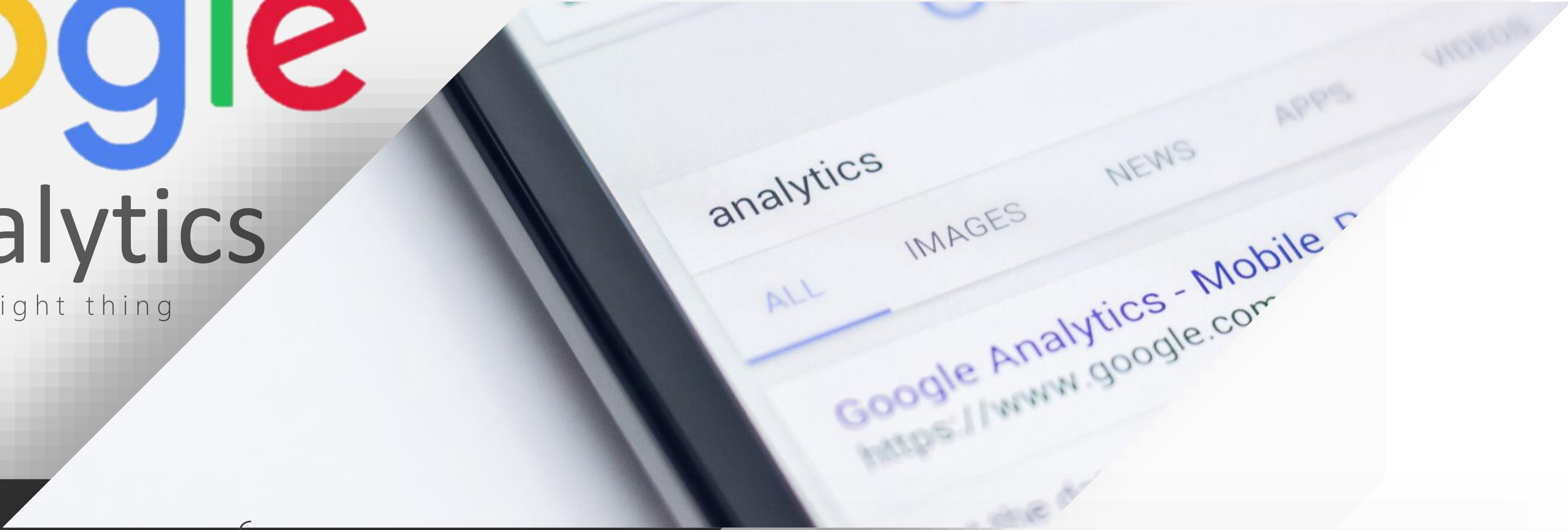
Google

Analytics

Do the right thing



Google Merchandise Store

Data Mining Principles

| Carrie Lu | Daniela Matinho |
| Hanna Kerr | Yuling GU |

THE UNIVERSITY OF
CHICAGO

Google

Do the right thing

AGENDA

01 Industry Overview & Business Problem

02 Data Preparation

03 Data Exploration & Feature Engineering

04 Modeling

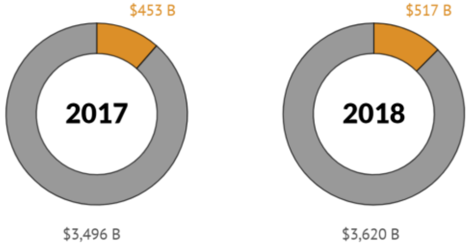
05 Results

06 Lessons learned & Future Work

07 Sources

THE UNIVERSITY OF CHICAGO

Industry Overview



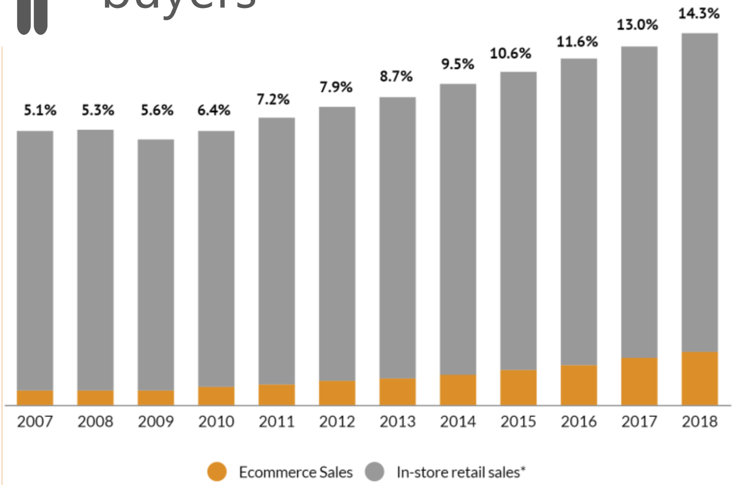
Sales in retail vs E-commerce

2017 – 2021

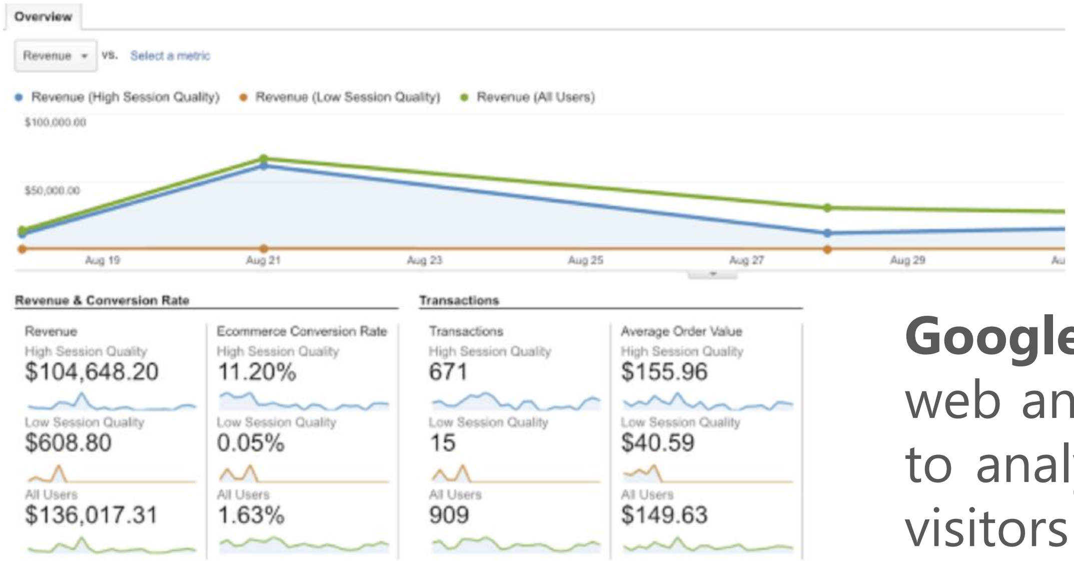
Retail e-commerce sales account for $2.3 trillion in 2017 and it is projected to grow to $4.88 trillion by 2021

2020

Over 2 billion digital buyers



Penetration rate retail vs ecommerce


Overview
Revenue ▾ vs. Select a metric
● Revenue (High Session Quality) ● Revenue (Low Session Quality) ● Revenue (All Users)
$100,000.00
$50,000.00
Aug 19
Aug 21
Aug 23
Aug 25
Aug 27
Aug 29
Au

Revenue & Conversion Rate

Revenue
High Session Quality
$104,648.20
Low Session Quality
$608.80
All Users
$136,017.31

Ecommerce Conversion Rate
High Session Quality
11.20%
Low Session Quality
0.05%
All Users
1.63%

Transactions

Transactions
High Session Quality
671
Low Session Quality
15
All Users
909

Average Order Value
High Session Quality
$155.96
Low Session Quality
$40.59
All Users
$149.63

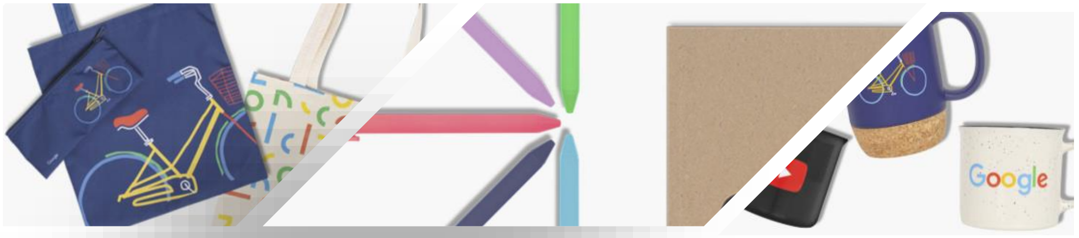
Google Analytics is a web analytics service that allows you to analyze in-depth detail about the visitors on your website.

THE UNIVERSITY OF
CHICAGO

Google Store

Do the right thing

The 80/20 rule has proven true for many businesses, only a small percentage of customers produce most of the revenue. Understanding how much each customer spends will allow companies to place actionable operations to better allocate their marketing budgets.



Problem Statement

Predict the natural log of the sum of all transactions per user

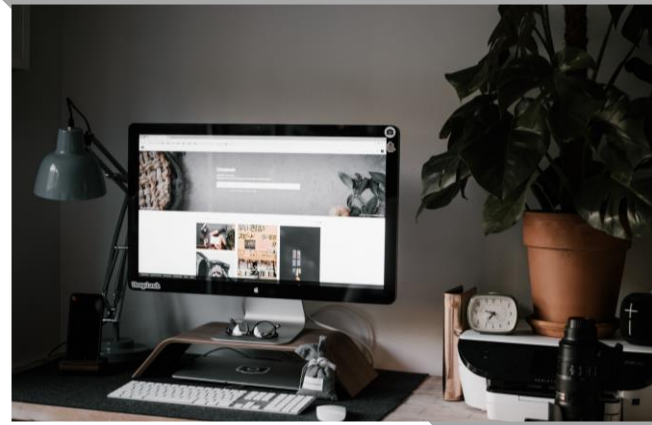
$$y_{user} = \sum_{i=0}^{n} transaction\ user_i \qquad target_{user} = \ln(y_{user} + 1)$$

THE UNIVERSITY OF CHICAGO

Data Preparation

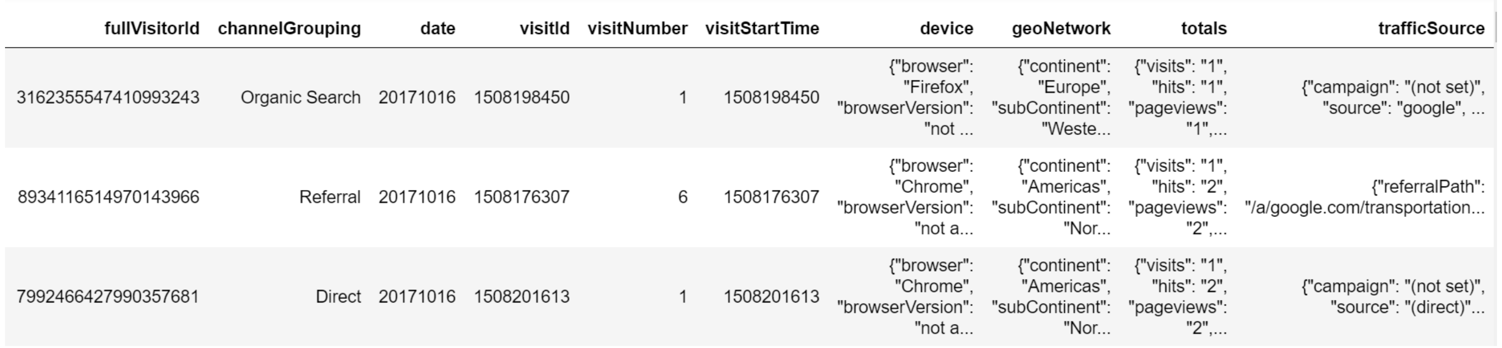
Dataset Columns

- ❑ fullVisitorId
- ❑ channelGrouping
- ❑ date
- ❑ visitId
- ❑ visitNumber
- ❑ visitStartTime



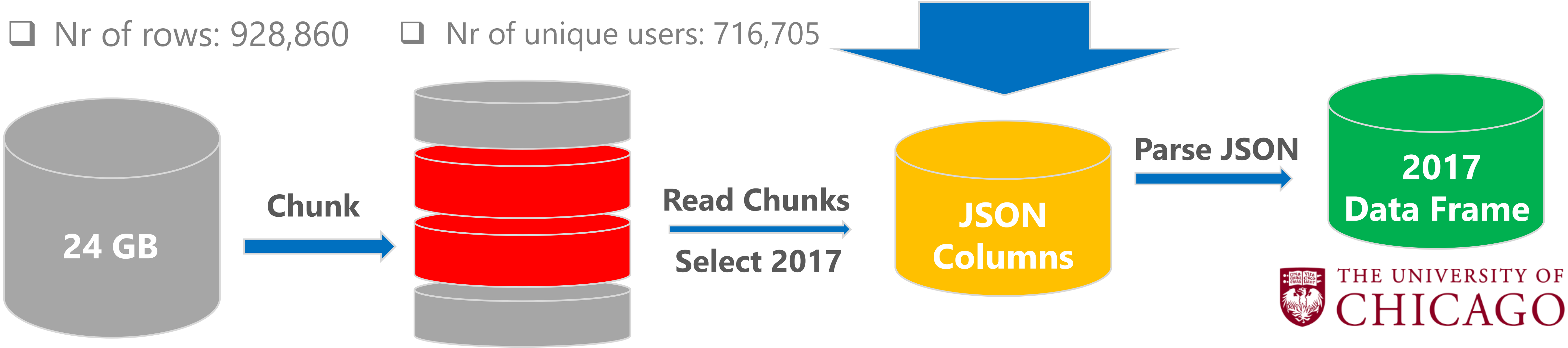
Jason Columns

- ❑ device
- ❑ geoNetwork
- ❑ totals
- ❑ trafficSource



fullVisitorId	channelGrouping	date	visitId	visitNumber	visitStartTime	device	geoNetwork	totals	trafficSource
3162355547410993243	Organic Search	20171016	1508198450	1	1508198450	{"browser": "Firefox", "browserVersion": "not ...	{"continent": "Europe", "subContinent": "Weste...	{"visits": "1", "hits": "1", "pageviews": "1",...	{"campaign": "(not set)", "source": "google", ...
8934116514970143966	Referral	20171016	1508176307	6	1508176307	{"browser": "Chrome", "browserVersion": "not a...	{"continent": "Americas", "subContinent": "Nor...	{"visits": "1", "hits": "2", "pageviews": "2",...	{"referralPath": "/a/google.com/transportation...
7992466427990357681	Direct	20171016	1508201613	1	1508201613	{"browser": "Chrome", "browserVersion": "not a...	{"continent": "Americas", "subContinent": "Nor...	{"visits": "1", "hits": "2", "pageviews": "2",...	{"campaign": "(not set)", "source": "(direct)"...

- ❑ Nr of rows: 928,860
- ❑ Nr of unique users: 716,705

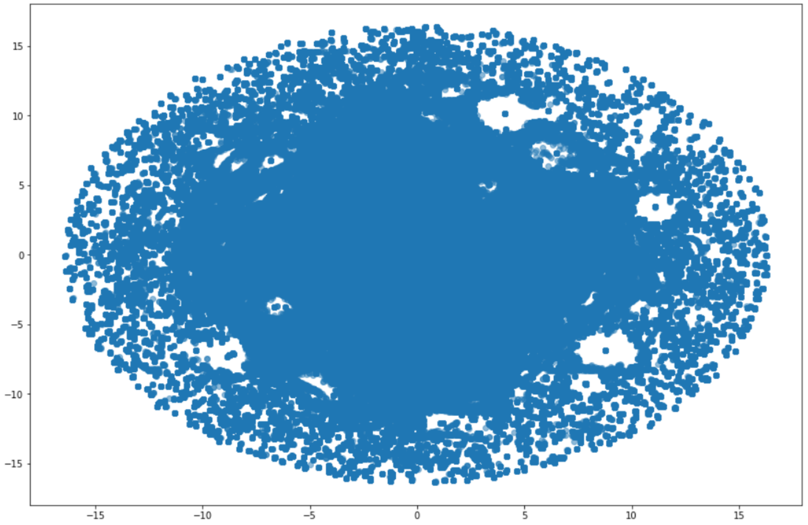


THE UNIVERSITY OF CHICAGO

Data Exploration & Feature Engineering

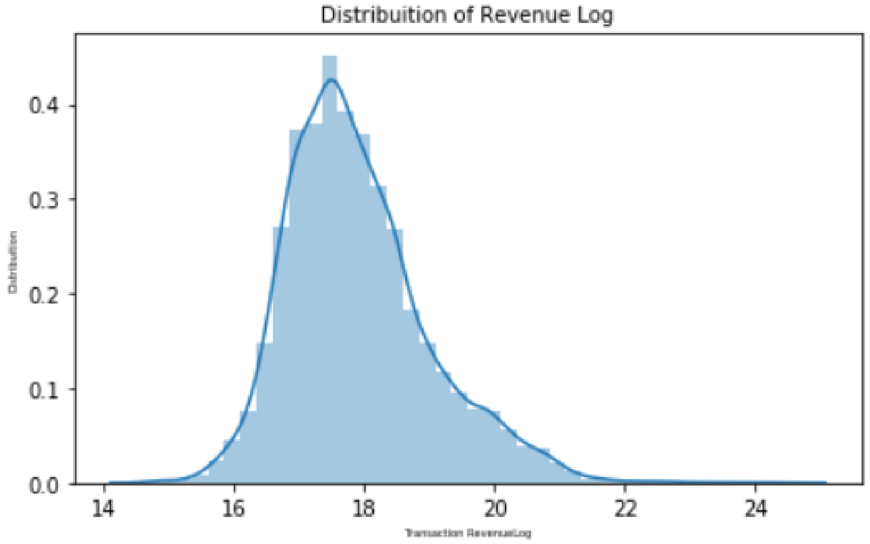
Tsne

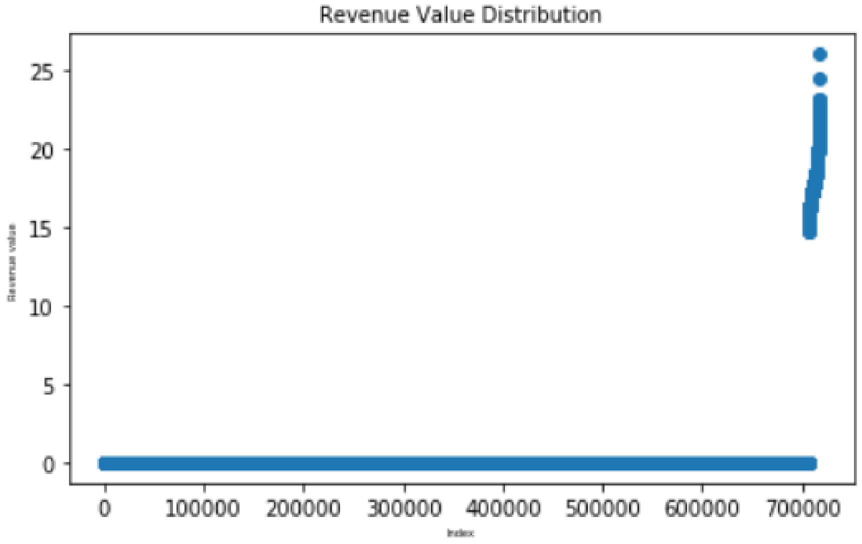
- □ Majority of transactions share similar attributes



Overview of total transaction revenue:

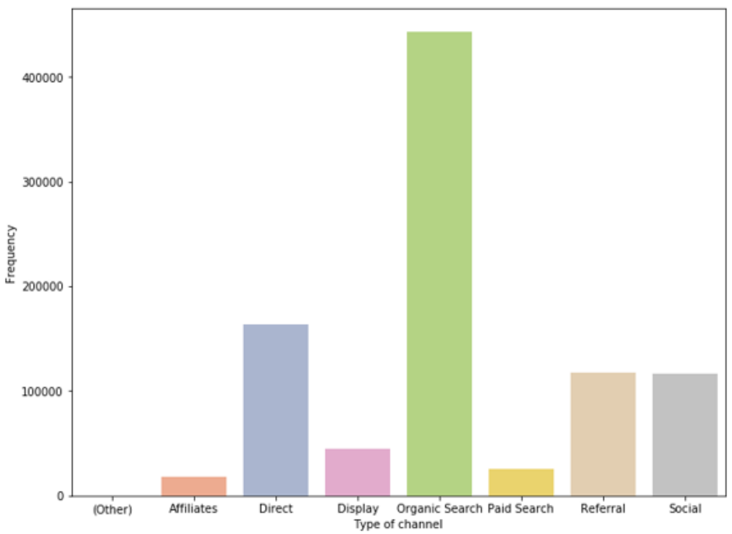
- □ For Log of revenue is slightly **skewed to right**
- □ Only **1.2% of the transactions** contribute to total revenue

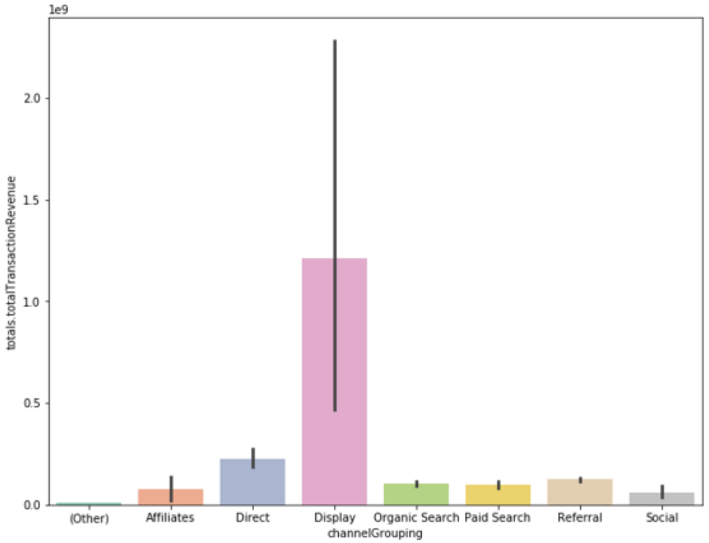


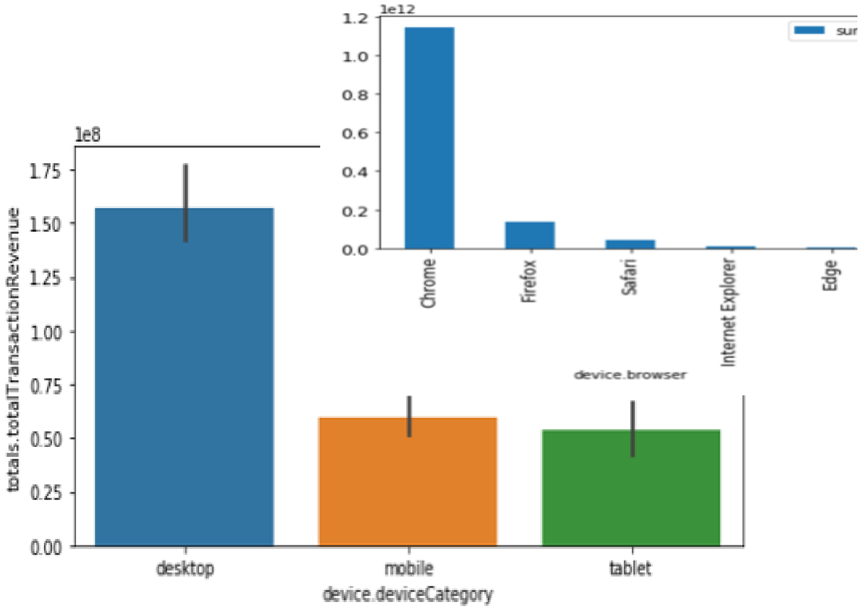


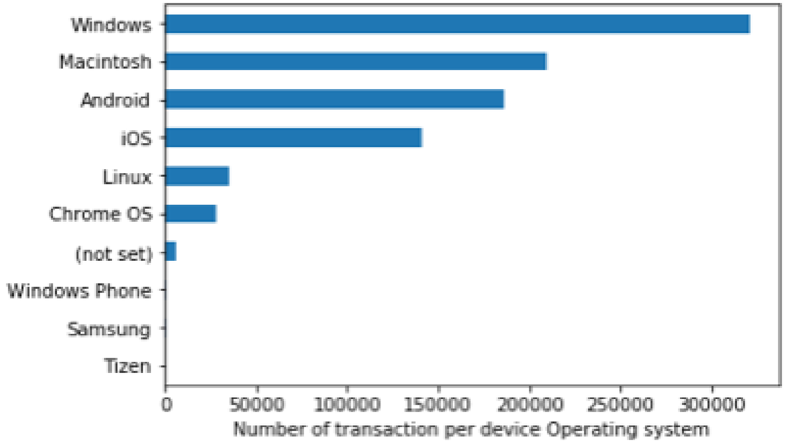
Data Exploration & Feature Engineering

- □ **Group channel:** Most common channel to access GStore: Organic Search; Direct, Referral and Social Media
- □ **Display:** the channel with the highest contribution in terms of revenue
- □ **Operating System:** the first 4 options represent more than 90% of the revenue generated
- □ **Device:** Desktop is the most used device
- □ **Browser:** Chrome is the most used







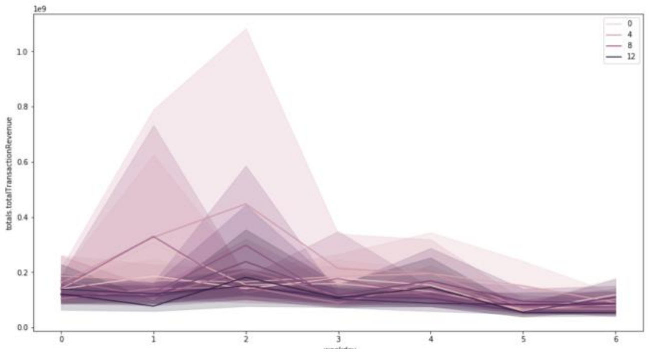


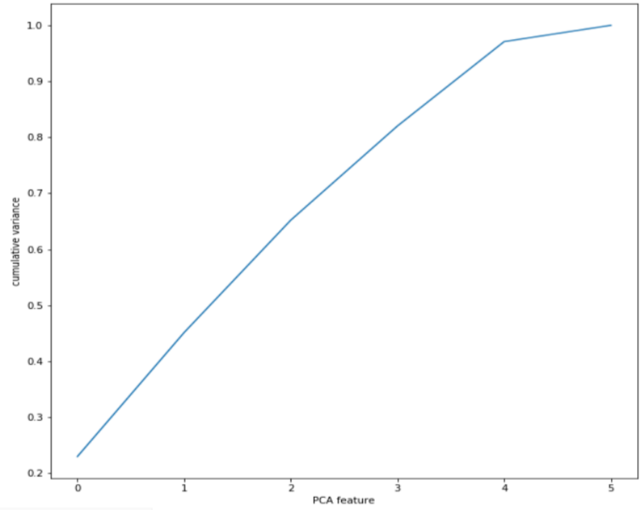
THE UNIVERSITY OF CHICAGO

Data Exploration & Feature Engineering

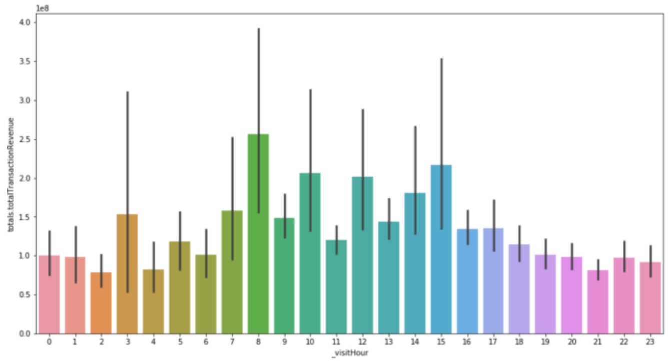
Visitor time analysis:

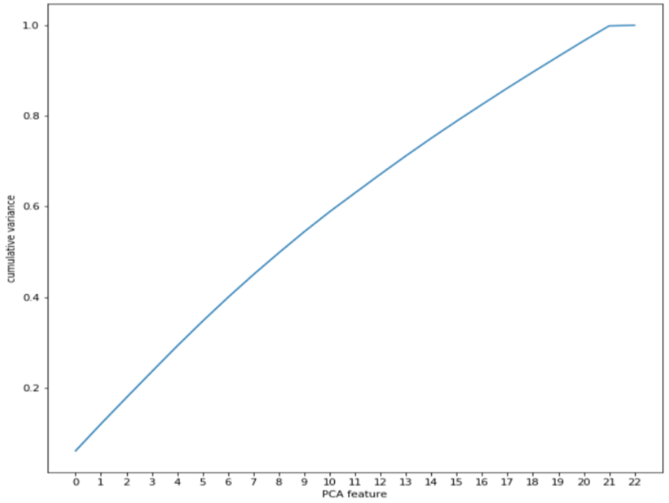
- ❑ Tuesday is the pick day



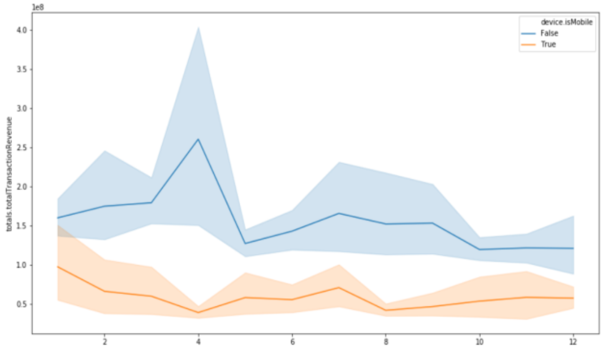


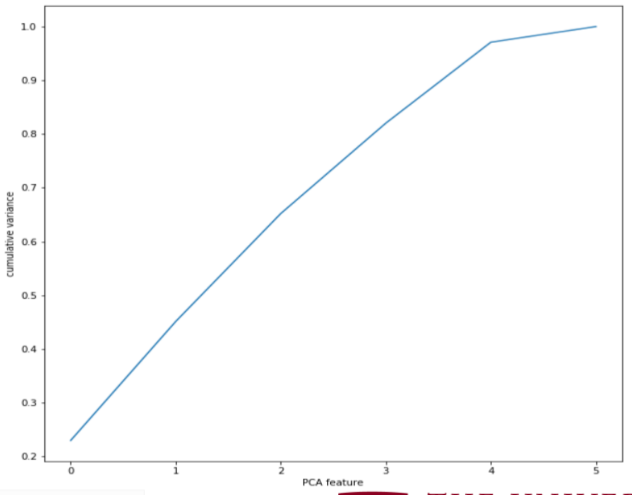
- ❑ Highest revenue: 3am, 8am, 10am, 12pm, 4pm, 5pm





- ❑ April is the pick month



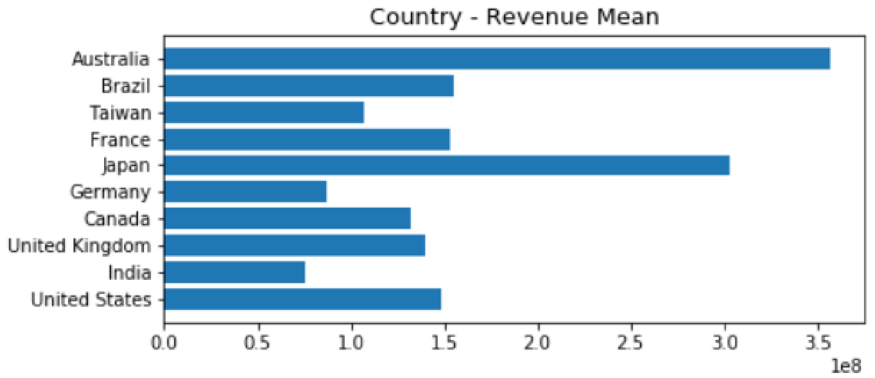


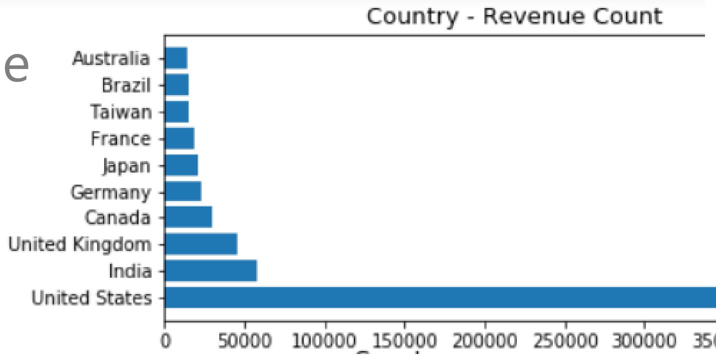
THE UNIVERSITY OF CHICAGO

Data Exploration & Feature Engineering

Country & Region:

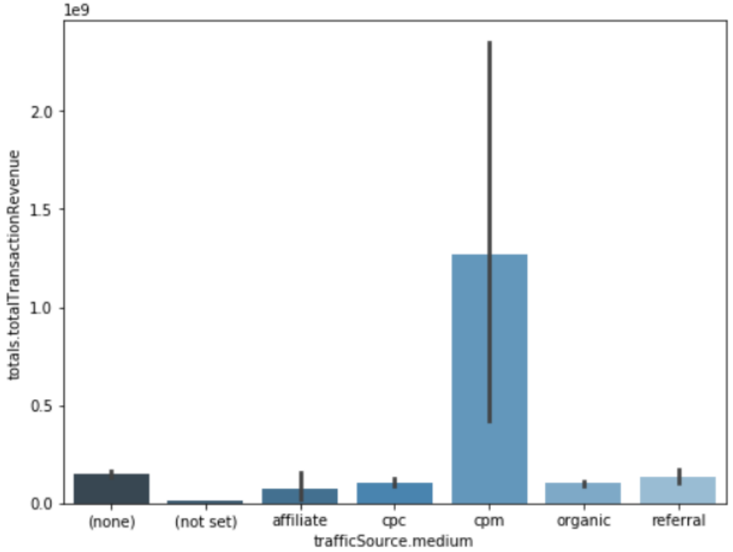
- ❑ United States contributed 95% of total revenue
- ❑ Australia has highest Revenue mean
- ❑ Create new feature called isUnitedStates, isAustralia

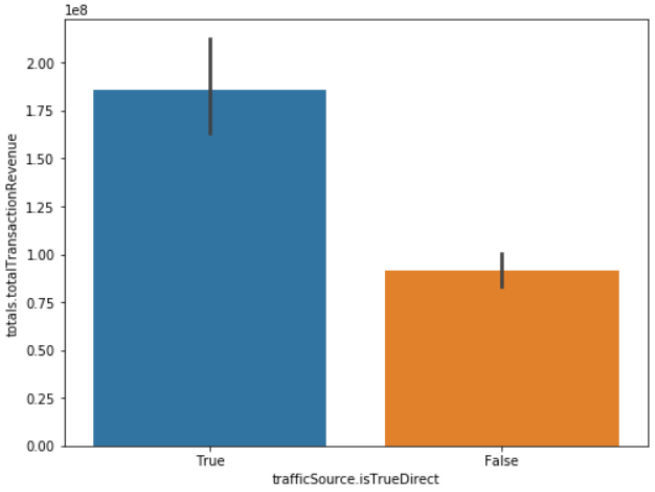




Traffic Source:

- ❑ CPM contributes main transaction revenues for medium of traffic source
- ❑ There is difference between True direct and False direct



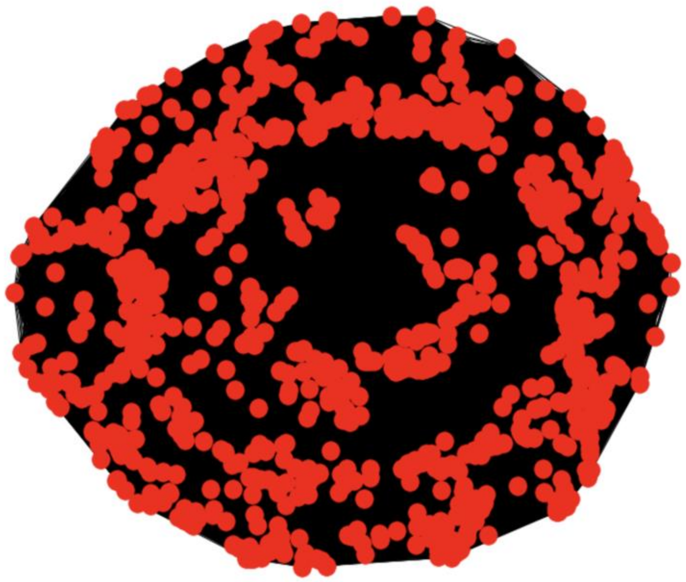


THE UNIVERSITY OF CHICAGO

Graph Analysis

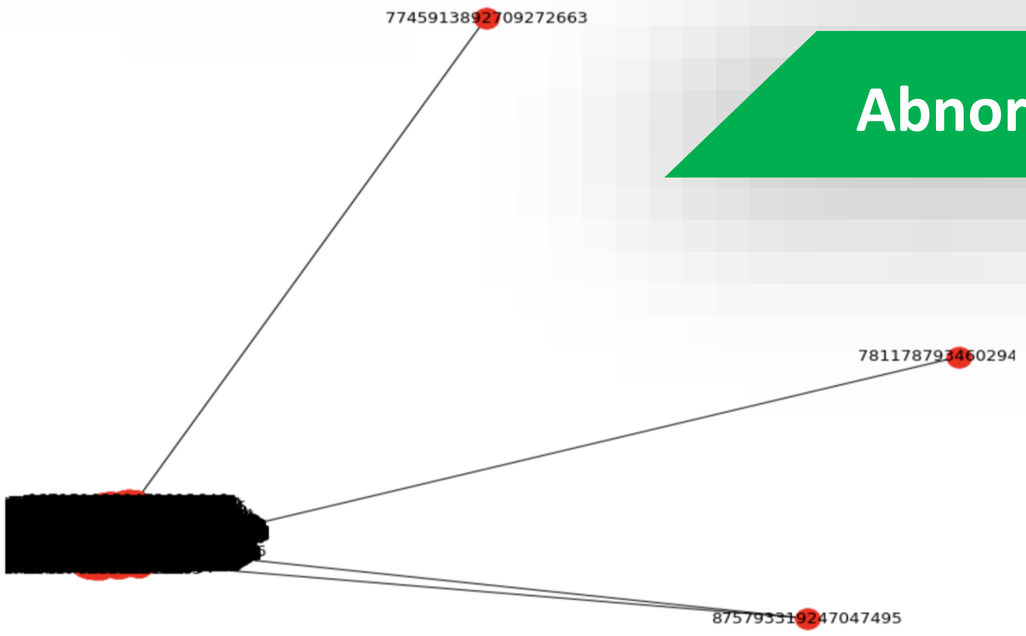
Revenue

- ❑ Node is User
- ❑ Edge revenue group for each transaction
  - Log revenue rounded to tens place
  - Remove zeros
- ❑ 1000 transactions



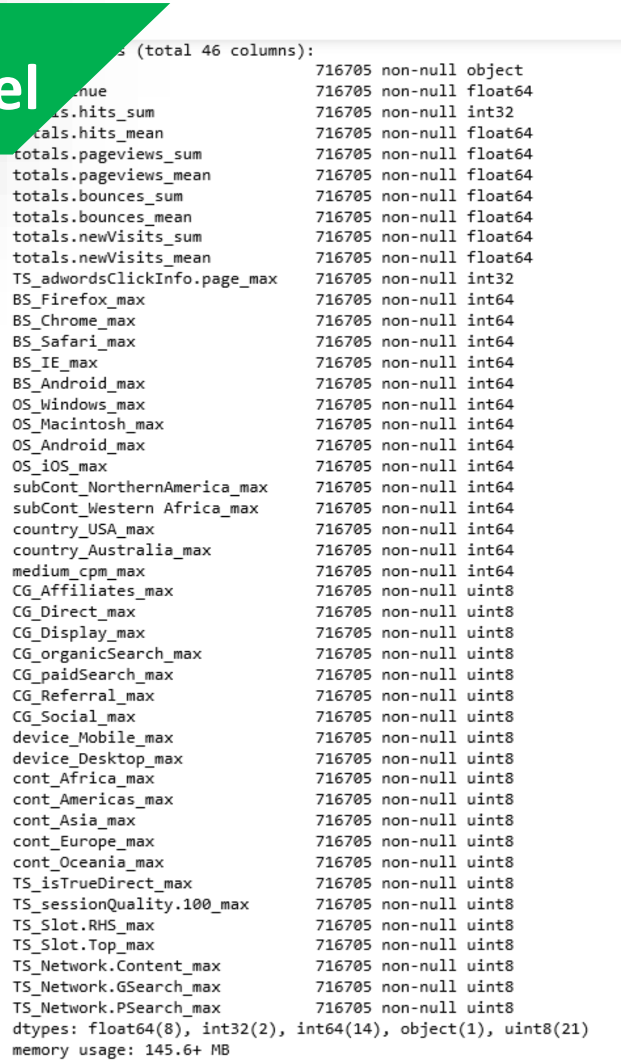
Abnormal User

- ❑ Subset of Users
  - Had transactions from more than one country
- ❑ Node and Edge similar to Revenue group
  - Edge not transformed
- ❑ (Not Set) country code



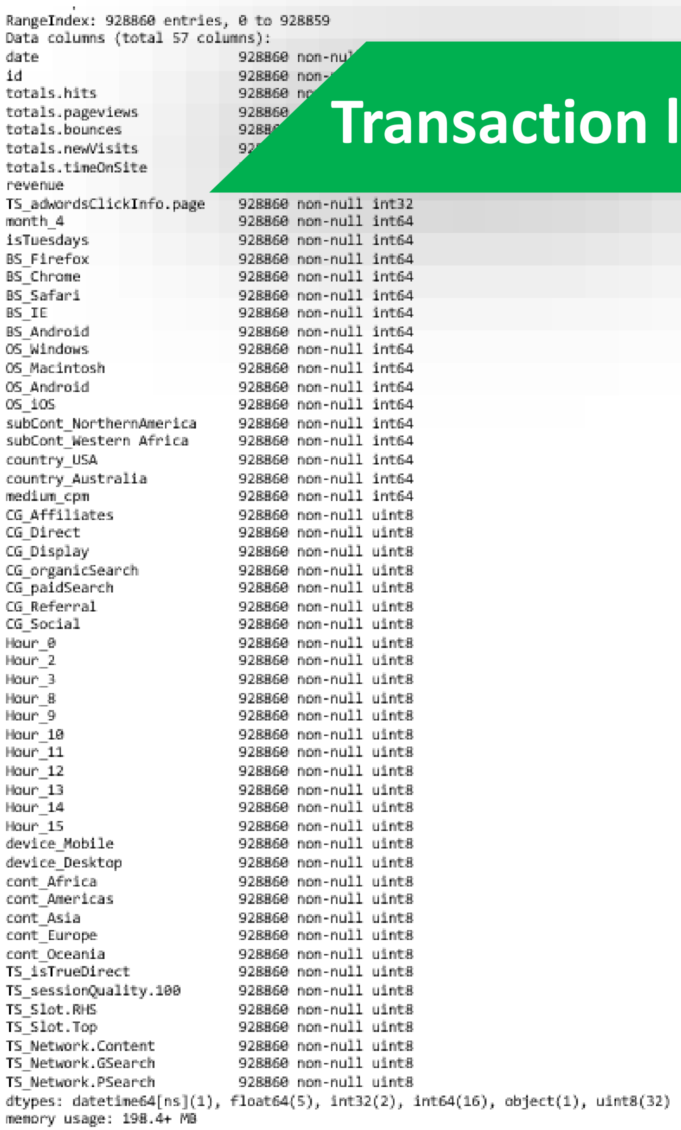
Two different approaches

Customer level



```
              s (total 46 columns):
                                716705 non-null object
         nue                    716705 non-null float64
     s.hits_sum                 716705 non-null int32
    tals.hits_mean              716705 non-null float64
totals.pageviews_sum            716705 non-null float64
totals.pageviews_mean           716705 non-null float64
totals.bounces_sum              716705 non-null float64
totals.bounces_mean             716705 non-null float64
totals.newVisits_sum            716705 non-null float64
totals.newVisits_mean           716705 non-null float64
TS_adwordsClickInfo.page_max    716705 non-null int32
BS_Firefox_max                  716705 non-null int64
BS_Chrome_max                   716705 non-null int64
BS_Safari_max                   716705 non-null int64
BS_IE_max                       716705 non-null int64
BS_Android_max                  716705 non-null int64
OS_Windows_max                  716705 non-null int64
OS_Macintosh_max                716705 non-null int64
OS_Android_max                  716705 non-null int64
OS_iOS_max                      716705 non-null int64
subCont_NorthernAmerica_max     716705 non-null int64
subCont_Western Africa_max      716705 non-null int64
country_USA_max                 716705 non-null int64
country_Australia_max           716705 non-null int64
medium_cpm_max                  716705 non-null int64
CG_Affiliates_max               716705 non-null uint8
CG_Direct_max                   716705 non-null uint8
CG_Display_max                  716705 non-null uint8
CG_organicSearch_max            716705 non-null uint8
CG_paidSearch_max               716705 non-null uint8
CG_Referral_max                 716705 non-null uint8
CG_Social_max                   716705 non-null uint8
device_Mobile_max               716705 non-null uint8
device_Desktop_max              716705 non-null uint8
cont_Africa_max                 716705 non-null uint8
cont_Americas_max               716705 non-null uint8
cont_Asia_max                   716705 non-null uint8
cont_Europe_max                 716705 non-null uint8
cont_Oceania_max                716705 non-null uint8
TS_isTrueDirect_max             716705 non-null uint8
TS_sessionQuality.100_max       716705 non-null uint8
TS_Slot.RHS_max                 716705 non-null uint8
TS_Slot.Top_max                 716705 non-null uint8
TS_Network.Content_max          716705 non-null uint8
TS_Network.GSearch_max          716705 non-null uint8
TS_Network.PSearch_max          716705 non-null uint8
dtypes: float64(8), int32(2), int64(14), object(1), uint8(21)
memory usage: 145.6+ MB
```

Transaction level



```
RangeIndex: 928860 entries, 0 to 928859
Data columns (total 57 columns):
date                        928860 non-null
id                          928860 non-
totals.hits                 928860 no
totals.pageviews            928860
totals.bounces              9288
totals.newVisits            92
totals.timeOnSite
revenue
TS_adwordsClickInfo.page    928860 non-null int32
month_4                     928860 non-null int64
isTuesdays                  928860 non-null int64
BS_Firefox                  928860 non-null int64
BS_Chrome                   928860 non-null int64
BS_Safari                   928860 non-null int64
BS_IE                       928860 non-null int64
BS_Android                  928860 non-null int64
OS_Windows                  928860 non-null int64
OS_Macintosh                928860 non-null int64
OS_Android                  928860 non-null int64
OS_iOS                      928860 non-null int64
subCont_NorthernAmerica     928860 non-null int64
subCont_Western Africa      928860 non-null int64
country_USA                 928860 non-null int64
country_Australia           928860 non-null int64
medium_cpm                  928860 non-null int64
CG_Affiliates               928860 non-null uint8
CG_Direct                   928860 non-null uint8
CG_Display                  928860 non-null uint8
CG_organicSearch            928860 non-null uint8
CG_paidSearch               928860 non-null uint8
CG_Referral                 928860 non-null uint8
CG_Social                   928860 non-null uint8
Hour_0                      928860 non-null uint8
Hour_2                      928860 non-null uint8
Hour_3                      928860 non-null uint8
Hour_8                      928860 non-null uint8
Hour_9                      928860 non-null uint8
Hour_10                     928860 non-null uint8
Hour_11                     928860 non-null uint8
Hour_12                     928860 non-null uint8
Hour_13                     928860 non-null uint8
Hour_14                     928860 non-null uint8
Hour_15                     928860 non-null uint8
device_Mobile               928860 non-null uint8
device_Desktop              928860 non-null uint8
cont_Africa                 928860 non-null uint8
cont_Americas               928860 non-null uint8
cont_Asia                   928860 non-null uint8
cont_Europe                 928860 non-null uint8
cont_Oceania                928860 non-null uint8
TS_isTrueDirect             928860 non-null uint8
TS_sessionQuality.100       928860 non-null uint8
TS_Slot.RHS                 928860 non-null uint8
TS_Slot.Top                 928860 non-null uint8
TS_Network.Content          928860 non-null uint8
TS_Network.GSearch          928860 non-null uint8
TS_Network.PSearch          928860 non-null uint8
dtypes: datetime64[ns](1), float64(5), int32(2), int64(16), object(1), uint8(32)
memory usage: 198.4+ MB
```

THE UNIVERSITY OF CHICAGO

Costumer Level Modeling

Goal

✓ Predict the log of the revenue per user

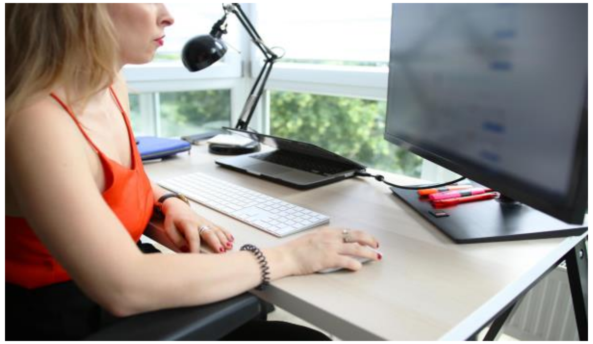
Steps

I. Aggregate data per user ID
II. Sum of the original revenue and log of the sum per user
III. Build Model

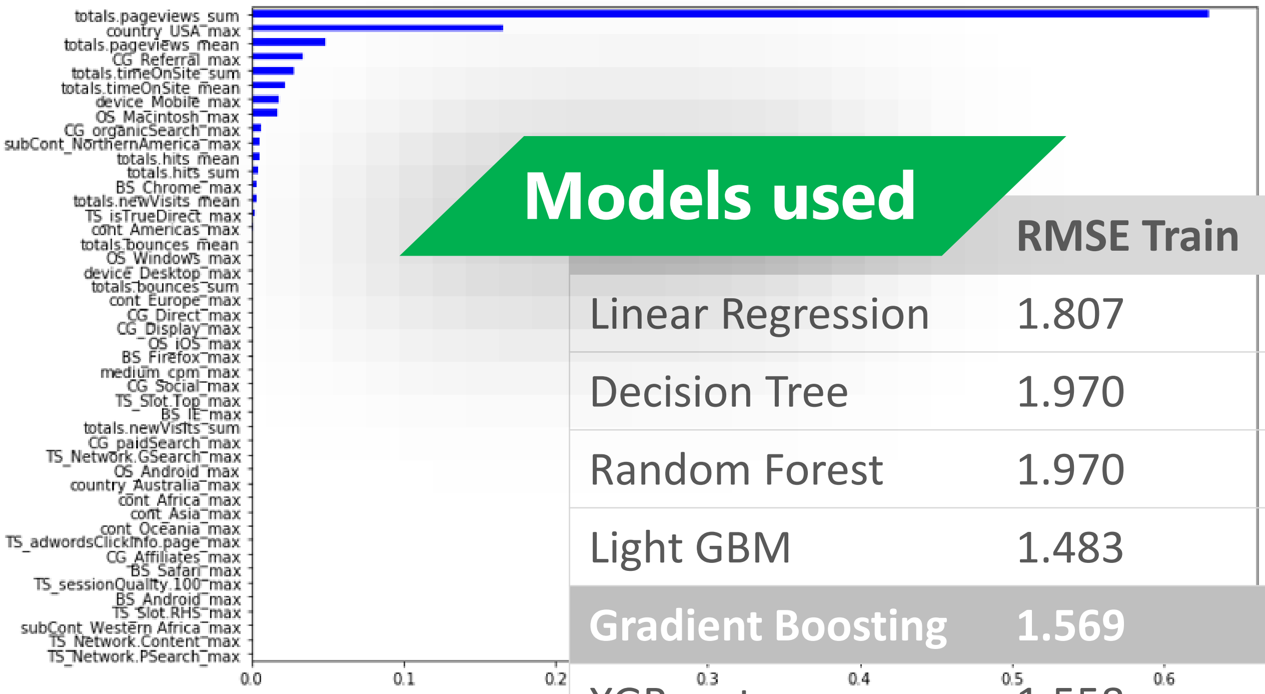
Y^t = The revenue at transaction level

Y^a = The revenue at user level

$$\sum Y^t{}_{user} = \xrightarrow{\text{Aggregate}} Y^a \xrightarrow{\text{Transform}} Y^a_{log} \xrightarrow{\text{Model}} \hat{Y}^a_{log}$$



Modeling: Customer level



Models used

	RMSE Train	RMSE Test	Cross Validation
Linear Regression	1.807	1.757	1.886
Decision Tree	1.970	1.894	1.938
Random Forest	1.970	1.894	1.970
Light GBM	1.483	1.532	1.571
Gradient Boosting	**1.569**	**1.549**	**1.589**
XGBoost	1.558	1.553	1.588

THE UNIVERSITY OF
CHICAGO

Transaction Level Modeling

Goal

✓ Want to predict the transaction level revenue as well as compare our results to the aggregated data

Steps

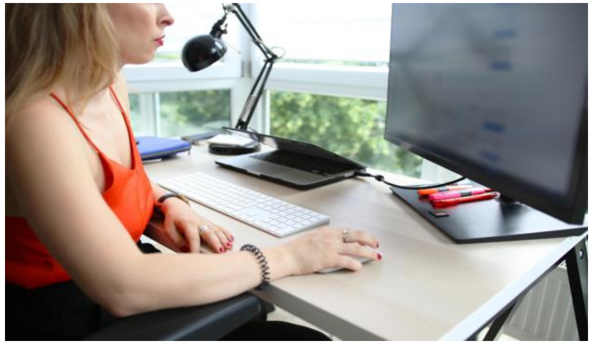
I. Log Revenue
II. Build Model
III. Exponentiate Predicted values
IV. Aggregate Sum over User ID
V. Log Predicted
VI. Compare values to Aggregate Log Revenue

$$Y^t = \text{The revenue at transaction level}$$
$$Y^a = \text{The revenue at user level}$$

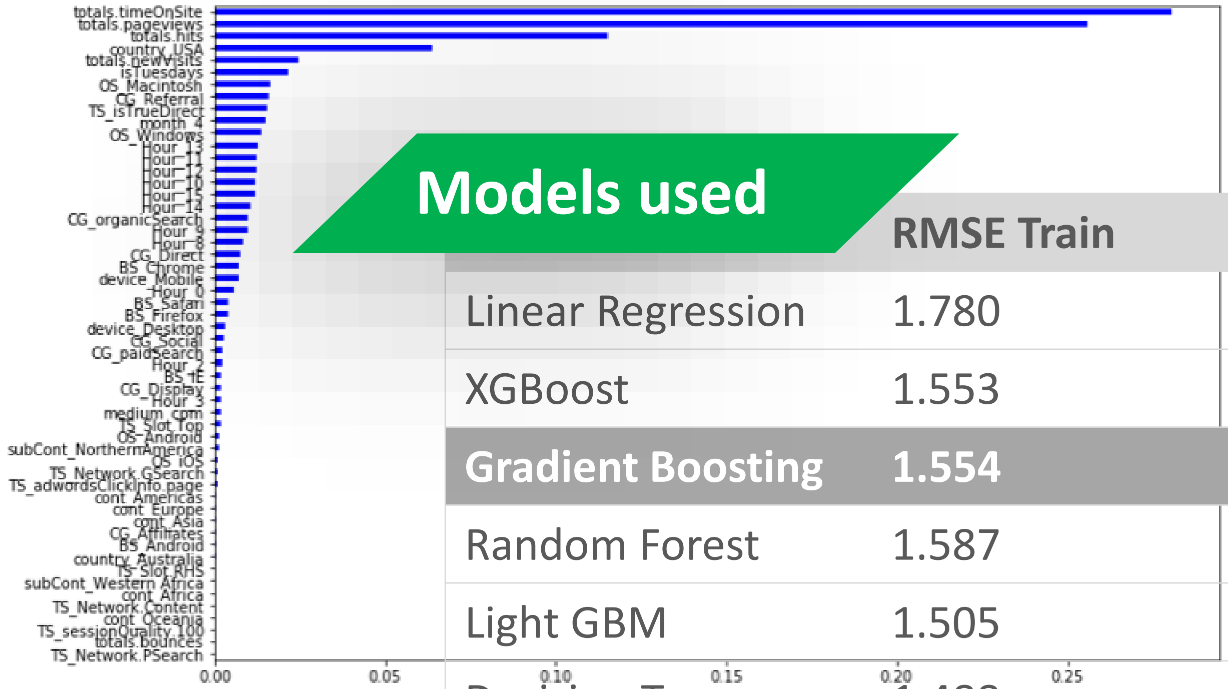
$$Y^t \xrightarrow{\text{Transform}} Y^t_{log} \xrightarrow{\text{Model}} \hat{Y}^t_{log} \xrightarrow{\text{Transform}} e^{\hat{Y}^t_{log}} = \hat{Y}^t \xrightarrow{\text{Aggregate}} \sum \hat{Y}^t_{user} = \hat{Y}^a \xrightarrow{\text{Transform}} \hat{Y}^a_{log}$$

$$\text{Error} = Y^a_{log} - \hat{Y}^a_{log}$$

THE UNIVERSITY OF CHICAGO



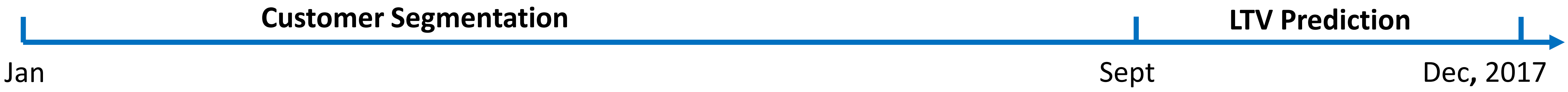
Modeling: Transaction Level



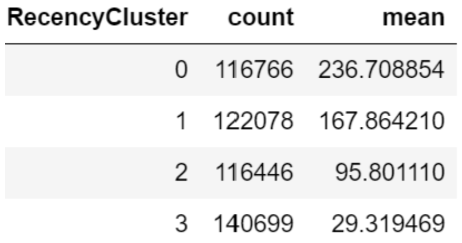
Models used

	RMSE Train	RMSE Test	Cross Validation
Linear Regression	1.780	1.724	1.781
XGBoost	1.553	1.556	1.573
Gradient Boosting	**1.554**	**1.557**	**1.575**
Random Forest	1.587	1.575	1.596
Light GBM	1.505	1.552	1.560
Decision Tree	1.488	2.188	2.226

Customer Segmentation & Life Time Value Prediction



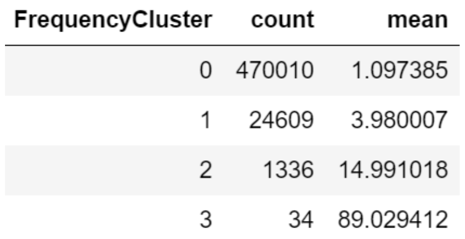
Recency



RecencyCluster	count	mean
0	116766	236.708854
1	122078	167.864210
2	116446	95.801110
3	140699	29.319469

+

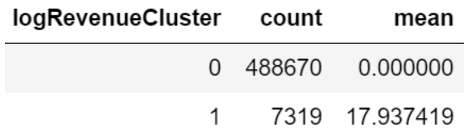
Frequency



FrequencyCluster	count	mean
0	470010	1.097385
1	24609	3.980007
2	1336	14.991018
3	34	89.029412

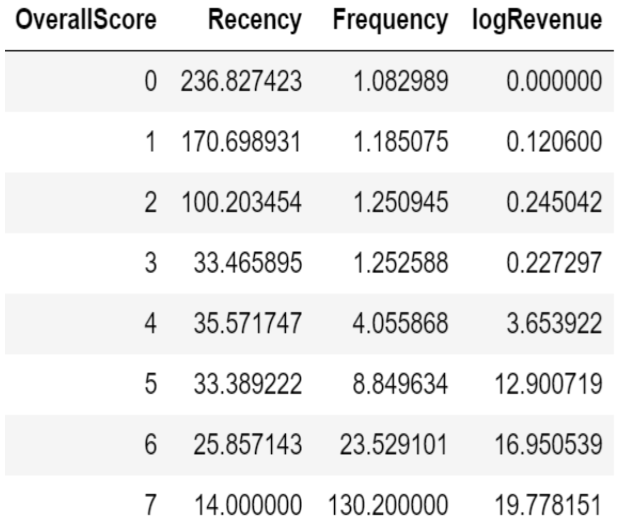
+

logRevenue



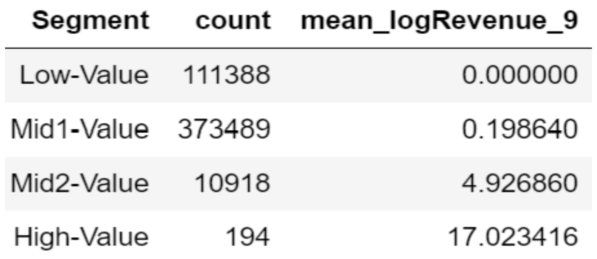
logRevenueCluster	count	mean
0	488670	0.000000
1	7319	17.937419

=



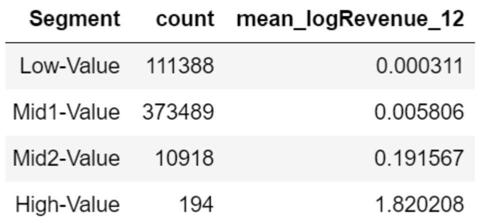
OverallScore	Recency	Frequency	logRevenue
0	236.827423	1.082989	0.000000
1	170.698931	1.185075	0.120600
2	100.203454	1.250945	0.245042
3	33.465895	1.252588	0.227297
4	35.571747	4.055868	3.653922
5	33.389222	8.849634	12.900719
6	25.857143	23.529101	16.950539
7	14.000000	130.200000	19.778151

↓



Segment	count	mean_logRevenue_9
Low-Value	111388	0.000000
Mid1-Value	373489	0.198640
Mid2-Value	10918	4.926860
High-Value	194	17.023416

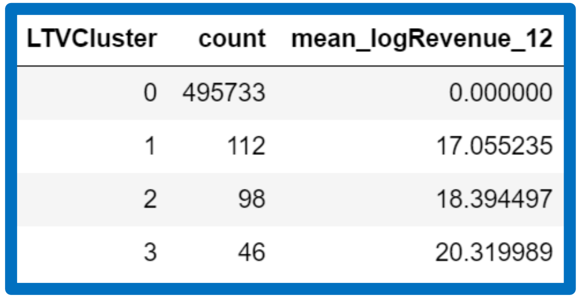
logRevenue (last 3 months)



Segment	count	mean_logRevenue_12
Low-Value	111388	0.000311
Mid1-Value	373489	0.005806
Mid2-Value	10918	0.191567
High-Value	194	1.820208

↓

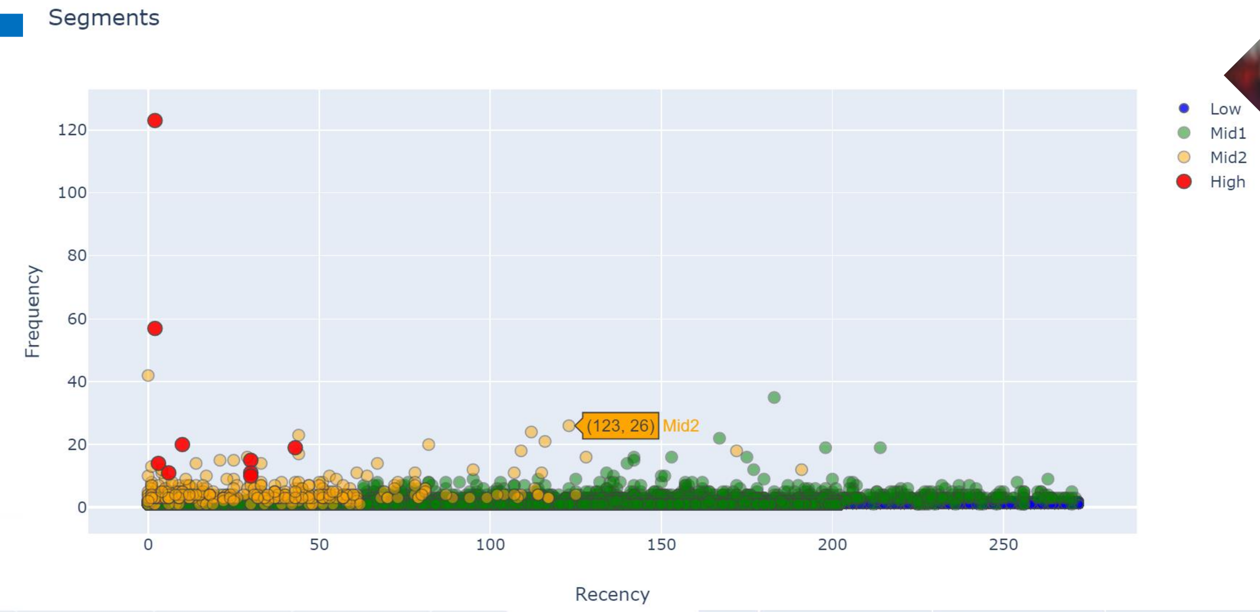
LTV Cluster



LTVCluster	count	mean_logRevenue_12
0	495733	0.000000
1	112	17.055235
2	98	18.394497
3	46	20.319989

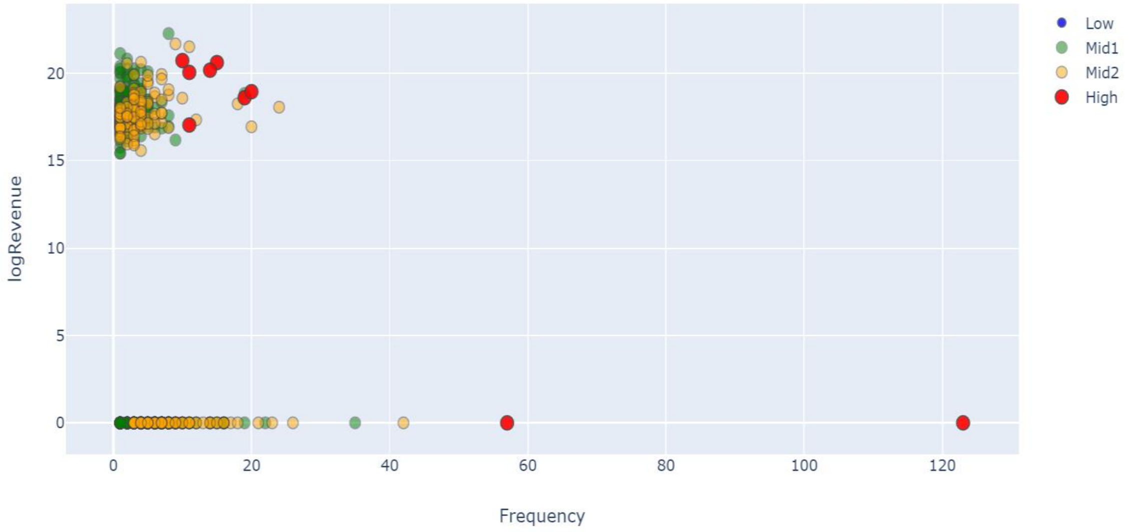
THE UNIVERSITY OF CHICAGO

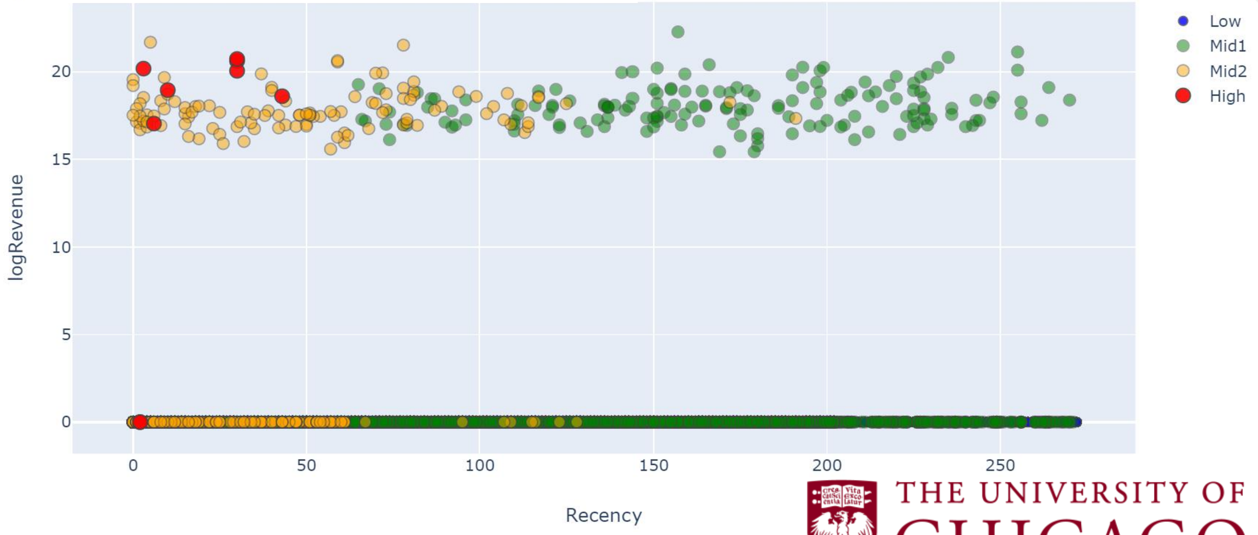
Customer Segmentation



Segments

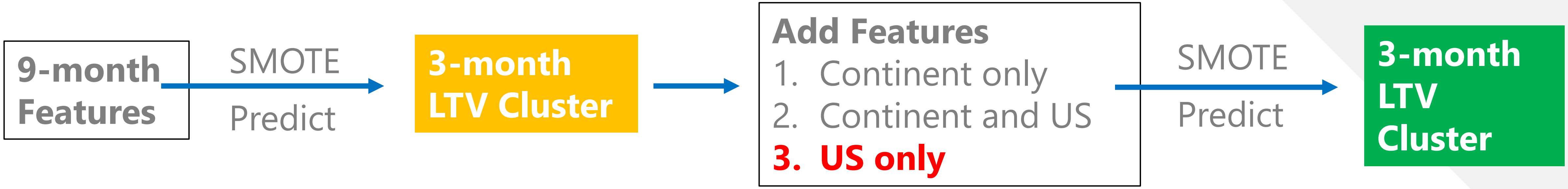
Segments





THE UNIVERSITY OF CHICAGO

Customer LTV Prediction



9-month Features

SMOTE

Predict

3-month LTV Cluster

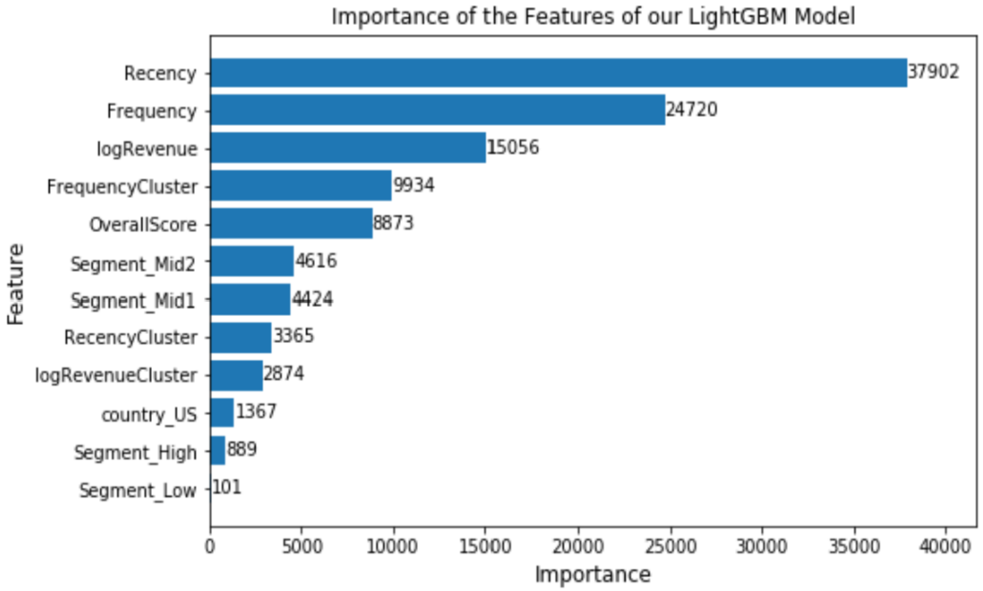
Add Features

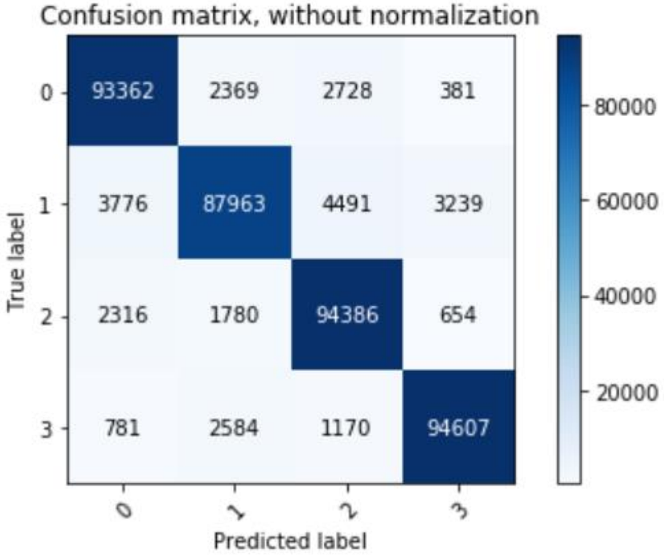
1. Continent only
2. Continent and US
3. **US only**

SMOTE

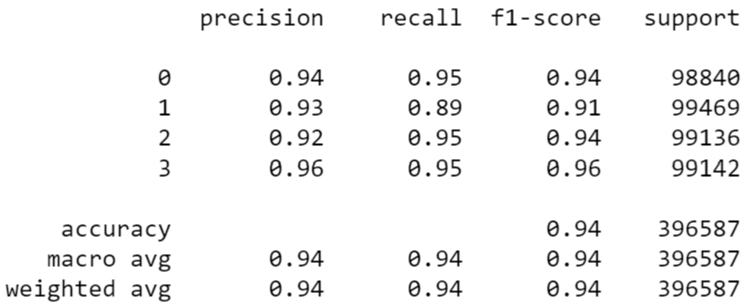
Predict

3-month LTV Cluster





	0	1	2	3
0	93362	2369	2728	381
1	3776	87963	4491	3239
2	2316	1780	94386	654
3	781	2584	1170	94607



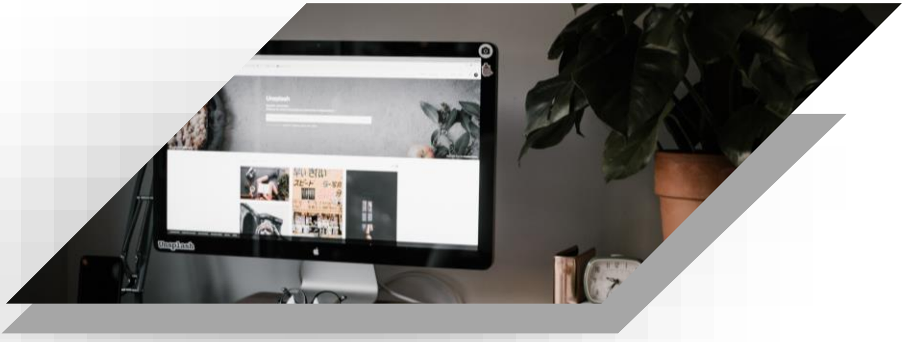
```
              precision    recall  f1-score   support

           0       0.94      0.95      0.94     98840
           1       0.93      0.89      0.91     99469
           2       0.92      0.95      0.94     99136
           3       0.96      0.95      0.96     99142

    accuracy                           0.94    396587
   macro avg       0.94      0.94      0.94    396587
weighted avg       0.94      0.94      0.94    396587
```

THE UNIVERSITY OF CHICAGO

Lessons Learned & Future Work

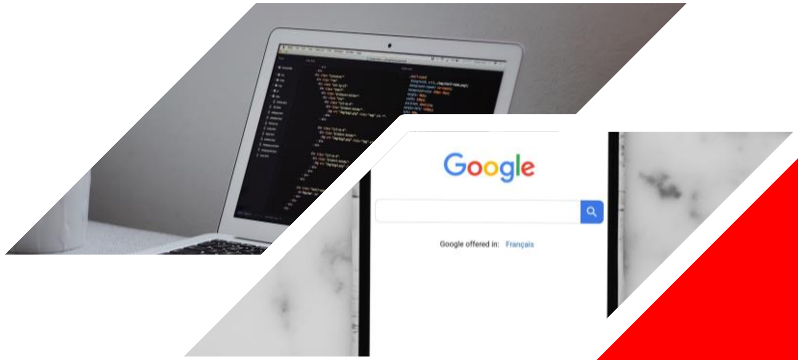


- ❑ Meaningful insights from large datasets sometimes more complicated
- ❑ Do proper research on models that better suit the specificities of our data
- ❑ Large data may limit the number of models to use
- ❑ Keep in mind the business goal throughout the project

- ❑ Use product level data for more insights
- ❑ Use several years of data – identify seasonality
- ❑ Apply more models to Customer segmentation – LTV
- ❑ Tune models more

THE UNIVERSITY OF CHICAGO

Sources



- ❑ Data: Google Analytics Customer Revenue Prediction – **link**
- ❑ Statistics digital buyers – **link**
- ❑ Statistics on ecommerce – **link** & **link**
- ❑ Customer lifecycle prediction – **link**

THE UNIVERSITY OF CHICAGO

Q&A

Google

Do the right thing

Analytics

Thank You

THE UNIVERSITY OF CHICAGO