Click Through Rate - Prediction

Praveen Selvaraj

MSDS

Batch-C5

Data Understanding

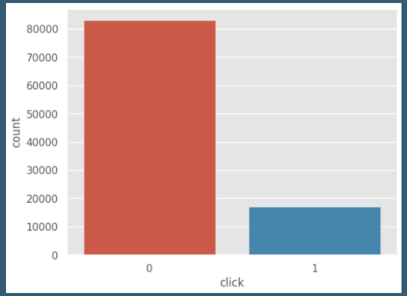
Variable	Description
click:	0/1 for non-click/click
hour: format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC.	
C1	anonymized categorical variable
banner_pos	position of the ad/banner on the page
site_id	unique id of the site on which the ad is shown
site_domain	unique domain of the site on which the ad is shown
site_category	category of the site on which the ad is shown
app_id	app id of the site on which the ad is shown
app_domain	app category of the site on which the ad is shown
app_category	category id of the site on which the ad is shown
device_id	device id on which the add was shown
device_ip	ip address of the device on which the ad was shown
device_model	model type of the device on which the ad was shown
device_type	the device type on which the ad was shown
device_conn_type	the connection type of the device on which the ad was shown
C14 - C21	anonymized categorical variable

Data Display

click	C1	banner_pos	site_id	site_domain	site_category	app_id	app_domain	app_category	device_id	device_ip	device_model	device_type	device_conn_type	C14	C15	C16	C17	C18	C19	C20	C21	month	dayofweek	day	hour	y
0	False	1005	1	856e6d3f	58a89a43	f028772b	ecad2386	7801e8d9	07d7df22	a99f214a	962c8333	be6db1d7	1	0	22683	320	50	2528	0	39	100075	221	10	1	28	14
1	True	1005	1	e151e245	7e091613	f028772b	ecad2386	7801e8d9	07d7df22	a99f214a	5b1f94b9	1b13b020	1	0	17037	320	50	1934	2	39	-1	16	10	2	22	19
2	False	1005	0	e3c09f3a	d262cf1e	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	a9a84f4c	9a45a8e8	1	0	22155	320	50	2552	3	167	100202	23	10	3	23	18
3	False	1002	0	0da94452	248e439f	50e219e0	ecad2386	7801e8d9	07d7df22	0fa578fd	88c62dad	ea6abc60	0	0	21591	320	50	2478	3	167	100074	23	10	2	22	19
4	True	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	1e5e0d0e	36d749e5	1	0	15708	320	50	1722	0	35	-1	79	10	1	21	8

High level inference of Dataset

- The data has total 99,999 rows and 27 columns
- The dataframe doesn't have null or missing values
- The data has total 18 numerical and 9 categorical columns
  - Numerical columns =['click', 'C1', 'banner_pos', 'device_type', 'device_conn_type', 'C14', 'C15', 'C16', 'C17', 'C18', 'C19', 'C20', 'C21', 'month', 'dayofweek', 'day', 'hour', 'y']
  - categorical columns = ['site_id', 'site_domain', 'site_category', 'app_id', 'app_domain', 'app_category', 'device_id', 'device_ip', 'device_model']
- The target variable is highly imbalanced as it has 83% of 0's and 17% of 1's



Target encoder for categorical features

C1	banner_pos	site_id	site_domain	site_category	app_id	app_domain	app_category	device_id	device_ip	device_model	device_type	device_conn_type	C14	C15	C16	C17	C18	C19	C21	dayofweek	day	hour
0	1005	1	0.03	0.03	0.18	0.20	0.19	0.20	0.17	0.15	0.19	1	0	22683	320	50	2528	0	39	221	1	28
1	1005	1	0.30	0.26	0.18	0.20	0.19	0.20	0.17	0.19	0.28	1	0	17037	320	50	1934	2	39	16	2	22
2	1005	0	0.05	0.03	0.21	0.20	0.19	0.20	0.17	0.15	0.10	1	0	22155	320	50	2552	3	167	23	3	23
3	1002	0	0.14	0.14	0.13	0.20	0.19	0.20	0.15	0.15	0.19	0	0	21591	320	50	2478	3	167	23	2	22
4	1005	0	0.20	0.20	0.21	0.20	0.19	0.20	0.17	0.28	0.22	1	0	15708	320	50	1722	0	35	79	1	21

Numerical variable inference

- Observations from numerical feature
  - Y and Click looks like same columns, after co-relation we can drop on of them.
  - month column has only 1 data entry, no exrtra information is added, can be dropped
  - banner pos, device conn, C20, C15, C16 looks like data is cenetered around certain values.

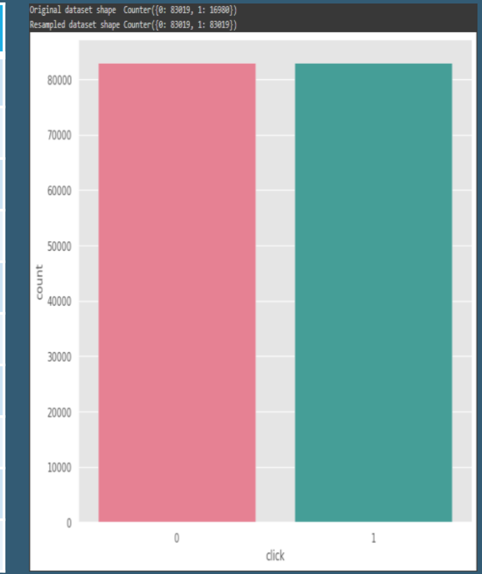
- Observations from Pearson correlation co-efficient
  - month has got no significance, better to drop it
  - y and click are same drop click column
  - C14 and C17 are highly co-related, later will remove one of them after the base model.
  - device type with C1 are highly co-related, later will remove one of them after the base model.
  - Removing C20 anomalised column, since it have got nearly 47% of values with -1. As a categorical variable it's not expected to have values as -1.

Feature Engineering - Inference

- P values and VIF looks good, will find the best threshold for classification
- Decision tree with right features seems data is overfitting. having the correct hyper parameter tuning help in interpretation and bit of over fitting of the model.
- VIF top10

Features	VIF
const	175.79
site_id	9.06
site_domain	8.97
app_id	2.50
app_category	2.34
app_domain	2.33
site_category	1.92
C21	1.76
C16	1.73
C18	1.61

Sampling target feature



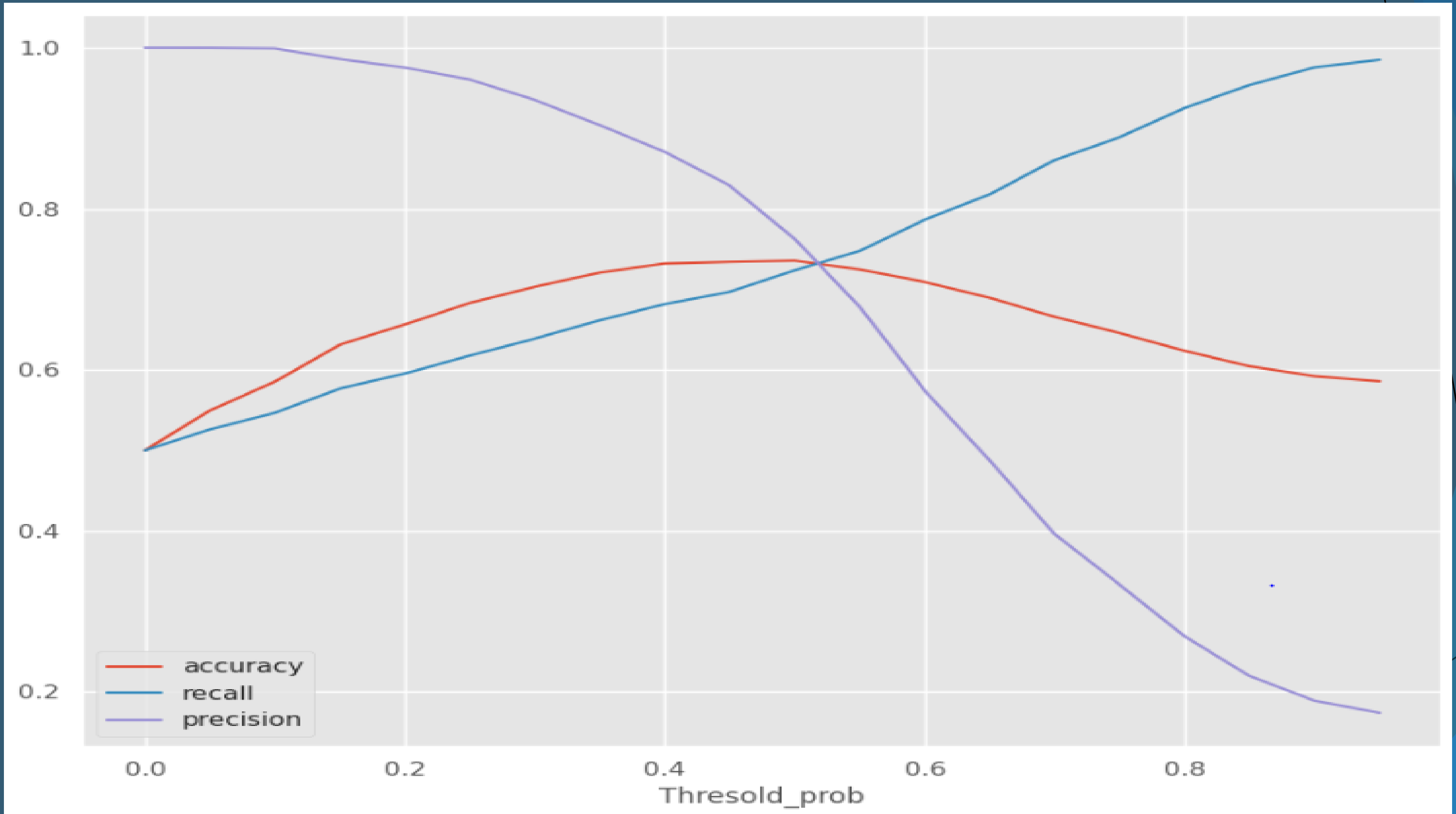
VIF after removing hour,day

Features	VIF
const	180.97
site_id	9.06
site_domain	8.97
app_id	2.51
app_domain	2.34
app_category	2.34
site_category	1.92
C21	1.76
C16	1.73
C18	1.61

VIF after removing dow, device_ip, C15

Features	VIF
const	68.67
site_id	9.05
site_domain	8.88
app_id	2.41
app_category	2.32
app_domain	2.31
site_category	1.91
C21	1.74
C16	1.73
C18	1.61

Threshold prob for Accuracy, Recall and Precision



Decision Tree & Random Forest - feature importance

features	importance
device_ip	0.85
hour	0.03
device_model	0.03
device_id	0.02
dayofweek	0.01
C17	0.01
site_domain	0.01
day	0.01
site_id	0.01
C19	0.00

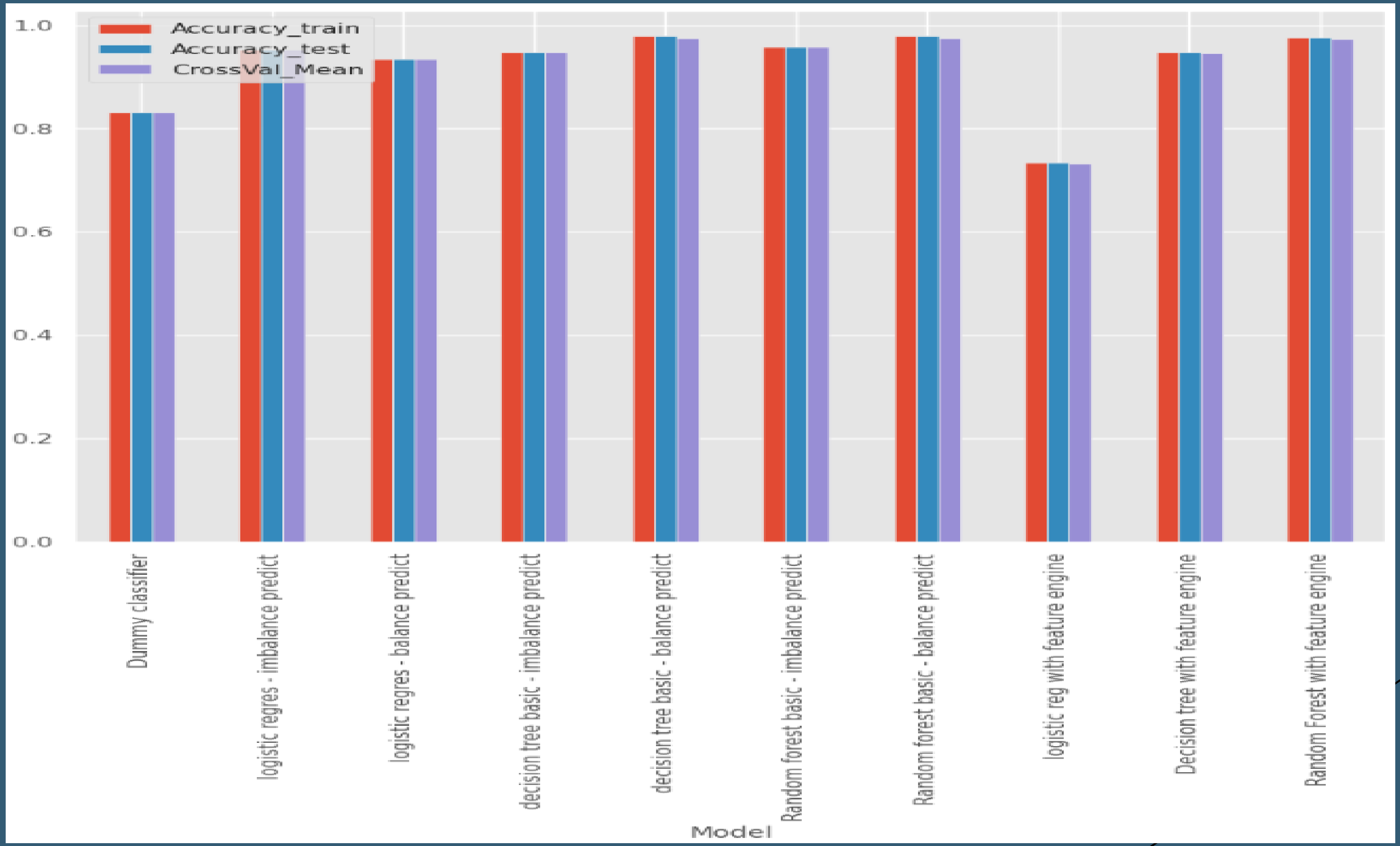
features	importance
device_ip	0.71
device_id	0.07
device_model	0.05
hour	0.03
site_id	0.03
site_domain	0.03
app_id	0.02
day	0.01
dayofweek	0.01
C17	0.01

Model building - Evaluation results

Model	Accuracy_train	recall_train	precision_train	Accuracy_test	recall_test	precision_test	CrossVal_Mean	CrossVal1	CrossVal2	CrossVal3	CrossVal4	CrossVal5
Dummy classifier	0.83	NaN	0.00	0.83	NaN	0.00	0.83	0.83	0.83	0.83	0.83	0.83
logistic regres - imbalance predict	0.95	0.89	0.82	0.95	0.89	0.82	0.95	0.95	0.95	0.95	0.95	0.95
logistic regres - balance predict	0.93	0.94	0.93	0.93	0.94	0.93	0.93	0.93	0.94	0.93	0.93	0.94
decision tree basic - imbalance predict	0.95	0.84	0.86	0.95	0.84	0.86	0.95	0.95	0.95	0.95	0.95	0.95
decision tree basic - balance predict	0.98	0.96	1.00	0.98	0.96	1.00	0.97	0.97	0.98	0.97	0.97	0.97
Random forest basic - imbalance predict	0.96	0.92	0.83	0.96	0.92	0.83	0.96	0.96	0.96	0.96	0.96	0.96
Random forest basic - balance predict	0.98	0.96	1.00	0.98	0.96	1.00	0.97	0.97	0.98	0.97	0.97	0.97
logistic reg with feature engine	0.73	0.73	0.75	0.73	0.73	0.75	0.73	0.73	0.73	0.73	0.73	0.73
Decision tree with feature engine	0.95	0.93	0.97	0.95	0.93	0.97	0.95	0.94	0.95	0.95	0.95	0.95
Random Forest with feature engine	0.98	0.96	0.99	0.98	0.96	0.99	0.97	0.97	0.97	0.97	0.97	0.97

Model building - Evaluation results

If we must select one model, Random forest classifier with feature engineering looks promising and best. although after the feature engineering the training and test results looks same as before feature engineering, but model is very robust with new features and rightly fitted for both training and test dataset. Decision tree and logistic regression classifier seems to have next best model to choose as the accuracy, precision, and recall is good, overall random forest classifier seems doing better with all aspects



Thank You! ☺