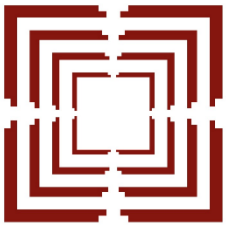

Ahmedabad University

CSE523 Machine Learning

Weekly Project Report

Date: 08-04-2023

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 Random Forest model as it was overfitting on the training.
    - The R^2 value on training set was 0.93 and on test set was 0.57.
    - There are many hyperparameters Random Forest 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. **max_features:** This hyperparameter controls the maximum number of features that are considered for splitting at each node. A larger value can lead to better performance, but also increases the risk of overfitting.
        4. **min_samples_split:** This hyperparameter specifies the minimum number of samples required to split an internal node. Increasing this value can help prevent overfitting.
        5. **min_samples_leaf:** This hyperparameter specifies the minimum number of samples required to be at a leaf node. Increasing this value can help prevent overfitting.
        6. **ccp_alpha:** This hyperparameter controls the complexity of the decision trees by imposing a penalty on each tree's total number of splits. Increasing this value can lead to simpler trees, which may improve generalization performance.
    - We tried different combinations of the values for the above-mentioned parameters to prevent the model from overfitting.

R-squared							
train	test	n_estimators	max_dept	max_featu	min_samp	min_samp	ccp_alpha
0.6261	0.5915	300	8	3	12	10	0.001
0.6419	0.6038	200	8	5	10	10	0.01
0.6449	0.6044	200	8	6	15	10	0.001
0.6207	0.6037	100	6	7	5	5	0.01
0.622	0.6046	200	7	7	50	30	0.01
0.6116	0.6035	200	7	7	100	70	0.01
0.6061	0.6018	200	8	8	150	100	0.01
0.6063	0.6023	200	9	8	150	100	0.01
0.6232	0.6046	1000	9	7	90	50	0.01
0.623	0.6049	200	9	7	90	50	0.01

- We got the best result by using the following values for the parameters:
    1. **n_estimators = 200**
    2. **max_depth = 7**
    3. **max_features = 7**
    4. **min_samples_split = 100**
    5. **min_samples_leaf = 70**
    6. **ccp_alpha = 0.01**
- The training set R^2 value is 0.6116 and test set R^2 value is 0.6035.
- We prevented the model from overfitting, but the accuracy decreased.

2. **Tasks to be performed in the upcoming week**
    - We will run all the models again with more improvements and do the analysis of all the models.
    - We will compare the results and errors of all the models.