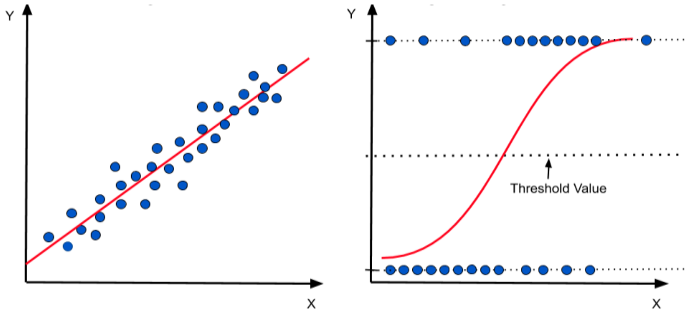
Click Through Rate Prediction

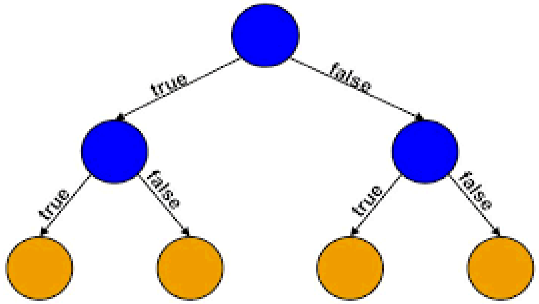
Click Through Rate Prediction

- CTR – Model Selection
- Understanding Features
- Optimizing Model and reducing error
- Picking the optimum Models
- Implication of Error Made in Prediction

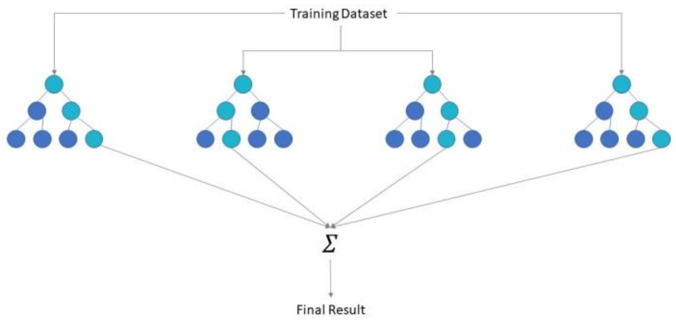
CTR – Model Selection



When the data was analyzed , as the outcome variable was binary , intention was to use *Logistic Regression (Classification Problem)*



Picked the decision tree model as the data was having categorical variables and *Decision tree* will choose between several courses of action.



Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems.

Understanding Features

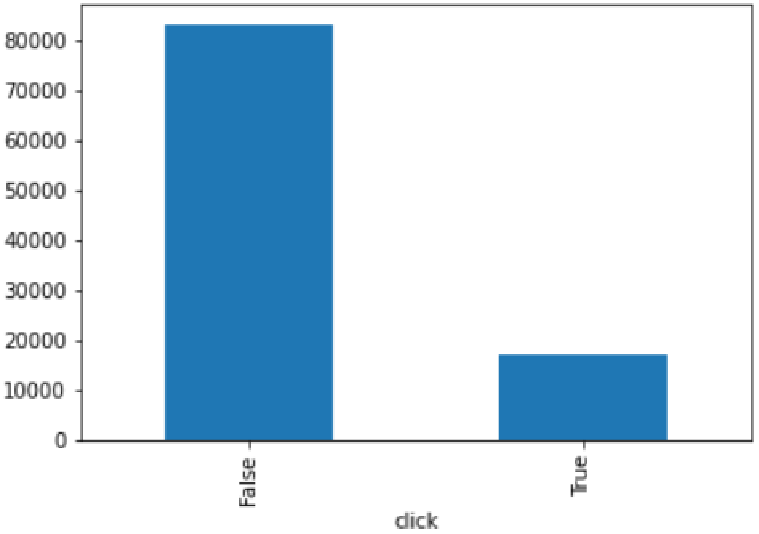
- Target Features Click > Site Features > Device Feature > Anonymous Feature

-site features : site_id, site_domain, site_category

-app feature: app_id, app_domain, app_category

-device feature: device_id, device_ip, device_model, device_type, device_conn_type

-anonymized categorical features: C14-C21



The overall click through rate is 17%, and approx. 83% is not clicked.

Understanding Features Continued

To select the best subset from the original feature space and in order to reduce the collinearity , multi – collinearity and finding the feature importance , we have used the below techniques , which provided insight into the columns that can be removed:

To achieve the above we used below techniques :

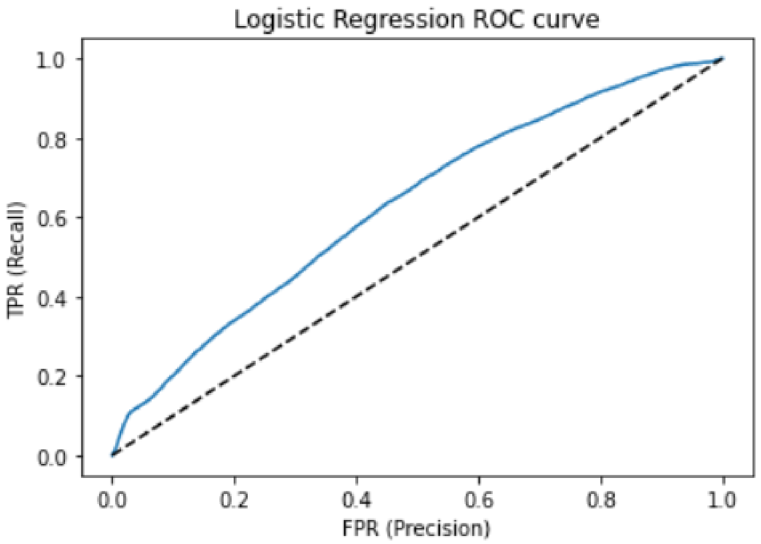
- VIF – Variance Inflation Factor calculation
- Feature importance using EXTRATREESCLASSIFIER available in SKLEARN library.
- Correlation heat map

-All the data in month is unique hence there is no information which can derived from the same.
-The outcome of the variables are in two column click and y and hence one of them is dropped.
-There were few anonymous categorical variable which were highly colinear , we picked couple of columns – C14 and C1

Logistic Regression

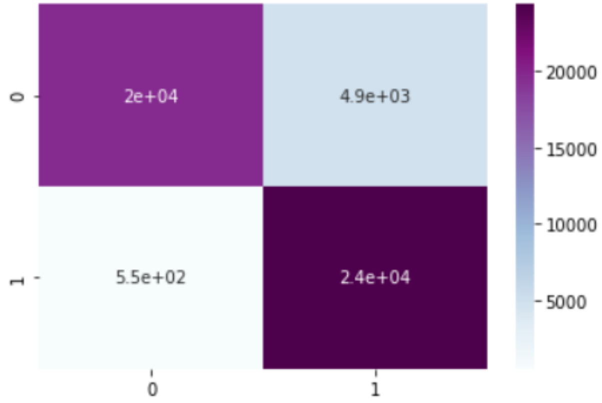
Reason to chose Logistic Regression :- As we can see, logistic regression is used to predict the likelihood of all kinds of “yes” or “no” outcomes. In our case of CTR prediction, by predicting such outcomes, logistic regression will help to make informed decisions , in which way advertisement should be launched , based on features like banner position , site categories , app related categories etc.

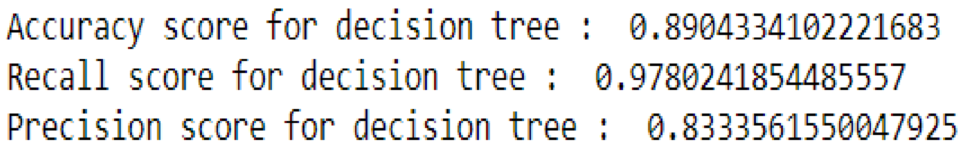
important details :- As we know the output of a logistic regression model is the probability of our input belonging to the class labeled with 1. And the complement of our model's output is the probability of our input belonging to the class labeled with 0



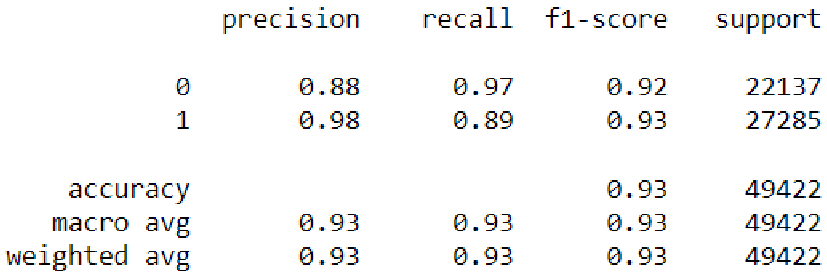
Decision Tree

- The predictability of decision tree model is quite good with TP and TN quite significant, it has less False Negative .The model has very high recall value and hence error in judgement is less.
- Decision trees are inherently highly unstable i.e. any change in the original dataset will have a huge difference in the model itself
- For data including categorical variables with different numbers of levels, information gain in decision trees is biased in favor of those attributes with more levels.
- Calculations can get very complex, particularly if many values are uncertain and/or if many outcomes are linked





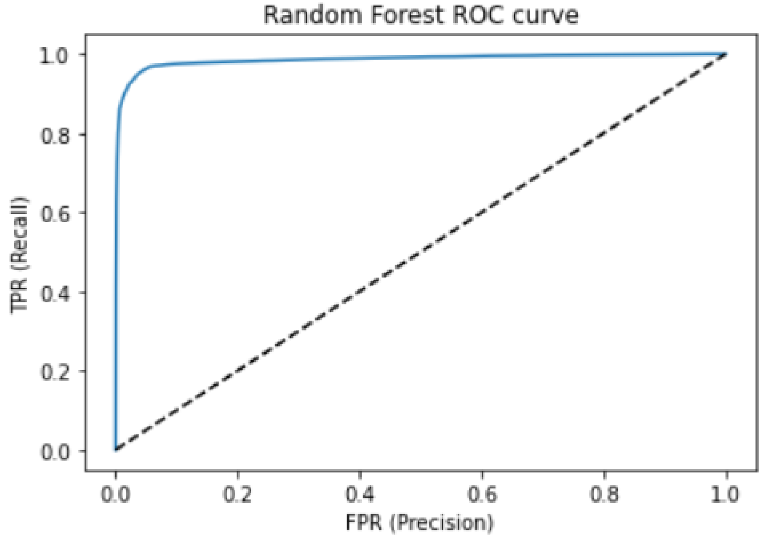
```
Accuracy score for decision tree :  0.890433410222168З
Recall score for decision tree :  0.978024185448555７
Precision score for decision tree :  0.833356155004792５
```



```
              precision    recall  f1-score   support

           0       0.88      0.97      0.92     22137
           1       0.98      0.89      0.93     27285

    accuracy                           0.93     49422
   macro avg       0.93      0.93      0.93     49422
weighted avg       0.93      0.93      0.93     49422
```



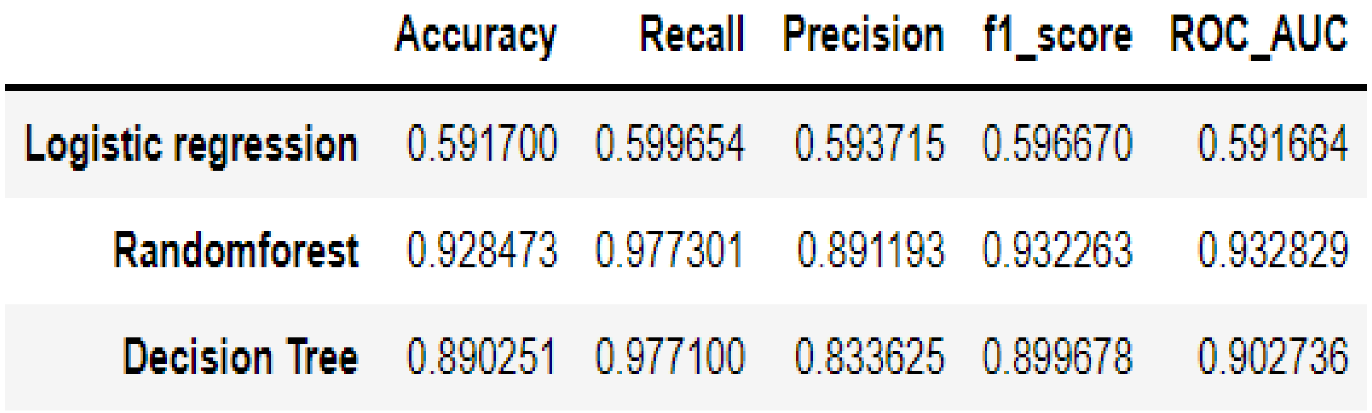
Model of Choice - Random Forest

The idea of CTR is better prediction as it is marketing related data. Predictability will, hold the key instead of interpretability, and hence we picked random forest for prediction.

Although we were able to achieve perfect accuracy in the Logistic regression model , but since we had to drop the features and after performing RandomSearchCV and picking the best penalty , we do not prefer Logistic regression for prediction.

Model Comparison

We clearly can see that Random Forest got highest accuracy out of all three although we got high accuracy in Logistic rogation but after applying regurization and drooping multicolinear feature It no longer hold highest number.



	Accuracy	Recall	Precision	f1_score	ROC_AUC
Logistic regression	0.591700	0.599654	0.593715	0.596670	0.591664
Randomforest	0.928473	0.977301	0.891193	0.932263	0.932829
Decision Tree	0.890251	0.977100	0.833625	0.899678	0.902736

Implication of Error in Model

We should be performing error on three levels – Prediction, Data, and Features.

Prediction

- For the random forest model , accuracy is 93 percent and recall is also 93 percent , obviously we can see on test data set that False Negative is hardly 6 percent of the total data. False positive is below 1 percent and hence the action we take on incorrect positive click rate in terms of launching any offer based on advertisement should be less.

Data

- Since the data was highly imbalanced , we had to use the random over sampler , which balanced out the click through rates.

Features

- Identify the relevance of feature using various techniques , before dropping the features , it might happen that some important features might get missed which could have led to better prediction.
- However, using the effective techniques of VIF , Feature importance and correlation we were able to use the features appropriately.