RETAILROCKET: E-COMMERCE USER PURCHASE PREDICTION

NURABIDAH JAMIL
5 AUGUST 2021

BACKGROUND

An e-commerce company wishes to understand visitors behaviour on its site and the factors that drive these site visitors to make a purchase. This is so that they can better target ads and improve future online sales.

PROBLEM STATEMENT

Using Retailrocket e-commerce dataset from Kaggle, the goal is to build and compare different binary classification models that would best predict whether a user will make a purchase on the e-commerce site.

Table of Contents

DATA

EVENT

	timestamp	visitorid	event	itemid	transactionid
0	1433221332117	257597	view	355908	NaN
1	1433224214164	992329	view	248676	NaN

CATEGORY TREE

	categoryid	parentid
0	1016	213.0
1	809	169.0

ITEM PROPERTIES

	timestamp	itemid	property	value
0	1435460400000	460429	categoryid	1338
1	1441508400000	206783	888	1116713 960601 n277.200

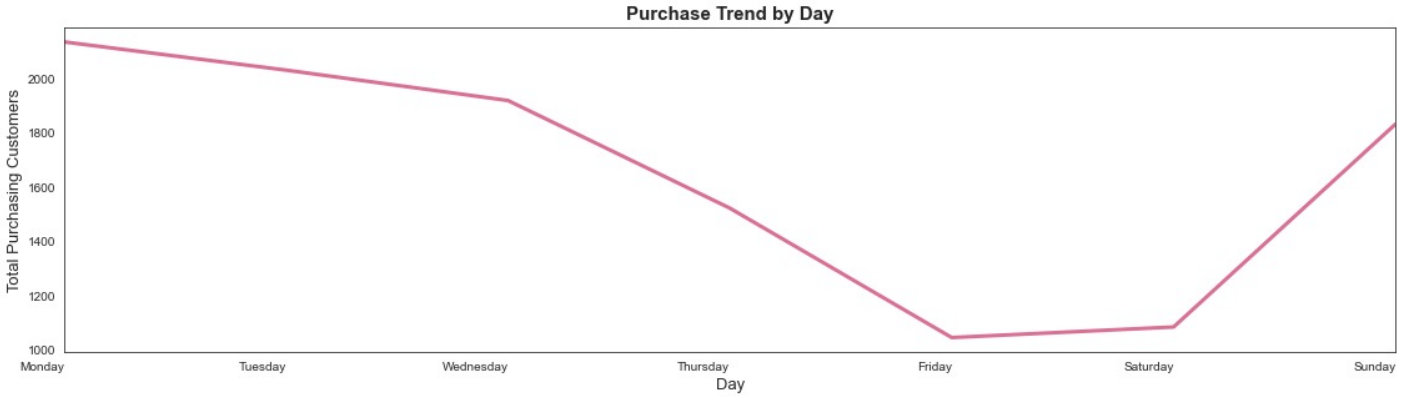
ITEM PROPERTIES

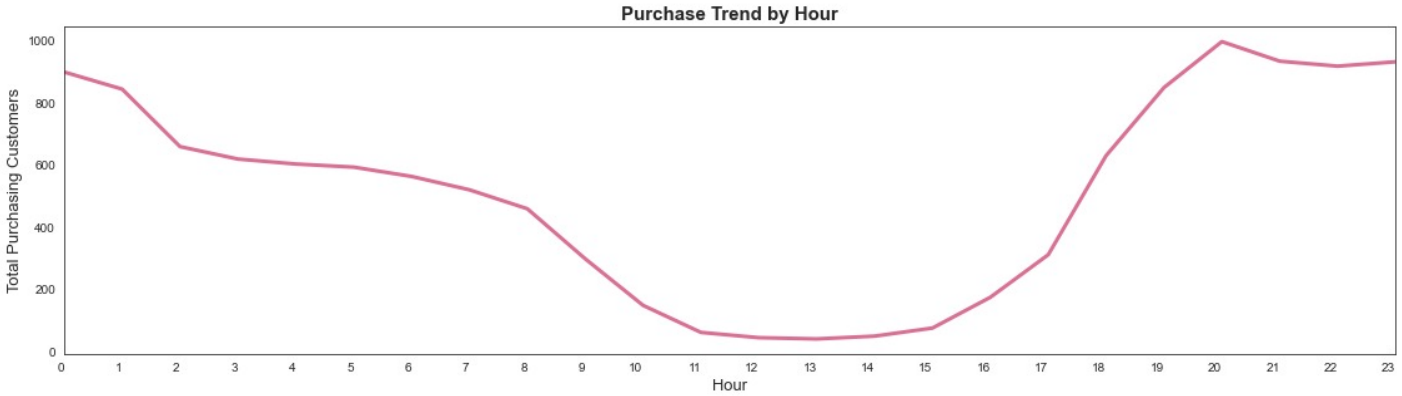
	timestamp	itemid	property	value
0	1433041200000	183478	561	769062
1	1439694000000	132256	976	n26.400 1135780

FINAL DATASET

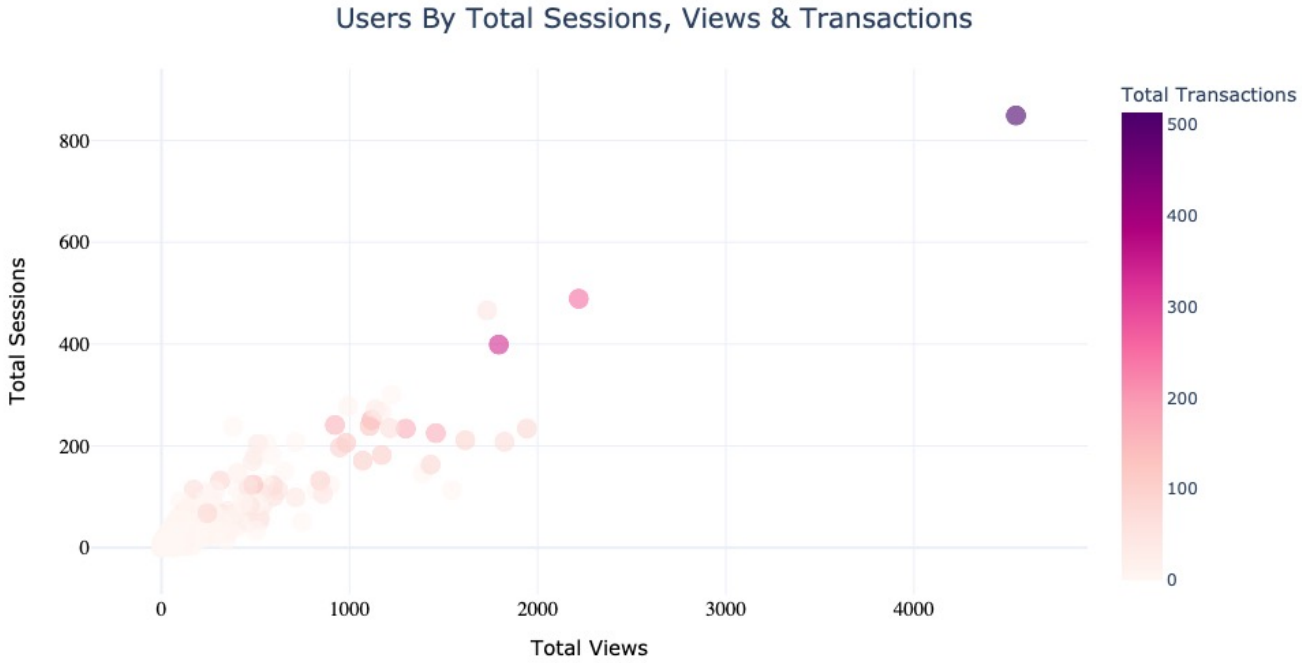
	visitorid	start_session	end_session	session_duration_m	cart	transaction	view	cart_cat	transaction_cat	view_cat
0	0	2015-09-11 23:49:49	2015-09-11 23:55:17	5.47	[]	[]	[285930, 357564, 67045]	[]	[]	[1188, 256, 333]
1	1	2015-08-13 20:46:06	2015-08-13 20:46:06	0.00	[]	[]	[72028]	[]	[]	[1192]
2	2	2015-08-07 20:51:44	2015-08-07 21:20:57	29.22	[]	[]	[325215, 325215, 259884, 216305, 342816, 34281...	[]	[]	[299, 299, 299, 299, 444, 444, 299, 299]
3	3	2015-08-01 10:10:35	2015-08-01 10:10:35	0.00	[]	[]	[385090]	[]	[]	[1171]
4	5	2015-07-17 04:45:56	2015-07-17 04:45:56	0.00	[]	[]	[61396]	[]	[]	[646]

PURCHASE TRENDS



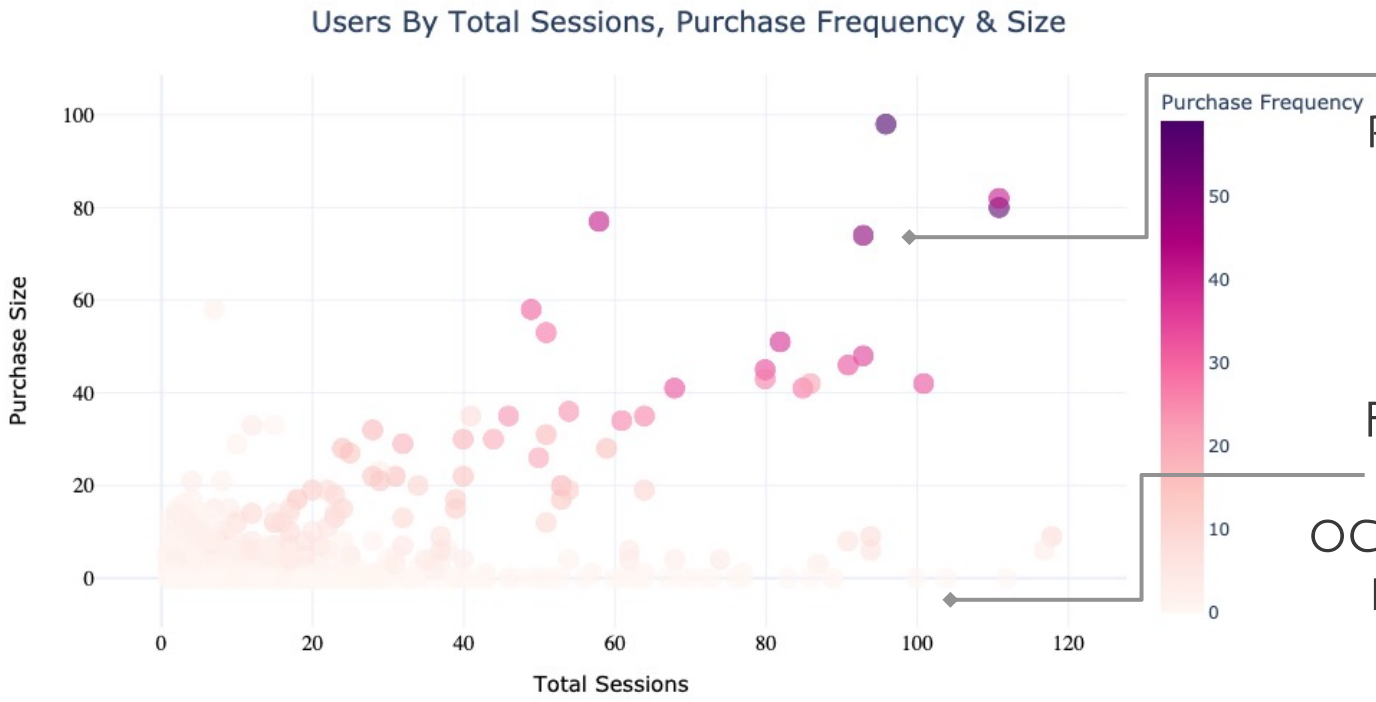


ABNORMAL USERS



- **DATA REPORTING PERIOD**
  Less than 4 months
- **ABNORMAL USERS**
  More than 150 total sessions and/or 150 total transactions

USERS



VISIT & PURCHASE OFTEN

FREQUENT VISIT & OCCASIONAL PURCHASE

User groups

K-means clustering

- Total clusters: 7

Clusters Example:

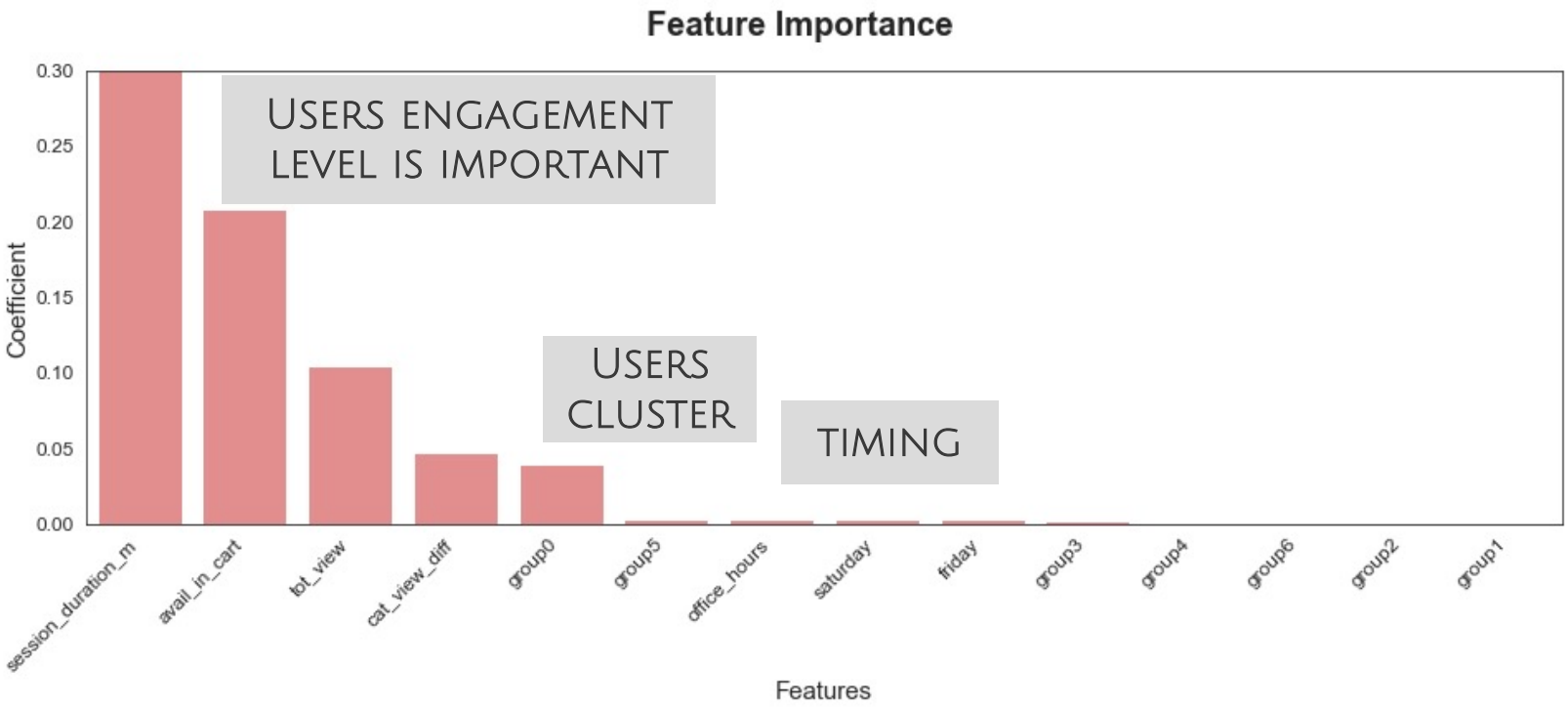
Group No.	Group 3	Group 1	Group 0
Visit Frequency	Almost daily	Almost daily	Once
Purchase Frequency	2-3 times a month	1-2 times in total	None yet

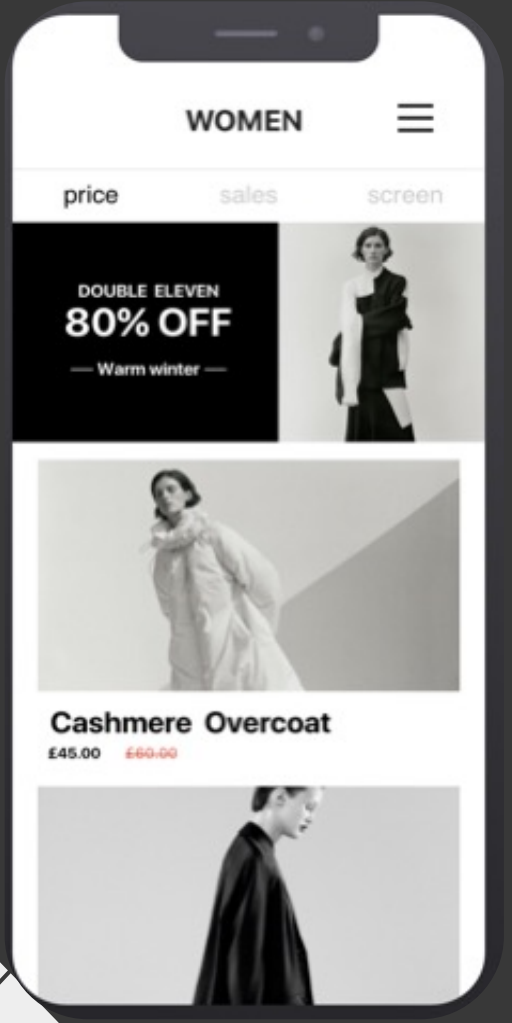
MODEL SUMMARY

MODEL	TRAIN			TEST		
	PRECISION	RECALL	F1	PRECISION	RECALL	F1
Random Forest	0.27	0.90	0.41	0.14	0.78	0.23
Logistic Regression	0.11	0.87	0.19	0.11	0.82	0.19
XG Boost	0.13	0.93	0.24	0.07	0.97	0.13
ADA Boost	0.11	0.94	0.19	0.05	0.98	0.10

* Best Parameters: {'rf__max_depth': 30, 'rf__max_features': 0.6, 'rf__min_samples_leaf': 2, 'rf__min_samples_split': 3, 'rf__n_estimators': 200}

MODEL SELECTED





Conclusions & Recommendations

- High false positive cases and model is not robust
- Timing, user engagement and item availability are important
- Avoid launching short-term campaigns at certain time
- Improve users engagement on the site

Conclusions & Recommendations

- Implement different marketing strategies for each user group:

Group 3

- Loyalty programs
- Market new products

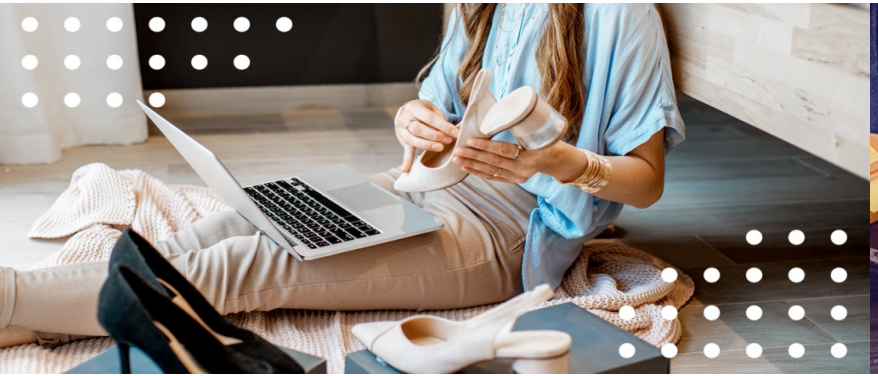
Group 1

- Target them for short-term campaigns like flash sales

Group 0

- Aggressive price incentives

Next steps



Incorporate other data

- The marketing channel the users come from
- Purchase value

Refresh customer segmentation

- Incorporate missing key information

Other models

- Neural Network

THANK YOU

Data dictionary

Variable	Description
group0	A dummy variable to indicate whether the users belong to group 0.
group1	A dummy variable to indicate whether the users belong to group 1.
group2	A dummy variable to indicate whether the users belong to group 2.
group3	A dummy variable to indicate whether the users belong to group 3
group4	A dummy variable to indicate whether the users belong to group 4.
group5	A dummy variable to indicate whether the users belong to group 5.
group6	A dummy variable to indicate whether the users belong to group 6.
friday	A dummy variable to indicate whether the session occurs on a Friday.
saturday	A dummy variable to indicate whether the session occurs on a Saturday.
office_hours	A dummy variable to indicate whether the session occurs within 8 am to 5 pm.
avail_in_cart	The percentage of items in the cart that are available.
tot_view	The total number of items viewed during a particular session.
cat_view_diff	A variable to show how different are the items being viewed (scaled between 0.005 and 1 , with 1 indicating that all items viewed are different).
session_duration_m	The duration during which there are regular active interactions coming from a user on the website.