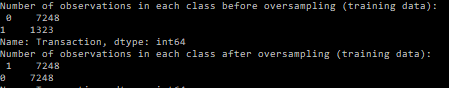
# Decision Tree Ensembles

The task is to create models to predict whether a transaction will take place or not for an E-commerce website. We will be evaluating 3 classification models and based on the evaluation scores, will decide the final model. For all the models, data needs to prepared first.

## Data Preparation

1. Converting Categorical features to numerical
2. Using in-built map function
3. Using a converter function to convert Boolean values (Weekend & Transaction) to numerical 1 & 0.
4. Diving dataset into features and labels (Transaction as target variable, rest as input variables)
5. Normalizing features so that value ranges are not different for variables
6. Train and Test set splitting
7. Since the target variable does not have balanced classes, using SMOTE for over sampling and creating synthetic samples.



## Model Creation

We will be creating three models – Random Forest Classifier, Adaptive Boosting Classifier, Gradient Boosting Classifier. All the models will be tuned and cross validated 5 times using GridsearchCV. Tuning parameters will be decided based on the evaluation chosen below –

## Evaluation Metrics

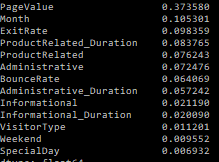
The focus of the model should be to minimize false negatives as the E-shop website’s management would not want to lose future True Transactions if the model is predicting them as False. Hence the recall score is evaluation score. The model with the best recall score would be the final model.

### Random Forest

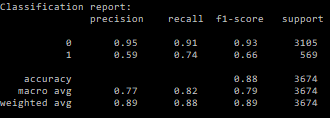
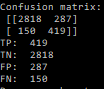
Appropriate number of tress for Random Forest Classifier if the scoring is ‘recall’ are – 150.



Building model using number of tress as 150, following feature importance is received with below classification report -



Classification report for Random Forest-

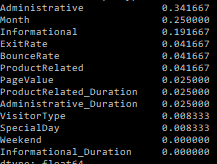
## Adaptive Boosting

Using scoring as ‘recall’, finding appropriate number of stumps for AdaBoost Classifier-

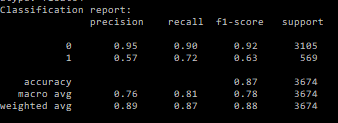
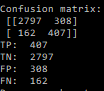
  

Using Number of stumps as 120, creating AdaBoost model.

Feature Importance



Classification report -

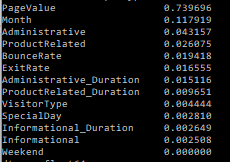
 

## Gradient Boosting Classifier

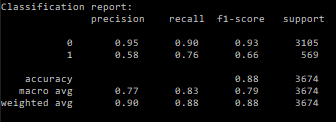
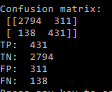
Using scoring as ‘recall’, finding appropriate depth, leaf nodes and for Gradient Boosting Classifier



Tuning Gradient Boosting Classifier with max\_depth = 11 (maximum number of nodes in the tree), max\_leaf\_nodes = 32 (best-first nodes) and boosting stages = 20.



Classification Report

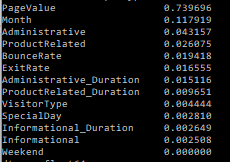
## Model Selection

Comparing the three models based on their recall scores, for classification 1 (as Positive for the model is 1 (True)).

Recall Score is best for Gradient Boosting Classification has it is the final model for this dataset.

## Interpretation and Conclusion

From the feature importance list of the final gradient boosting model, it is evident that PageValue, Month and Administrative are the key features to determine whether a user would make the Transaction on the E-shop or not.



PageValue is intended to give you an idea of which page in the site contributed more to the site's revenue. If the page was not involved in an ecommerce transaction for the website in any way, then the PageValue for that page will be 0 since the page was never visited in a session where a transaction occurred. So, for this E-shop, it is critical that a user navigates to or returns to as many pages as possible. Since Shopping cart is the most common page with high Page Value, the Eshop can have pop-ups build to give user easy access to that page.

Month: Most businesses run well during the holiday seasons (summer vacations, Christmas holiday etc). For this E-shop, the month with highest transactions is November followed by May and December. It can be assumed the users were Christmas gift shopping in November and December, it is advisable for Eshop to have more exciting offers or promotions around these months.