feb. 4, 2020

BIG DATA PROJECT

SPOTIFY.

TO SKIP OR NOT TO SKIP?

JANUZZI, TOMASELLO

SPOTIFY

SPOTIFY IS A MUSIC STREAMING PLATFORM FOUNDED IN 2006 BY DANIEL ELK AND MARTIN LORENTZON.

FREE VERSION

- AD-SUPPORTED
- LIMITED FEATURES
- 6 "SKIPS" PER HOUR

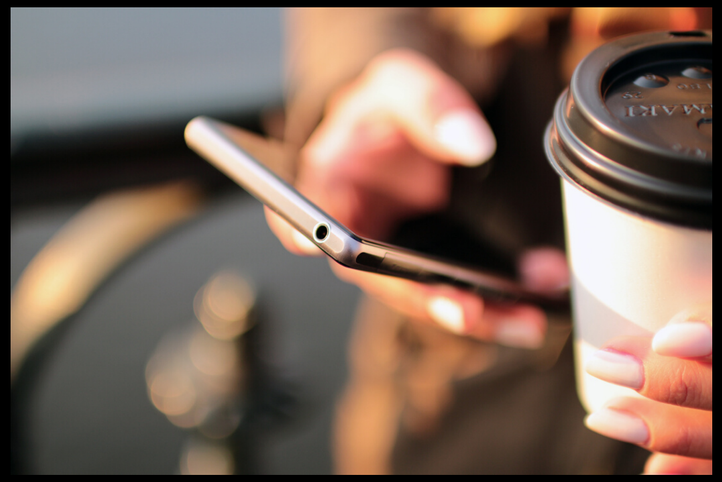
PREMIUM VERSION

- NO ADS
- FULL USE OF FEATURES
- UNLIMITED "SKIPS"

TO SKIP OR NOT TO SKIP?

...

THE GOAL FOR OUR MODELS IS TO PREDICT THE USER BEHAVIOR FOR SKIPPING OR NOT A SONG, USING AS OUR Y THE VARIABLE ‘NOT_SKIPPED’, UNDERSTANDING IF AND HOW MUCH OUR OUTPUT VARIABLE IS DEPENDENT ON THE FEATURES INCLUDED IN OUR DATASET.



dataset.

OUR INITIAL DATASET IS COMPOSED OF 99999 OBSERVATIONS FOR 21 VARIABLES:

session_id
session_length
track_id_clean
skip_1
skip_2
skip_3
not_skipped
context_switch
no_pause_before_play
short_pause_before_play
long_pause_before_play

hist_user_behavior_n_seekfwd
hist_user_behavior_n_seekback
hist_user_behavior_is_shuffle
hour_of_day
date
premium
context_type
hist_user_behavior_reason_start
hist_user_behavior_reason_end

dataset.

WE FOUND THE DATASET FOR OUR PROJECT ON AIC (ARTIFICIAL INTELLIGENCE CROWD)

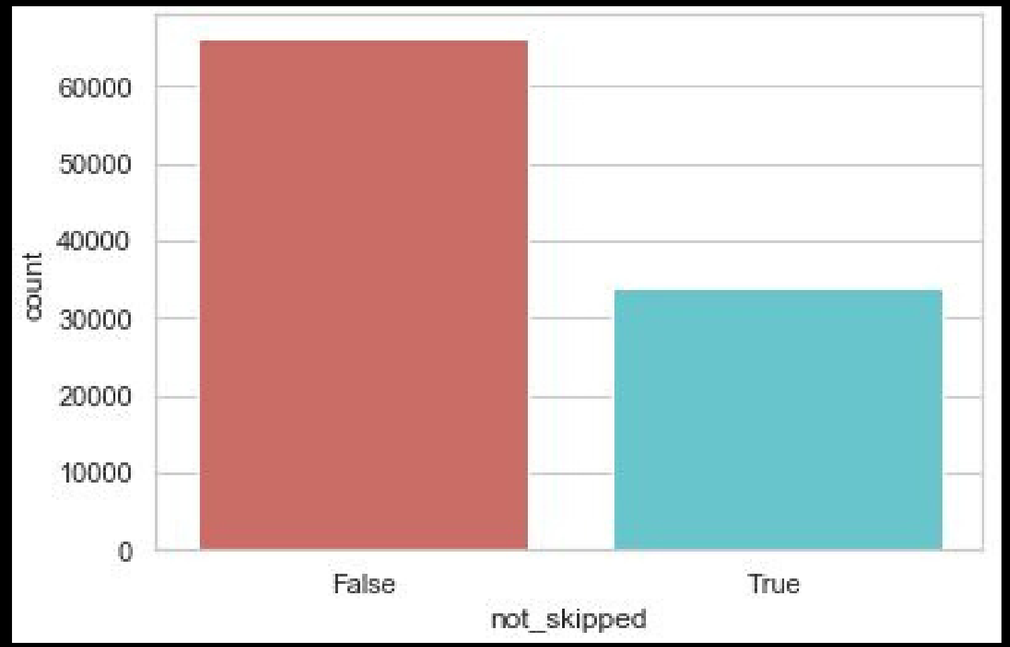
session_position	session_id	session_length	track_id_clean	skip_1	skip_2	skip_3	not_skipped	context_switch	no_pause_before_play	short_pause_before
1	1_0000015a-8bee-425e-bc96-25042a1a1cab	20	t_f56065b2-b26f-4080-a121-ff3bf5d25fd6	False	False	False	True	0	0	
2	1_0000015a-8bee-425e-bc96-25042a1a1cab	20	t_447a276d-5b3d-44cf-bcbe-834ad91e7b72	False	False	True	False	0	1	
3	1_0000015a-8bee-425e-bc96-25042a1a1cab	20	t_07a8a863-2f51-4e64-ae3c-2fac451a0651	False	False	True	False	0	1	
4	1_0000015a-8bee-425e-bc96-25042a1a1cab	20	t_d2a13f32-05a8-4f7b-8633-4ed8a74ca560	False	False	True	False	0	1	
5	1_0000015a-8bee-425e-bc96-25042a1a1cab	20	t_0368bff8-85c0-4162-b1ca-18c526b14d3a	False	False	True	False	0	1	

data exploration.

EXPLORING OUR DATA GAVE US A BETTER UNDERSTANDING OF THE STRUCTURE OF OUR DATA

THE PERCENTAGE OF SONGS SKIPPED IS 66.127

THE PERCENTAGE OF SONGS FULLY PLAYED IS 33.873



models.

SUPERVISED MACHINE LEARNING

LOGISTIC REGRESSION

WE APPLIED THIS MODEL TO OUR DATASET OF CATEGORICAL DATA. WE ALSO PERFORMED RIDGE AND LASSO TO UNDERSTAND IF WE COULD HAVE IMPROVED OUR RESULTS APPLYING PENALTIES

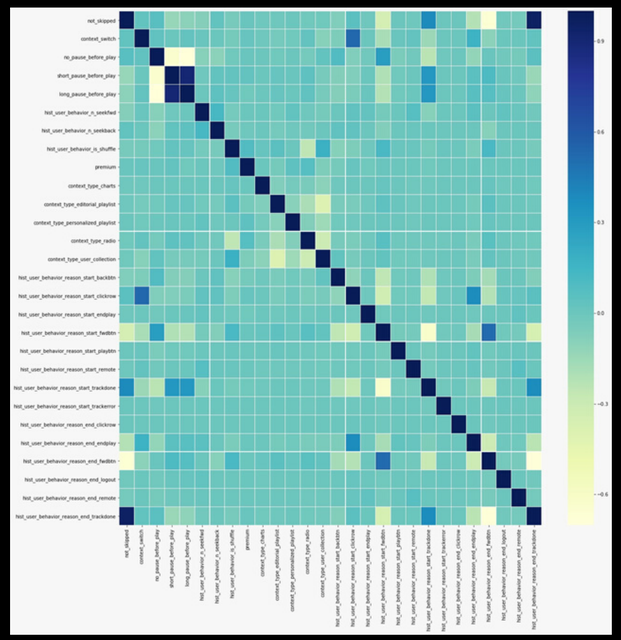
REGRESSION TREE

WE APPLIED THE DECISION TREE MODEL TO VISUALIZE OUR VARIABLES, WHILE OUR RANDOM FOREST BUILT MULTIPLE DECISION TREES AND MERGED THEM TOGETHER TO GET A MORE ACCURATE AND STABLE PREDICTION.

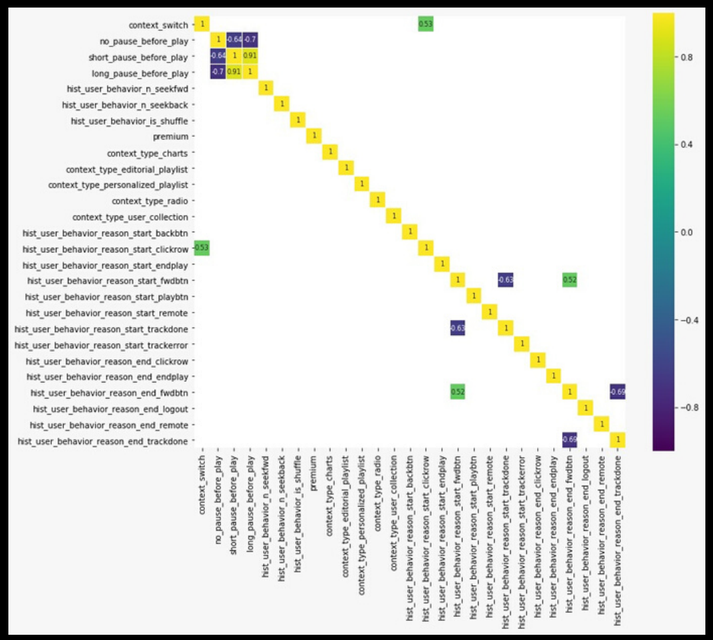
logistic regression.

LIKE ALL REGRESSION ANALYSES, THE LOGISTIC REGRESSION IS A PREDICTIVE ANALYSIS. IT IS USED TO DESCRIBE DATA AND TO EXPLAIN THE RELATIONSHIP BETWEEN ONE DEPENDENT BINARY VARIABLE AND THE INDEPENDENT VARIABLES.

heatmap for correlation



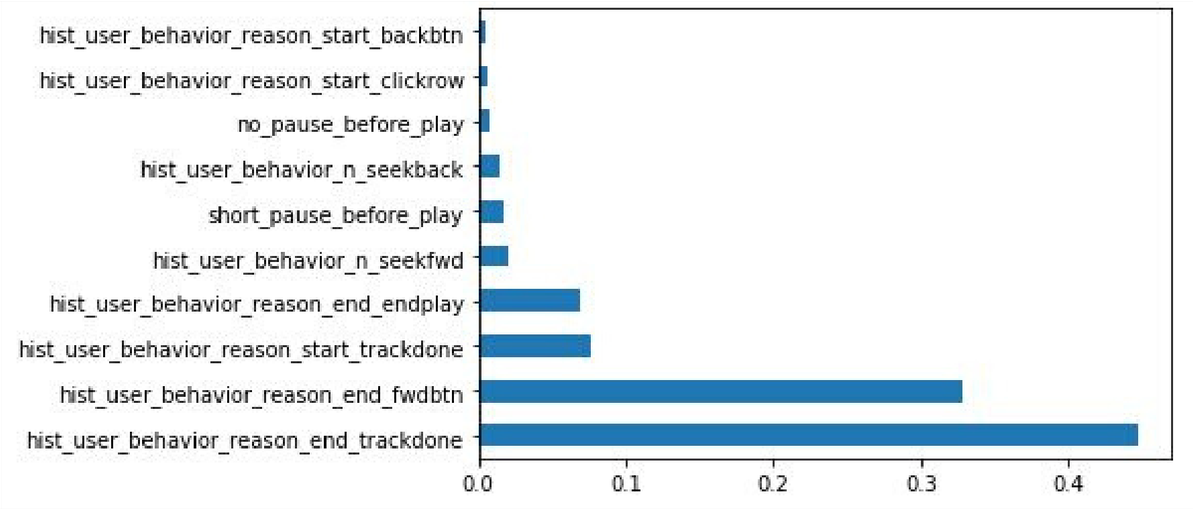
heatmap for multicollinearity



logistic regression.

IN ORDER TO KEEP JUST THE RELEVANT VARIABLES WE PERFORMED FEATURE SELECTION:

Built-in feature importance

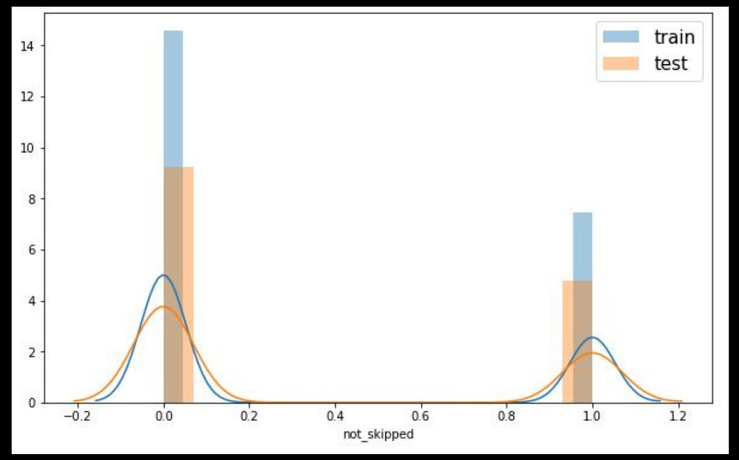


VIF: variance inflation factor

constant	25.53
hist_user_behavior_reason_start_backbtn	1.45
hist_user_behavior_reason_start_clickrow	1.44
no_pause_before_play	1.77
hist_user_behavior_n_seekback	1.06
short_pause_before_play	1.91
hist_user_behavior_n_seekfwd	1.04
hist_user_behavior_reason_end_endplay	2.55
hist_user_behavior_reason_start_trackdone	1.73
hist_user_behavior_reason_end_fwdbtn	4.85
hist_user_behavior_reason_end_trackdone	4.57

dtype: object

logistic regression.



SPLIT IN TRAIN AND TEST

SUMMARY OF THE MODEL
ALGORYTHM CONVERGED

Results: Logit

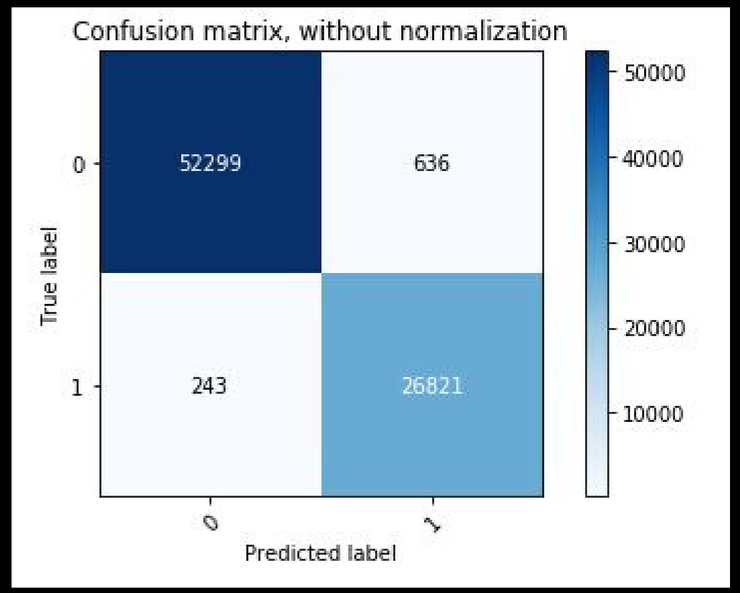
Model:	Logit	Pseudo R-squared:	0.890
Dependent Variable:	not_skipped	AIC:	11309.4721
Date:	2020-02-01 16:50	BIC:	11402.3698
No. Observations:	79999	Log-Likelihood:	-5644.7
Df Model:	9	LL-Null:	-51192.
Df Residuals:	79989	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	10.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
hist_user_behavior_reason_start_backbtn	-2.0669	0.1006	-20.5446	0.0000	-2.2641	-1.8697
hist_user_behavior_reason_start_clickrow	-1.4094	0.1040	-13.5501	0.0000	-1.6133	-1.2056
no_pause_before_play	-2.4347	0.0663	-36.6998	0.0000	-2.5647	-2.3046
hist_user_behavior_n_seekback	1.2978	0.0438	29.6180	0.0000	1.2119	1.3837
short_pause_before_play	-2.7393	0.1050	-26.0849	0.0000	-2.9452	-2.5335
hist_user_behavior_n_seekfwd	-7.5939	0.1243	-61.0839	0.0000	-7.8376	-7.3503
hist_user_behavior_reason_end_endplay	-2.2394	0.1150	-19.4644	0.0000	-2.4649	-2.0139
hist_user_behavior_reason_start_trackdone	-1.3349	0.0758	-17.6070	0.0000	-1.4835	-1.1863
hist_user_behavior_reason_end_fwdbtn	-3.8220	0.1030	-37.0897	0.0000	-4.0239	-3.6200
hist_user_behavior_reason_end_trackdone	7.2307	0.0908	79.5987	0.0000	7.0527	7.4087

logistic regression.

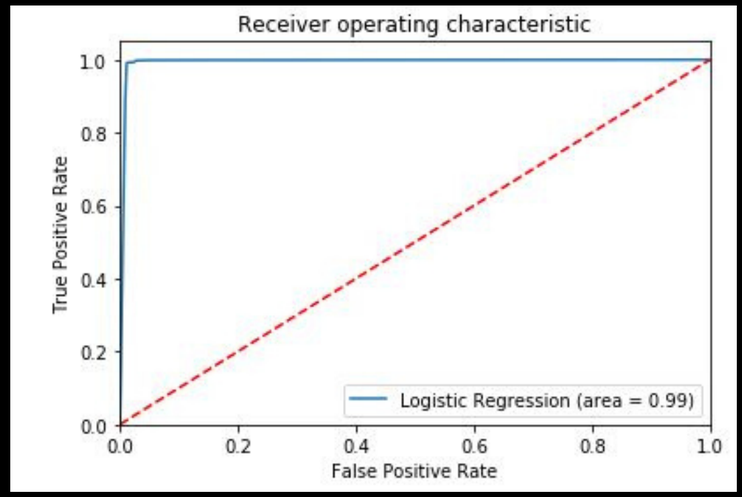
RESULTS ON THE TRAINING SET

Confusion matrix



	precision	recall	f1-score	support
0	1.00	0.99	0.99	52935
1	0.98	0.99	0.98	27064
accuracy			0.99	79999
macro avg	0.99	0.99	0.99	79999
weighted avg	0.99	0.99	0.99	79999

ROC curve

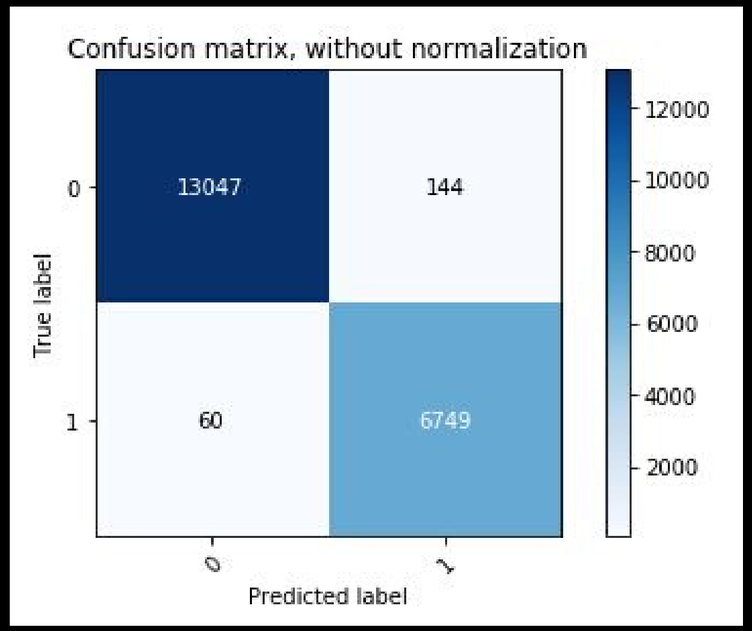


pg. 10

logistic regression.

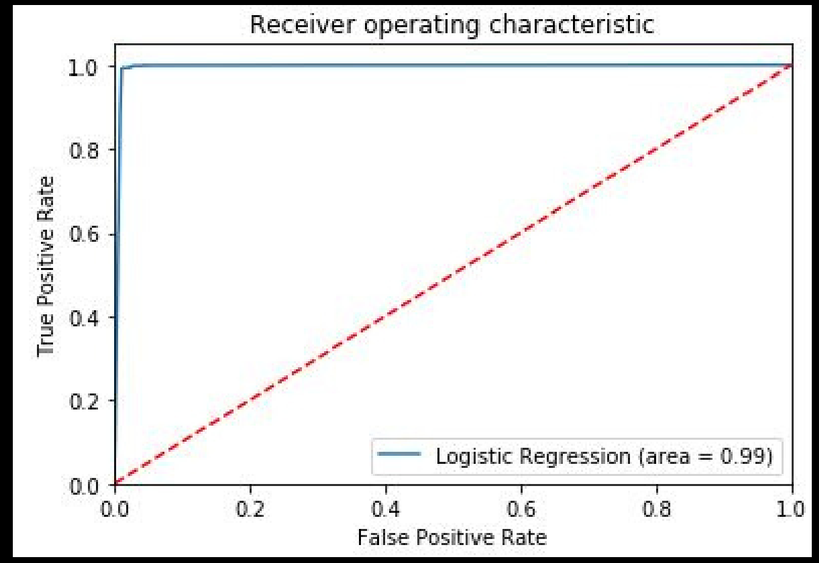
RESULTS ON THE TEST SET

Confusion matrix



	precision	recall	f1-score	support
0	1.00	0.99	0.99	13191
1	0.98	0.99	0.99	6809
accuracy			0.99	20000
macro avg	0.99	0.99	0.99	20000
weighted avg	0.99	0.99	0.99	20000

ROC curve



pg. 11

logistic regression.

WE HAVE APPLIED RIDGE REGRESSION AND LASSO TO UNDERSTAND HOW PENALTIES COULD INFLUENCE OUR ACCURACY:

	Penalty	Train_Accuracy	Test_Accuracy	Train_Precision	Test_Precision	Train_Recall	Test_Recall
0	none	0.989000	0.98980	0.986104	0.987298	0.989476	0.990100
0	l1	0.989025	0.98975	0.986132	0.987226	0.989504	0.990062
0	l2	0.989012	0.98980	0.986114	0.987298	0.989494	0.990100

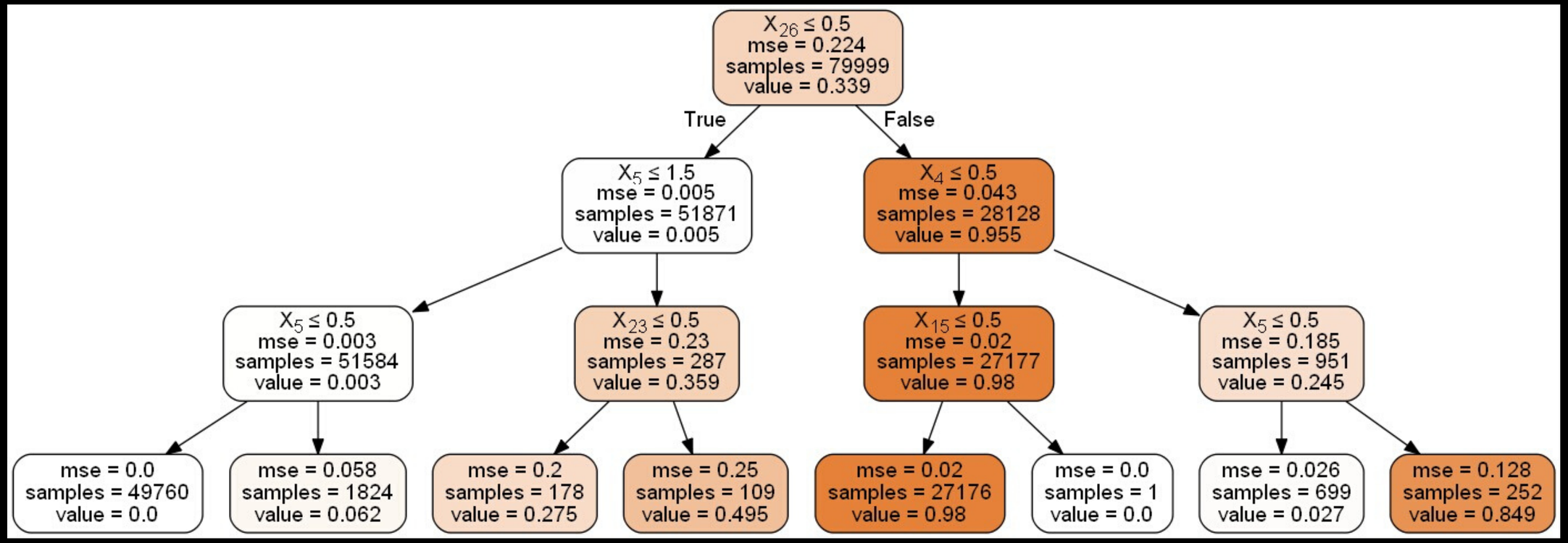
AS WE CAN SEE FROM THE RESULTS,NEITHER WITH LASSO NOR WITH RIDGE WE SEE A RELEVANT CHANGE IN ACCURACY.

regression tree.

THE TREE STRUCTURE IN THE DECISION MODEL HELPS IN DRAWING A CONCLUSION FOR ANY PROBLEM WHICH IS MORE COMPLEX IN NATURE. TREE BASED LEARNING ALGORITHMS ARE CONSIDERED TO BE ONE OF THE BEST AND MOSTLY USED SUPERVISED LEARNING METHODS. TREE BASED METHODS EMPOWER PREDICTIVE MODELS WITH HIGH ACCURACY, STABILITY AND EASE OF INTERPRETATION

WE USED K_FOLD CROSS VALIDATION BEFORE PERFORMING THE REGRESSION. K-FOLD CV IS WHERE A GIVEN DATA SET IS SPLIT INTO A K NUMBER OF SECTIONS/FOLDS WHERE EACH FOLD IS USED AS A TESTING SET AT SOME POINT

regression tree.



regression tree.

AS OUR VARIABLES IN OUR TREE WE HAVE
X26 - HIST_USER_BEHAVIOR_REASON_END_TRACKDONE
X5 -HIST_USER_BEHAVIOR_N_SEEKBACK
X4 - HIST_USER_BEHAVIOR_N_SEEKFWD
X15 - HIST_USER_BEHAVIOR_REASON_START_ENDPLAY
X23 - HIST_USER_BEHAVIOR_REASON_END_FWDBTN

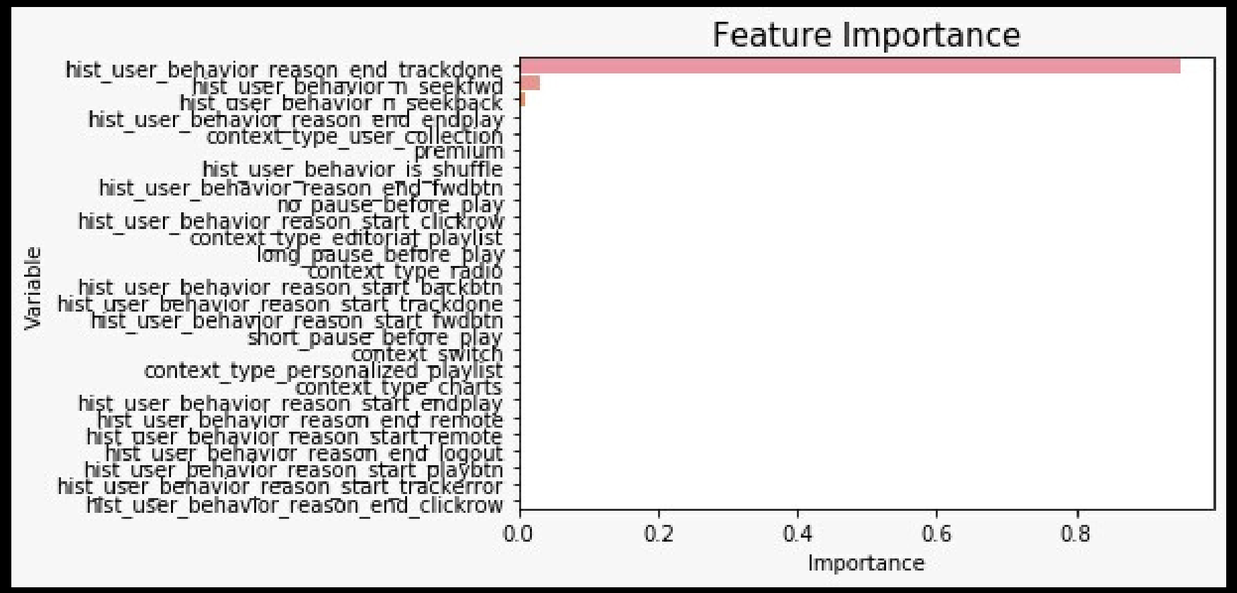
```
DecisionTreeRegressor(criterion='mse', max_depth=3, max_features=None,
                      max_leaf_nodes=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      presort=False, random_state=0, splitter='best')


Root Mean Squared Error on train set: 0.09882687922411729
Root Mean Squared Error on test set: 0.10529649555232867
Mean of y_train: 0.3387333873338733
```

PERFORMANCE OF THE DECISION TREE

random forest.

RANDOM FOREST BUILDS MULTIPLE DECISION TREES AND MERGES THEM TOGETHER TO GET A MORE ACCURATE AND STABLE PREDICTION. WE USE AS OUR N_ESTIMATER = 50



random forest.

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
                      max_features='auto', max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, n_estimators=50,
                      n_jobs=None, oob_score=False, random_state=0, verbose=0,
                      warm_start=False)


Root Mean Squared Error on train set: 0.08721280950953403
Root Mean Squared Error on test set: 0.1051392623351552
Mean of y_train: 0.3387333873338733
```

WE CAN OBSERVE THAT THE ACCURACY OF THE RANDOM FOREST IS ALMOST THE SAME OF THE LOGISTIC REGRESSION.

```
Accuracy of RF classifier on training set: 0.99
Accuracy of RF classifier on test set: 0.99
```

conclusions.

best model.

IN THE END, GIVEN THE RESULTS OF THE LOGISTIC REGRESSION AND THE RANDOM FOREST, WE HAVE SEEN THAT THE ACCURACY IN BOTH OF THE APPROACHES STAYS ALMOST THE SAME,WITH A VARIATION OF +0,01 IN THE RANDOM TREE.

ANYWAY WITH THE DECISION TREE WE CAN VISUALIZE BETTER OUR RESULTS, AND UNDERSTAND WHICH VARIABLES AFFECT ON THE RESULT.

conclusions.

relevant features.

- HIST_USER_BEHAVIOR_REASON_END_TRACKDONE
- HIST_USER_BEHAVIOR_REASON_END_FWDBTN
- HIST_USER_BEHAVIOR_N_SEEKBACK
- HIST_USER_BEHAVIOR_N_SEEKFWD
- HIST_USER_BEHAVIOR_REASON_START_ENDPLAY

BIG DATA PROJECT

THANK YOU!

Maira Januzzi

Beatrice Tomasello

MIBE
Master in International Business & Entrepreneurship

feb. 4, 2020