# CONDÉ NAST

Yanghui Liu | Yiyun (Rona) Luo | Suzy Gao | Tianying (Jessie) Gong

Yumeng Gu | Li Su | Yuhan (Jennifer) Wang | Mo Yang

# AGENDA

- Executive Summary
- Data Preparation
- Descriptive Analysis
- Predictive Analysis
- Recommendation

# EXECUTIVE SUMMARY

## PREP

- Collected company data
- Conducted data pre-processing
- Finalized dataset

## METHODS

- Evaluated effectiveness of models
- Conducted descriptive analysis to present models' changes over time & key variable analysis
- Persona analysis to discover patterns of high-quality users
- Text mining to reveal contents of interest
- Predictive Analysis

# 1 Data Prep

Collect Data from Internal Database

Aggregate and join datasets

Drop, add and alter columns

Fill in missing values

Unify data format

## Final Dataset:

Time range: 2019/1/3 - 2019/3/26
18 columns:

Infiniity_id, brand, category, channel, location, country, browser, scrolldepth, scroll_depth_0_1, referrer, first_author, title, keyword, score, score_date, bucket, model

# 2 Descriptive Analysis: Summary

**Model performance:**

Models' performance has differed over time. Not all of them are considered effective. Threshold needs to be set to identify effective predicting modeling attempts. In this project we proposed a threshold value of auc equals to 0.7.

**Key variables:**

Key indicators for luxury male, luxury retail and luxury watch model: brand_brides_bin, topic_127_bin, topic_137_bin, topic_139_bin, country_UnitedStates_dma_638_bin, country_UnitedStates_dma_756_bin
Key indicators for luxury jewelry model: brand_vogue_bin

**Marketing insights:**

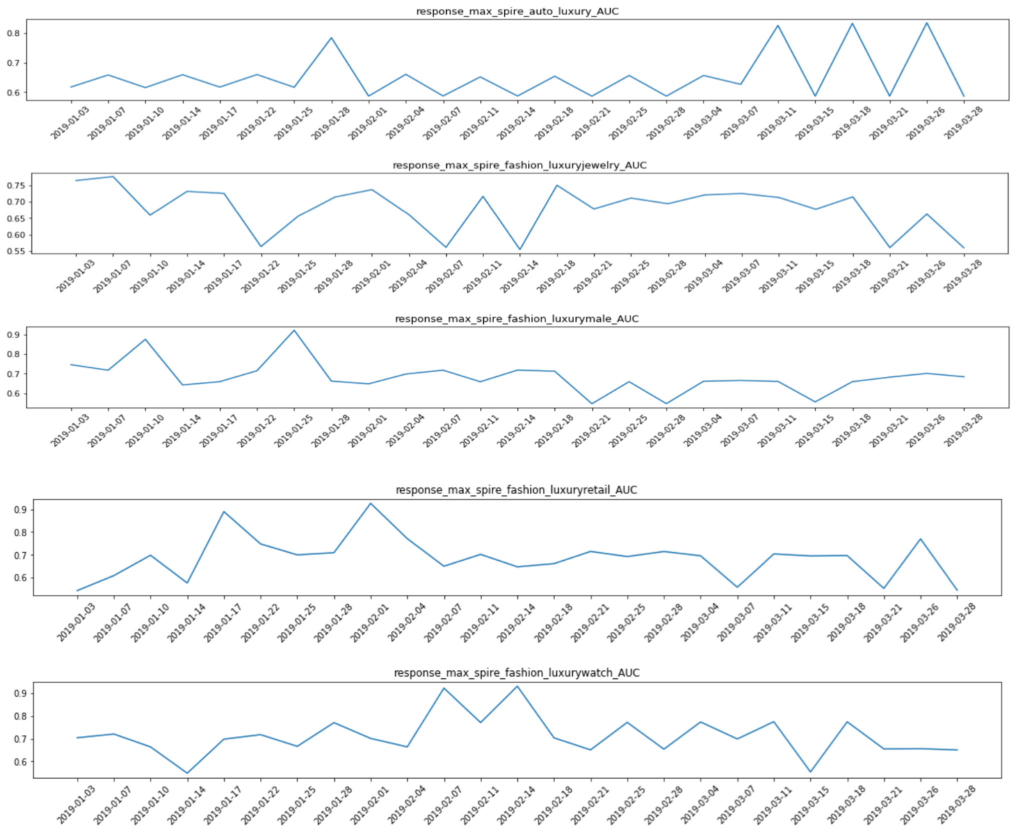
Brands of magazines, devices of visitors using, and how visitors access to a website are three major factors that could affect visitors response to specific information.

# 2.1 Variable Coefficients and AUCs

Inspect AUC data of the five different models

Baseline: auc > 0.7*

*Rice, Marnie & Harris, Grant. (2005). Rice ME, Harris GTComparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. Law Hum Behav 29: 615-620. Law and human behavior. 29. 615-20. 10.1007/s10979-005-6832-7.



# 2.1 Variable Coefficients and AUCs

## Dates for Effective Results

| | January | | | | | | | | February | | | | | | | | | March | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 03 | 07 | 10 | 14 | 17 | 22 | 25 | 28 | 01 | 04 | 07 | 11 | 14 | 18 | 21 | 25 | 28 | 04 | 07 | 11 | 18 | 26 |
| auto_luxury_model | | | | | | | | √ | | | | | | | | | | | | √ | √ | √ |
| fashion_luxuryjewelry_model | √ | √ | | √ | √ | | | √ | √ | | | √ | | √ | | √ | | √ | √ | √ | √ | |
| fashion_luxurymale_model | √ | √ | √ | | | √ | √ | | | | √ | | √ | √ | | | | | | | | √ |
| fashion_luxuryretail_model | | | | | √ | √ | √ | √ | √ | √ | | √ | | | √ | | √ | | | √ | | √ |
| fashion_luxurywatch_model | √ | √ | | | | √ | | √ | √ | | √ | √ | √ | √ | | √ | | √ | | √ | √ | |

**Link to Coef !**

# 2.1 Variable Coefficients and AUCs

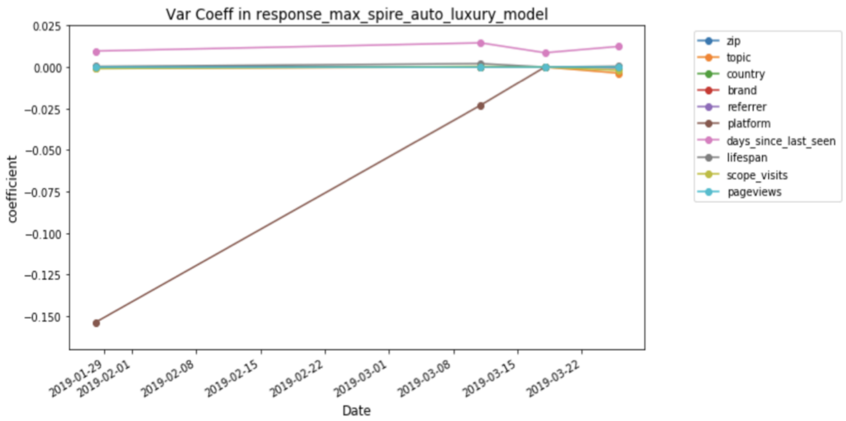
- Analyze the coefficient of each category of variables in the models when the models function effectively (auc > 0.7)

- Method
  - aggregated same category of variables by median
  - aggregated same category of variables by  mean

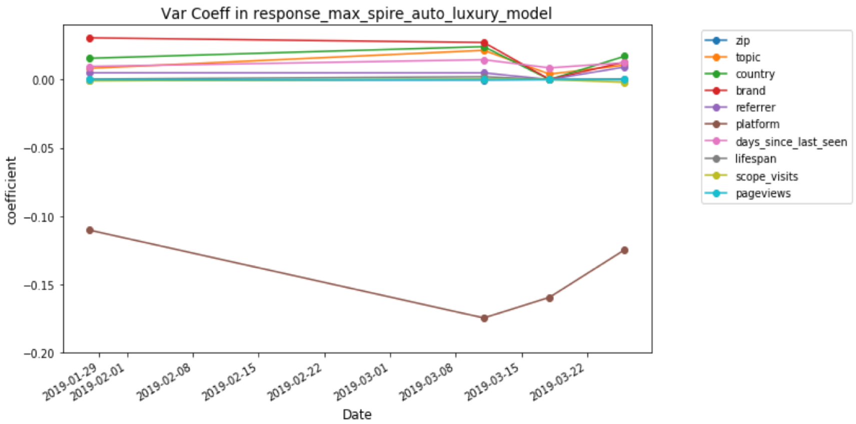
| Category | #of variables |
|---|---|
| zip | 521 |
| topic | 301 |
| country | 224 |
| brand | 48 |
| referrer | 7 |
| platform | 3 |
| days_since_last_seen | 1 |
| lifespan | 1 |
| scope_visits | 1 |
| pageviews | 1 |

# 2.2 Variable Coefficient Changes in Luxury Model

Aggregate by Median

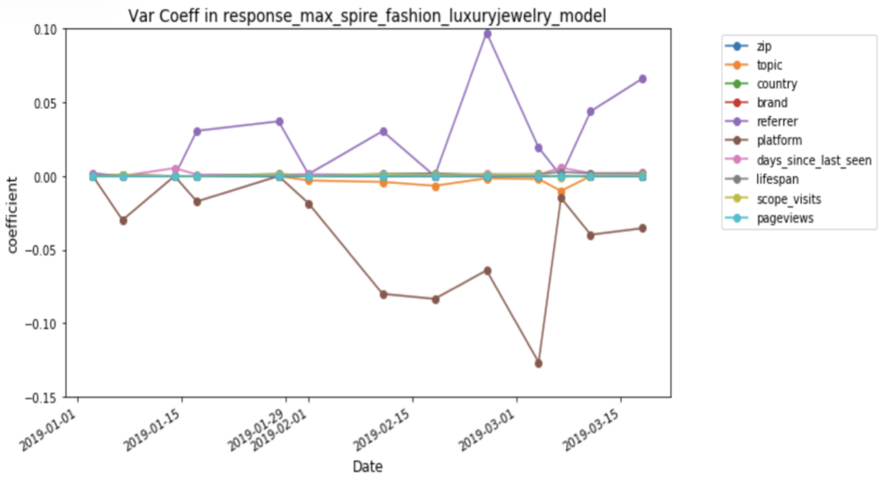


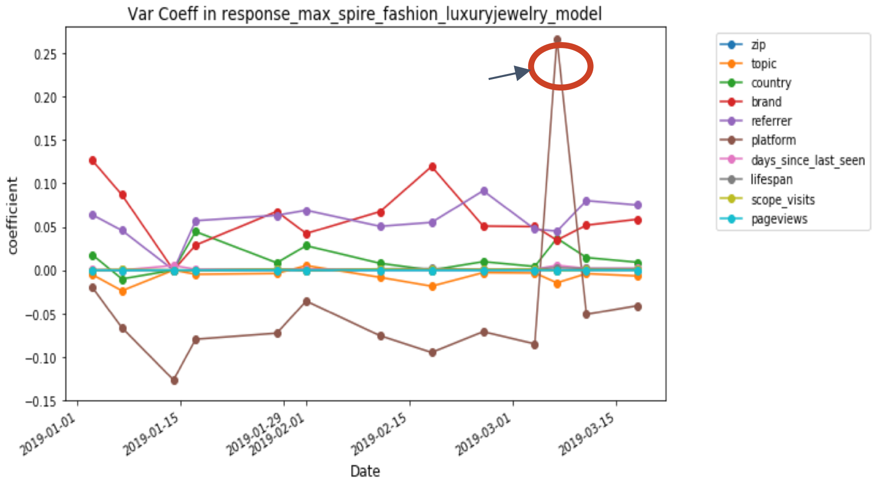

Aggregate by Mean




- The response_max_spire_auto_luxury_model can only function well in 4 days
- Variables of ‘platform’ negatively impact the model
- The trend of the variable coefficient are the same in the response_max_spire_auto_luxury_model the except for the variables of ‘platform’

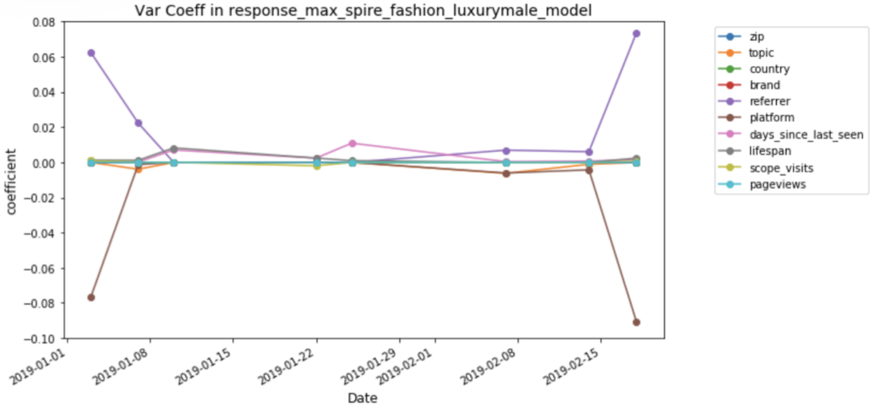
# 2.2 Variables Coefficient Changes in Luxuryjewelry Model

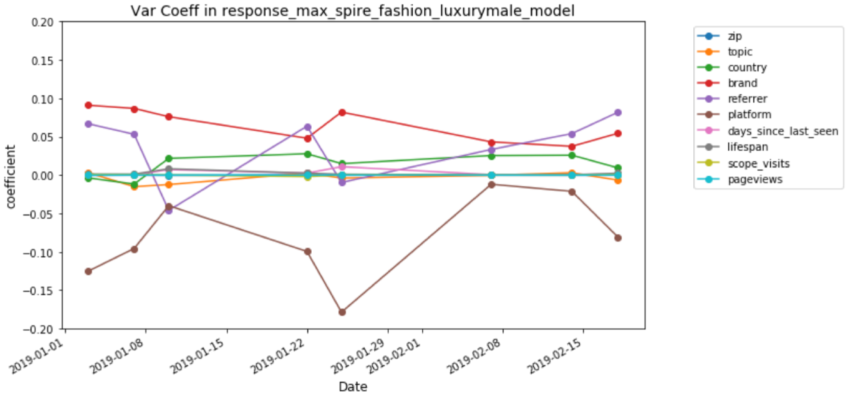



- Variables of referrer and variables of platform have an opposite impact to the response_max_spire_fashion_luxuryjewelry_model
- On 2019-03-07, ‘platform_mobile_bin’ coefficient is extremely high and drive the mean of ‘platform’ coefficient to over 0.25

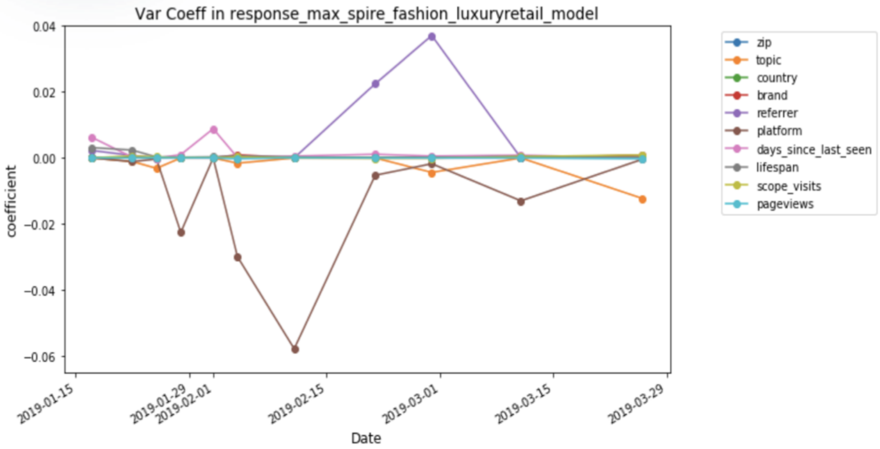
# 2.2 Variables Coefficient Changes in Luxurymale Model

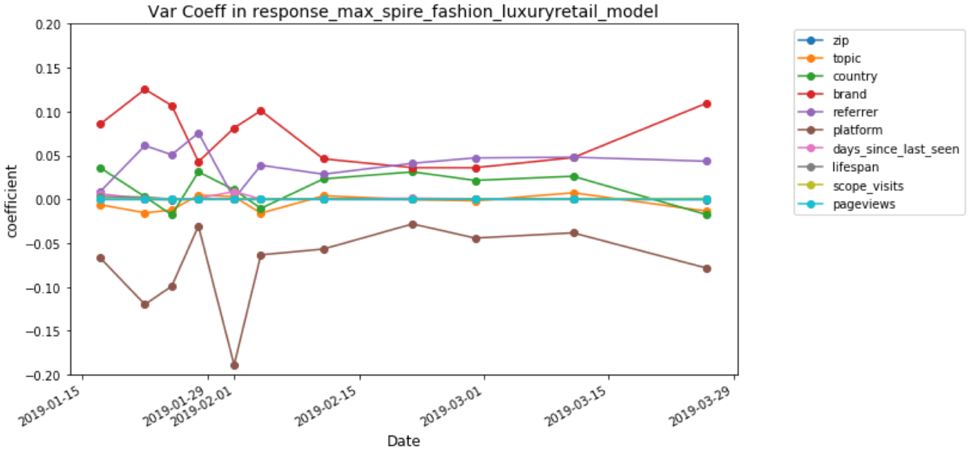



- Variables of referrer and variables of platform have an opposite impact to the response_max_spire_fashion_luxurymale_model
- Variables of 'Brand' positively impact the response_max_spire_fashion_luxurymale_model

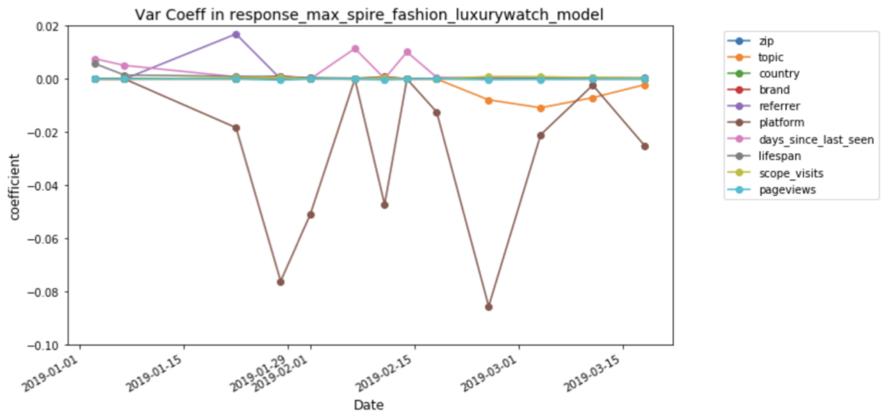
# 2.2 Variables Coefficient Changes in Luxuryretail Model




- Variables of ‘referrer’ and variables of ‘platform’ have an opposite impact on the response_max_spire_fashion_luxuryretail_model
- The coefficient of ‘topic’ variables drop in a great degree on 2019-03-28
- Except the variables of ‘platform’ and “topic”, other variables usually have positive impact on the response_max_spire_fashion_luxuryretail_model
- The variables of ‘Brand’ and ‘referrer’ have more impact over other variables

# 2.2 Variables Coefficient Changes in Luxurywatch Model





- The variables of ‘brand’, ‘referrer’ and ‘days_since_last_seen’ have positive impact on the response_max_spire_fashion_luxurywatch_model

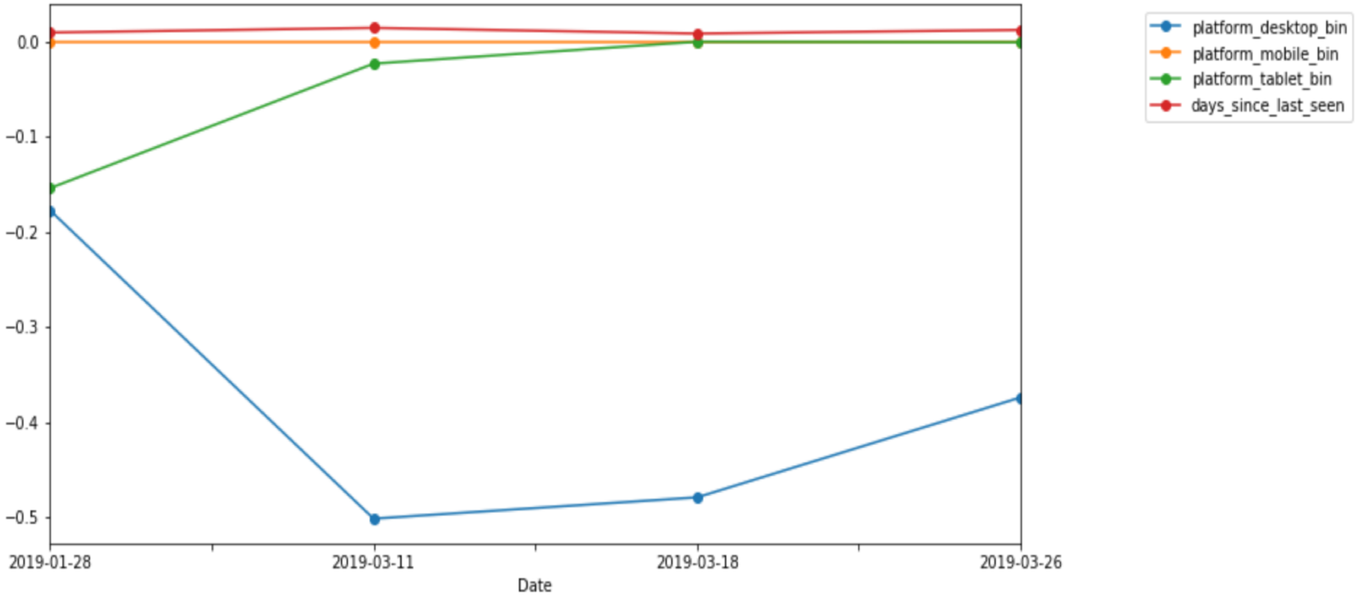
# Variables in Top Categories for Each Model

Examine the changes of variables’ coefficients over time within identified top categories for each model:

- TOP categories of variables for Auto_luxury_model: 'days_since_last_seen’ and 'platform'.
- TOP categories of variables for luxury_jewelry_model, luxury_male_model, luxury_ retail_model, luxury_watch_model: 'days_since_last_seen’ and 'referrer'.

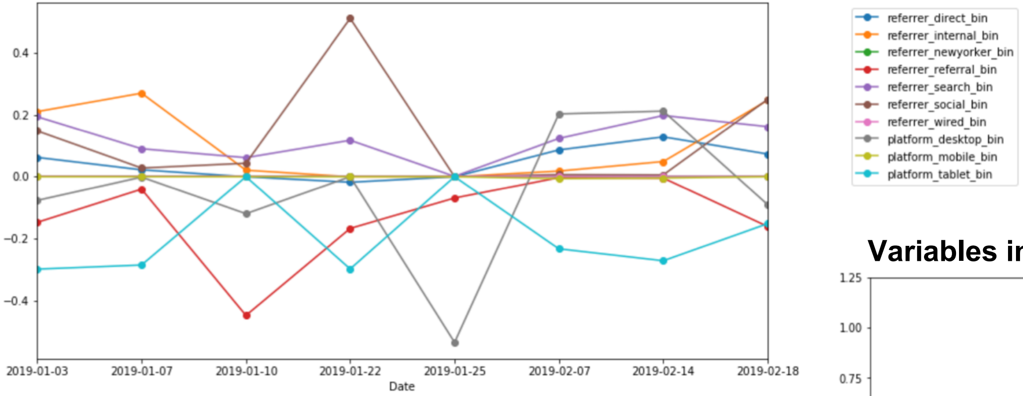
# 2.3 Variables in Top Categories for Each Model

**Variables in TOP Categories - Coeff in auto_luxury_model - 'days_since_last_seen' , 'platform'**

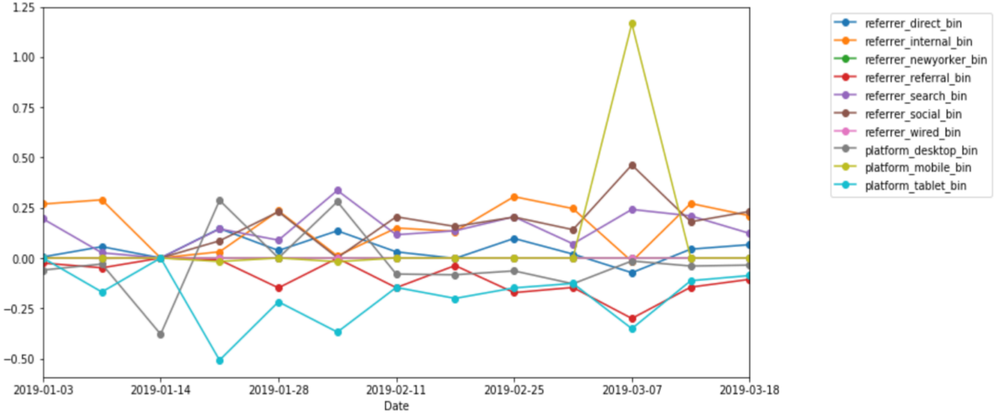


# 2.3 Variables in Top Categories for Each Model

**Variables in TOP Categories - Coef in response_max_spire_fashion_luxurymale_model**

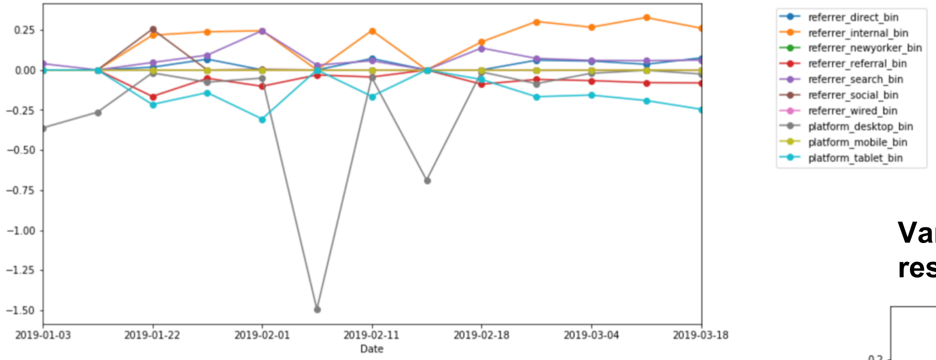



**Variables in TOP Categories - Coef in luxuryjewelry_model**

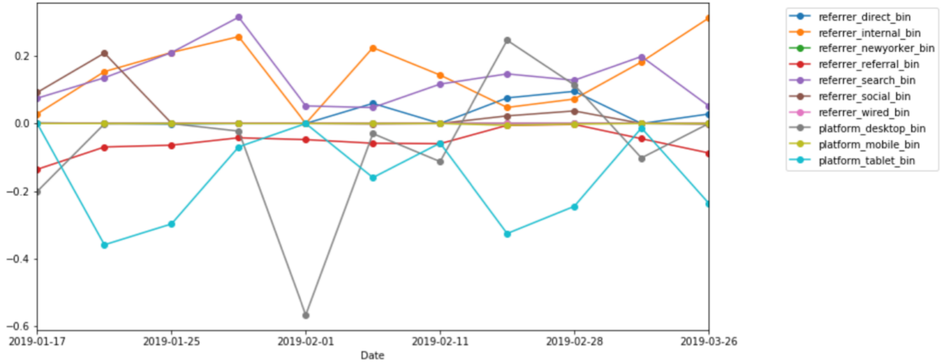


# 2.3 Variables in Top Categories for Each Model

### Variables in TOP Categories - Coeff in response_max_spire_fashion_luxurywatch_model




### Variables in TOP Categories - Coef in response_max_spire_fashion_luxuryretail_model



# 2.4 Examine 3 Dimensions of Coefficients

We will look into 3 different dimensions of coefficients:

- Significance of impact of variables (Top 20 median)
- Consistency of impact of variables (95% CI)
- Correlation of coefficients change and auc change (Top 20)

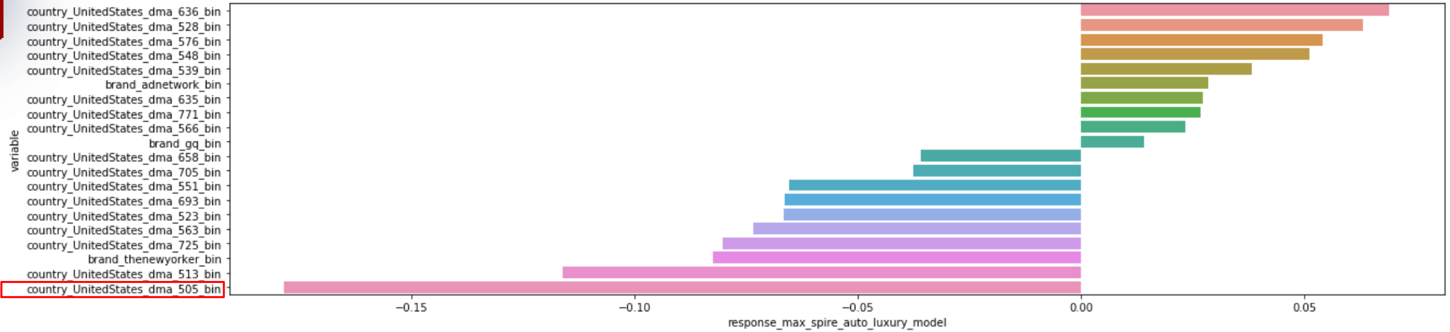
**Goal**: To see if any variable has high significance of impact, consistent impact and is highly correlated with auc change.

2.4

# Examine 3 Dimensions of Coefficients

*Auto_luxury_model*

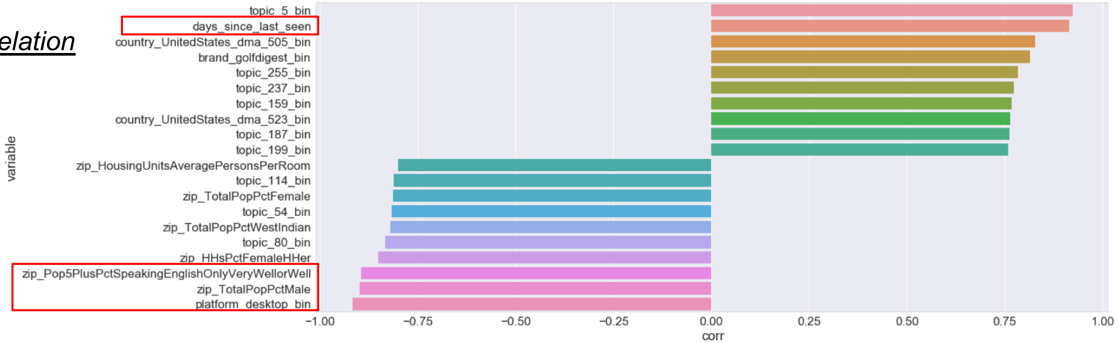
*Significance*




*Consistency*

**Variables that always have positive impact**: ['zip_AsianAlonePopPctAsianIndian', 'zip_RenteroccupiedHousingUnitsPct12501499', 'zip_TotalPopPctWestIndian'].
**Variables that always have negative impact**: ['country_UnitedStates_dma_505_bin', 'zip_EmployedPop16PlusPctAgriculture', 'zip_HousingUnitsPctMobileHomes', 'zip_OccupiedHousingUnitsPctBottledTankOrLPGas', 'zip_OccupiedHousingUnitsPctUtilityGas', 'zip_RenteroccupiedHousingUnitsPct400499'].

*Correlation*

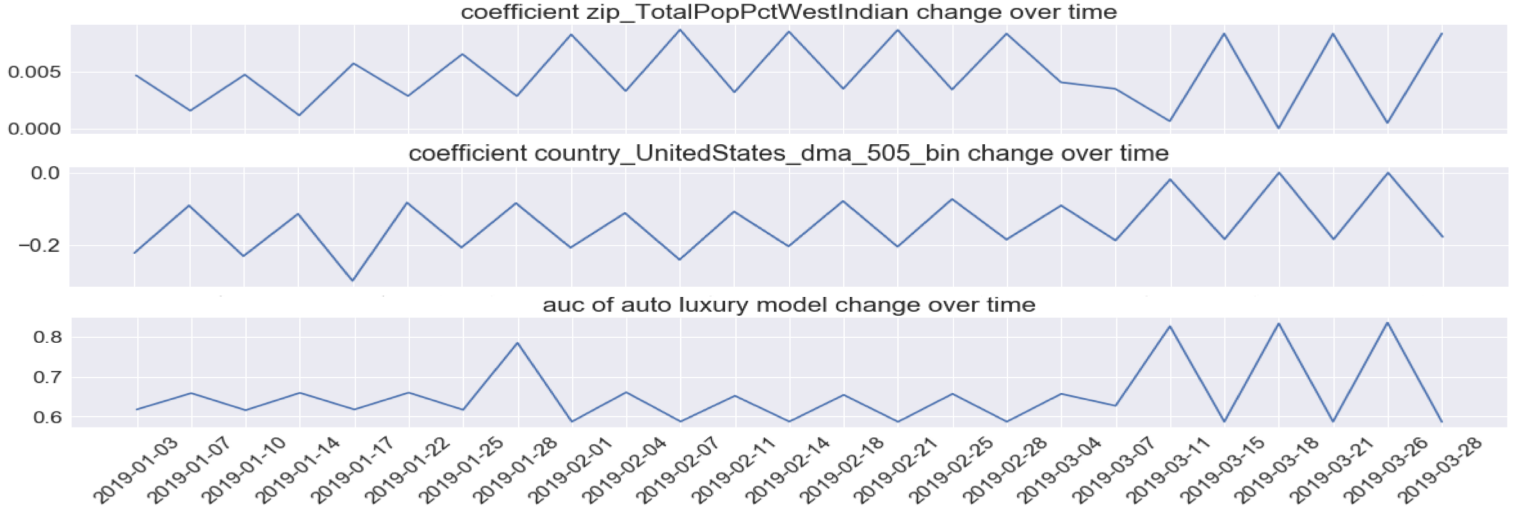


# Examine 3 Dimensions of Coefficients



**Result:**

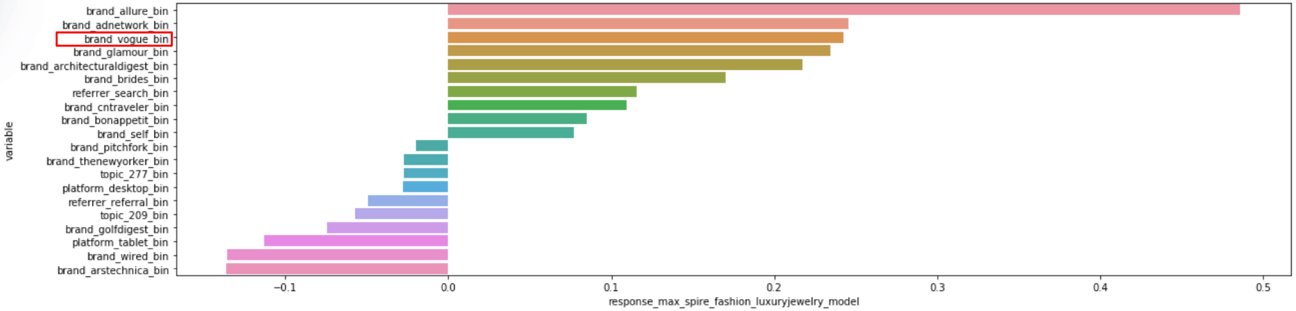
**zip_TotalPopPctWestIndian** is highly negatively correlated with auc, has consistent positive impact, not significant impact on prop score.
**country_UnitedStates_dma_505_bin** is highly positively correlated with auc, has consistent negative impact, has significant negative impact on prop score.



# Examine 3 Dimensions of Coefficients

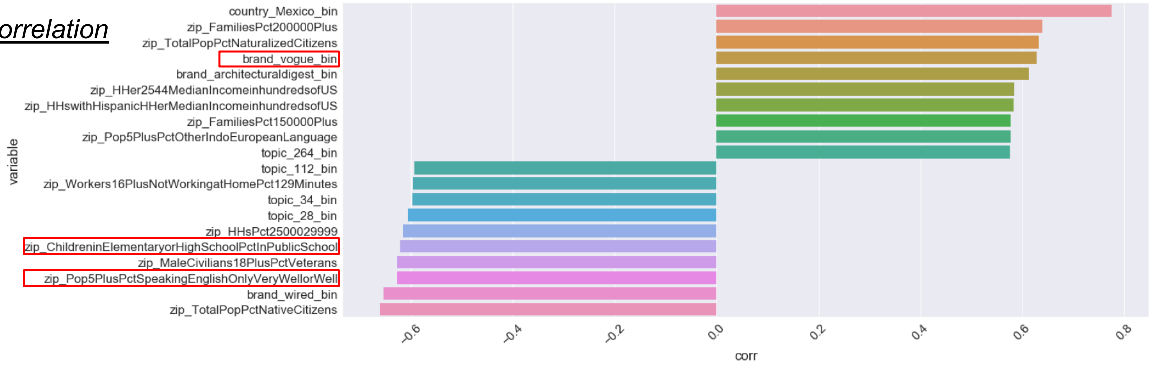
*Fashion_luxuryjewelry_model*

*Significance*




*Consistency* **Variables that always have positive impact**: ['lifespan'].

*Correlation*

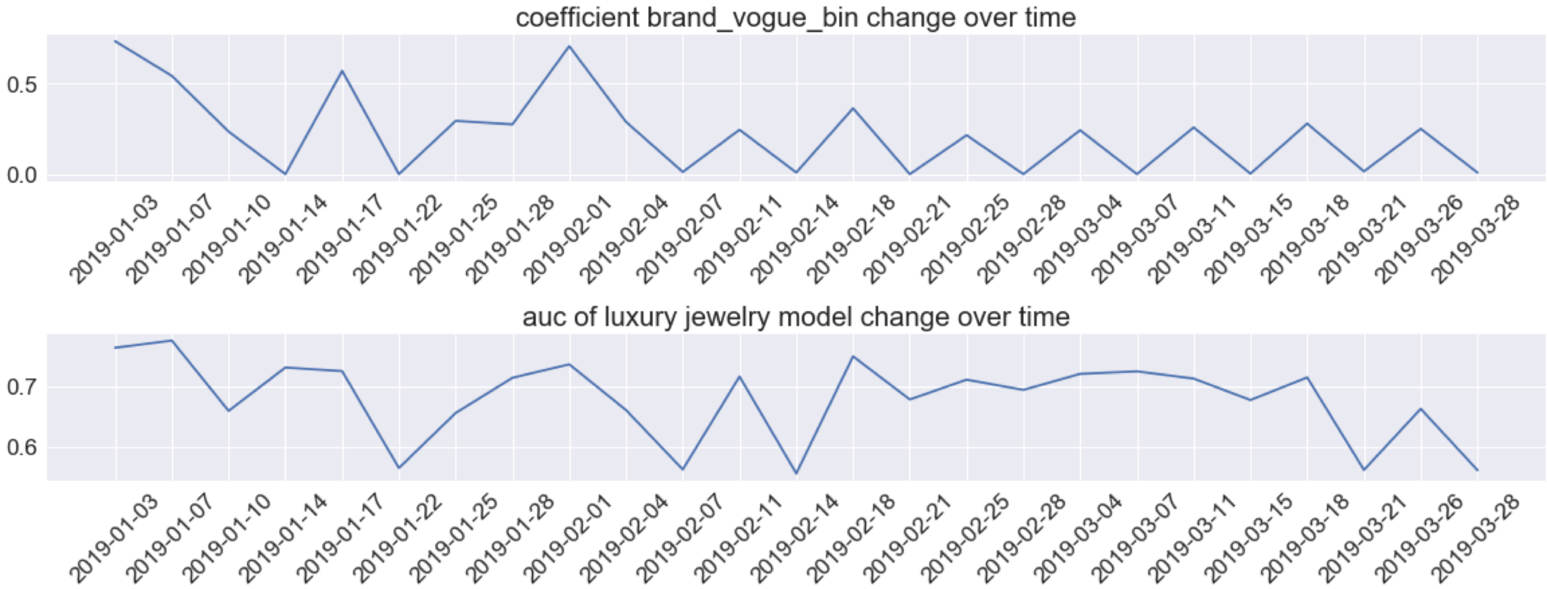


# Examine 3 Dimensions of Coefficients

*Fashion_luxuryjewelry_model*

**Result:**

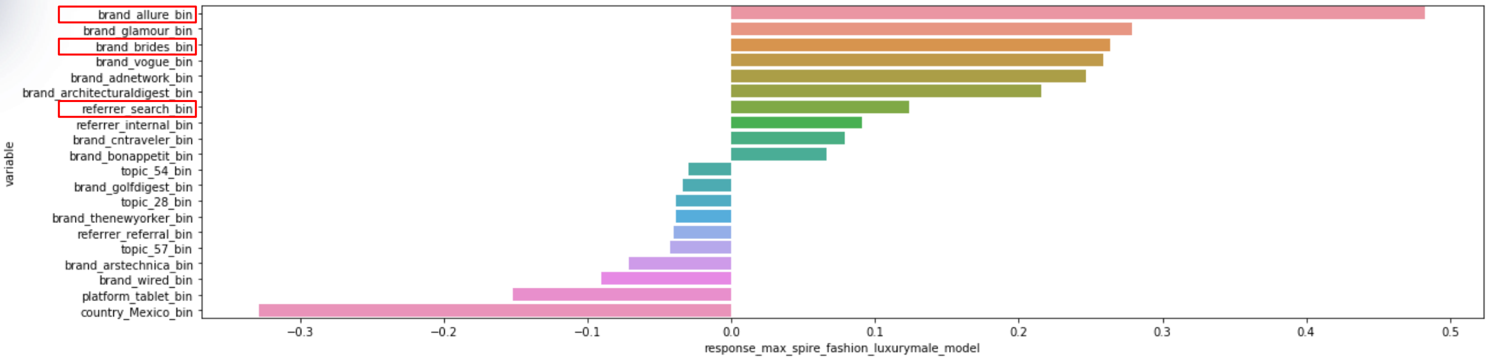
**Brand_vogue_bin** is highly correlated with auc, has high significance on prop score, 80% consistent positive impact.



# 2.4 Examine 3 Dimensions of Coefficients

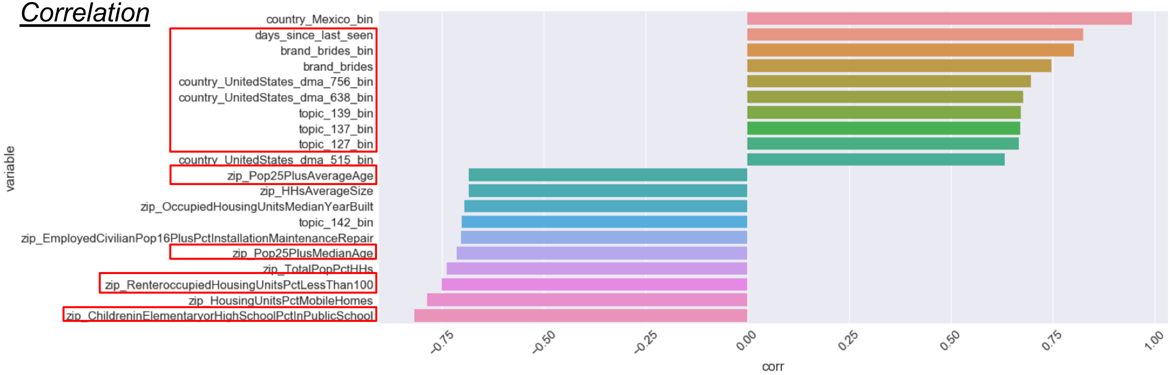
*Fashion_luxurymale_model*

*Significance*




*Consistency* **Variables that always have positive impact**: ['referrer_search_bin'].

*Correlation*



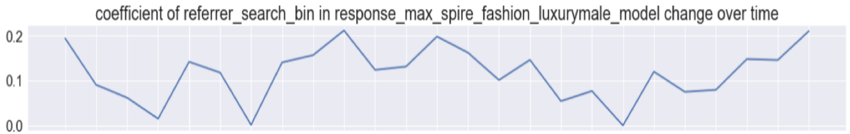
# 2.4 Examine 3 Dimensions of Coefficients

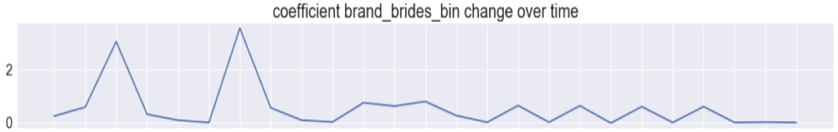
*Fashion_luxurymale_model*

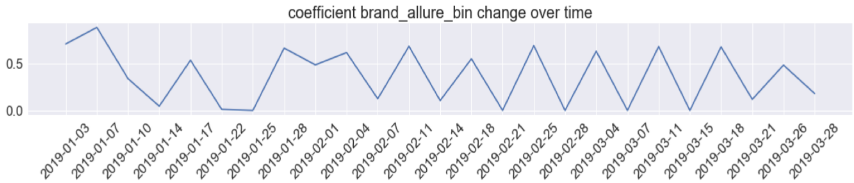
**Result:**

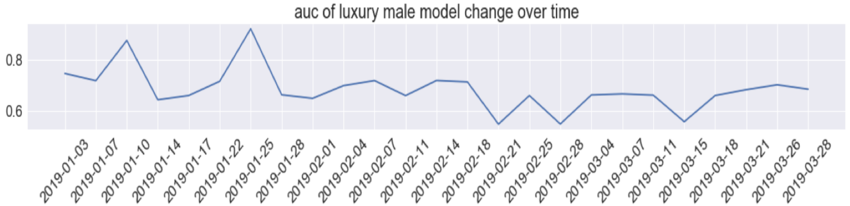
**Referrer_search_bin** and **brand_allure_bin** (90%CI) always have significant positive impact on propensity score, but the auc score doesn’t correlate with coefficients change (corr = -0.06).

**Brand_brides_bin** is highly correlated with auc, has high significance on prop score, 80% consistent positive impact.



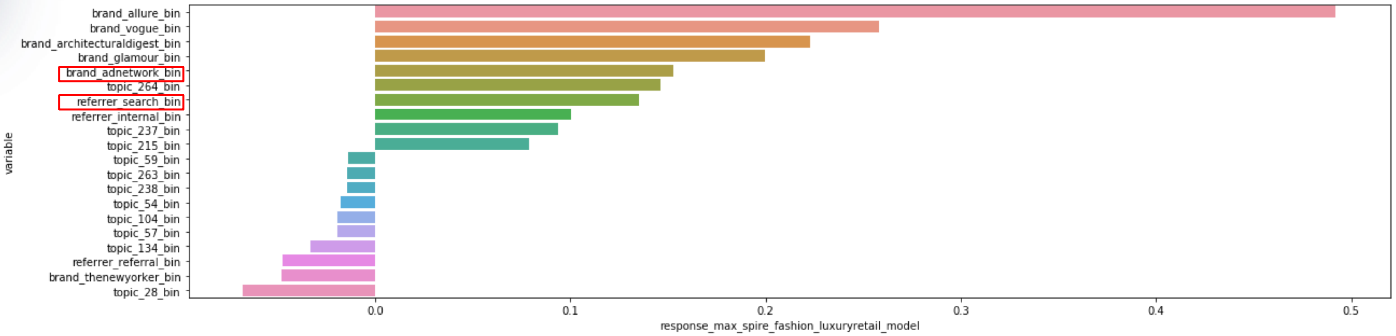



# 2.4 Examine 3 Dimensions of Coefficients

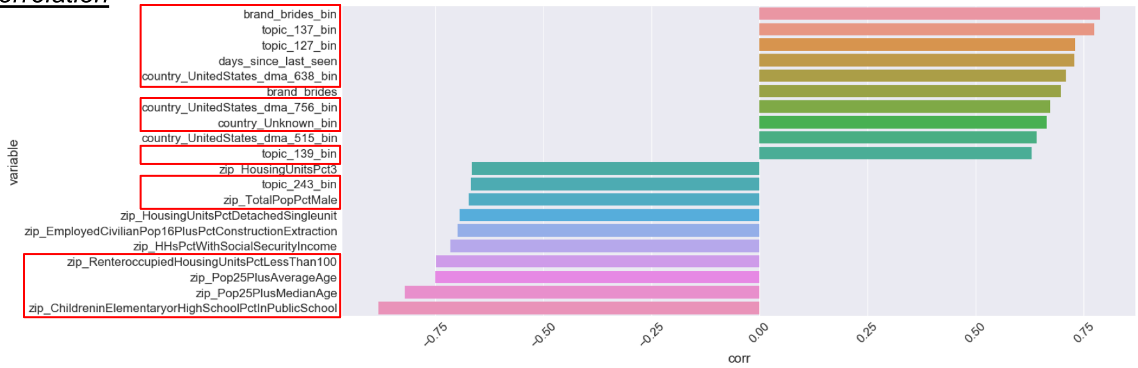
*Fashion_luxuryretail_model*

*Significance*




*Consistency* **Variables that always have positive impact**: ['brand_adnetwork_bin', 'lifespan', 'referrer_search_bin'].

*Correlation*

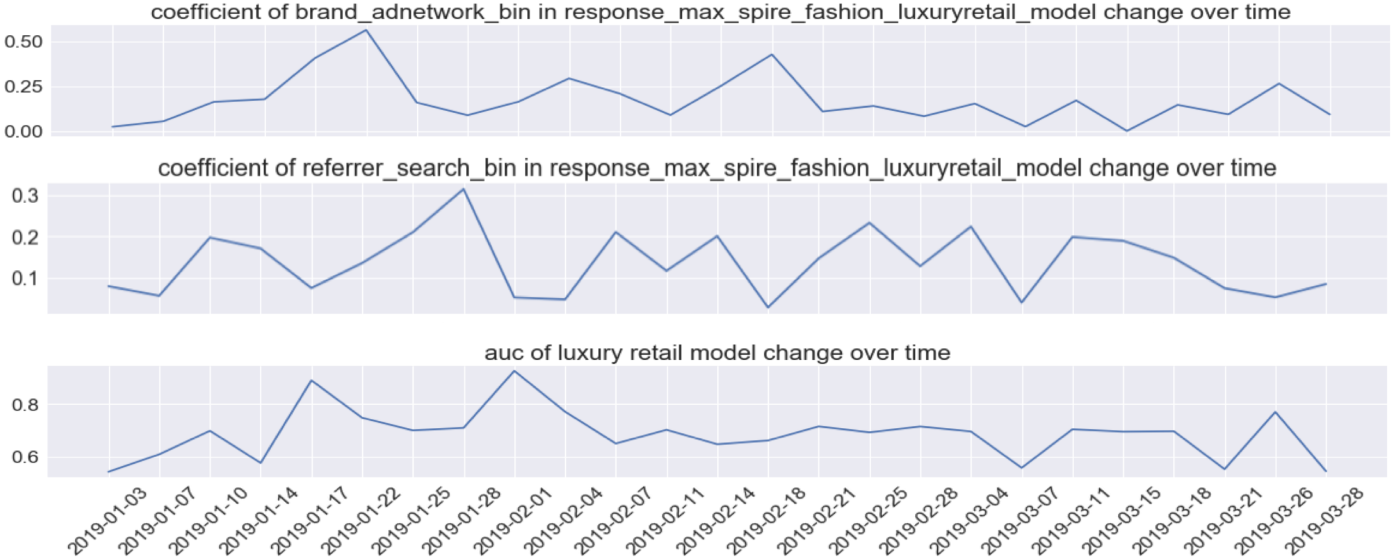


# 2.4 Examine 3 Dimensions of Coefficients

*Fashion_luxuryretail_model*

**Result:**

**Brand_adnetwork_bin** and **referrer_search_bin** always have significant positive impact on propensity score, the auc score doesn't correlate with coefficients change.

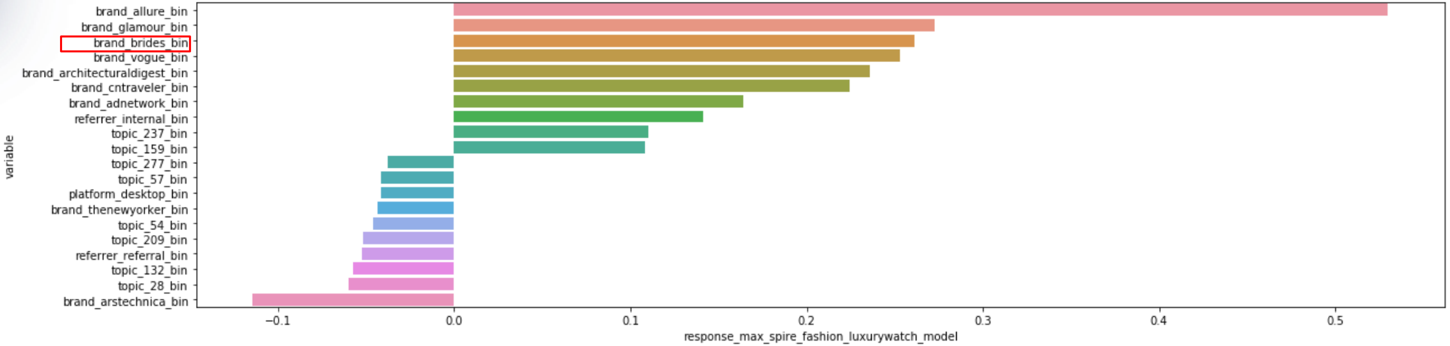


2.4

# Examine 3 Dimensions of Coefficients

*Fashion_luxurywatch_model*

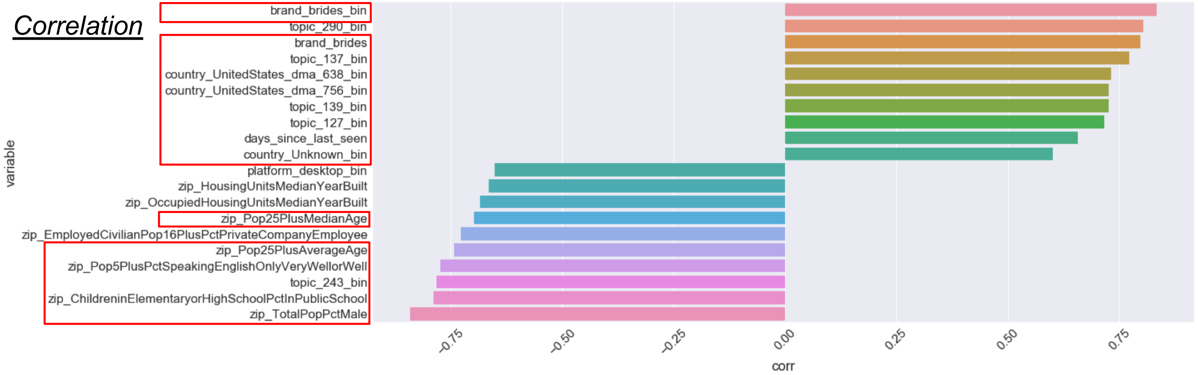
*Significance*




*Consistency*

**Variables that always have positive impact**: ['lifespan'].
**Variables that always have negative impact**: ['platform_desktop_bin', 'zip_Pop5PlusPctSpeakingEnglishOnlyVeryWellorWell', 'zip_TotalPopPctMale'].

*Correlation*



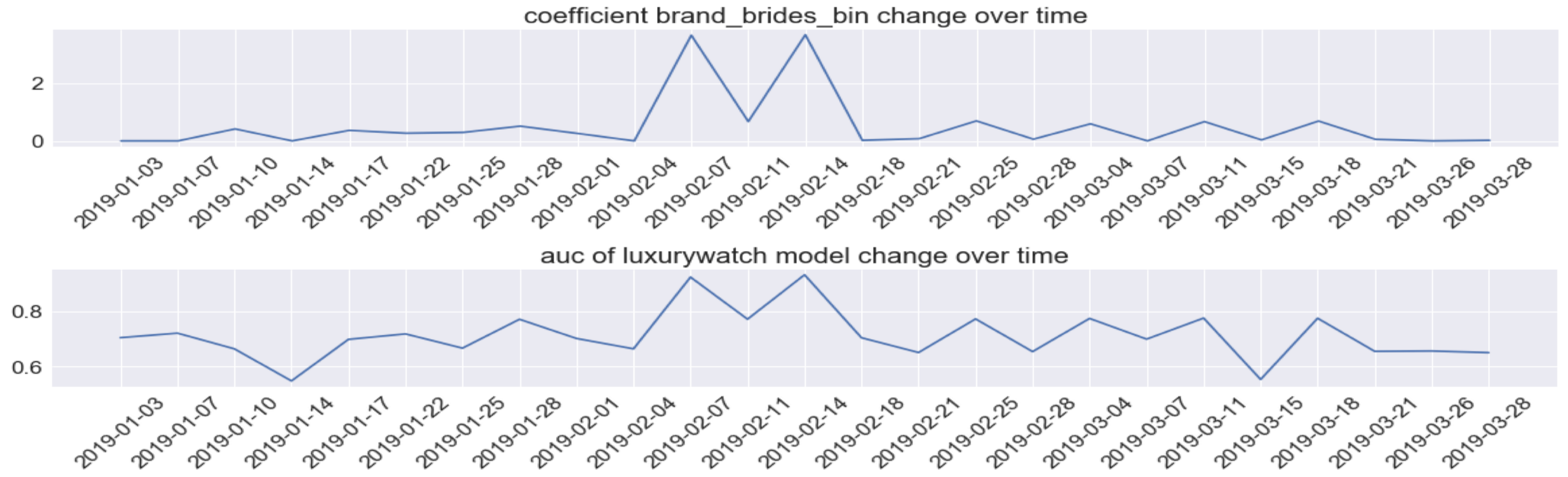
# Examine 3 Dimensions of Coefficients



**Result:**

**Brand_brides_bin** is highly correlated with auc, has high significance on prop score, 80% consistent positive impact

Valentine Week has impact on luxury watch model?



# 2.4 Examine 3 Dimensions of Coefficients



## Common impact on three models

Common highly correlated variables among luxury watch, luxury retail and luxury male model:

**Brand_brides_bin, topic_127_bin, topic_137_bin, topic_139_bin, country_UnitedStates_dma_638_bin, country_UnitedStates_dma_756_bin**

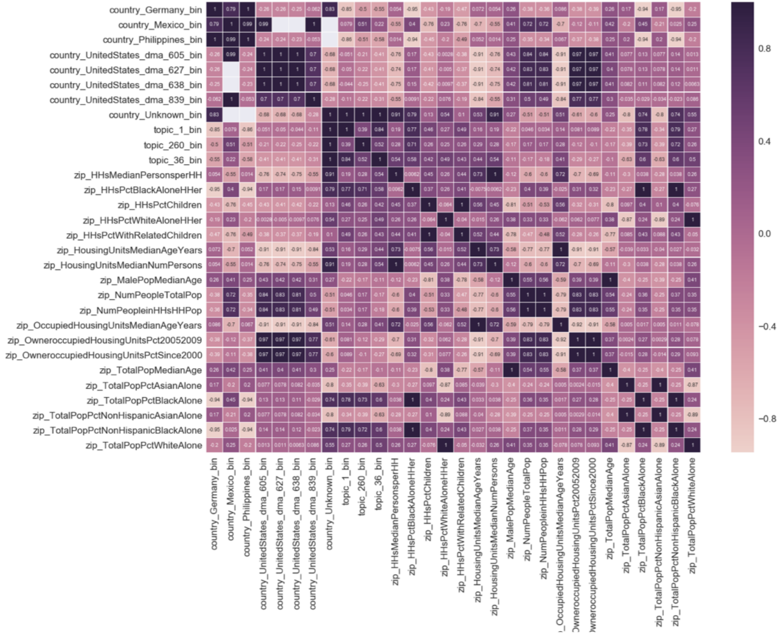
The three models are all highly correlated to these variables.

These are significant indicators of prop score.

Special dates for luxury retail model: 1.17 and 2.1.
Special dates for luxury male model: 1.10 and 1.25.
Special dates for luxury watch model: 2.7 and 2.14.

# 2.5 Correlation Between Variable Trends

Check the correlation between variable trends, to see if any pairs of variable change together. It turns out there are many variables' trends highly correlated in all models.



# 3.1 Persona Analysis on Total Most Valuable Customers

- **Who are the most valuable customers?**

The most valuable customers refer to customers in  bucket1. ‘Conde Nast group all the scores into deciles, and bucket 1 will be the first decile and will have the highest propensity scores for that model’.

- **Why should we study this group of people?**

Propensity score is the likelihood that a visit will purchase/search for an advertiser’s product. Thus people in bucket 1 with high propensity scores can be considered as the most valuable customers for both Conde Nast and advertisers.

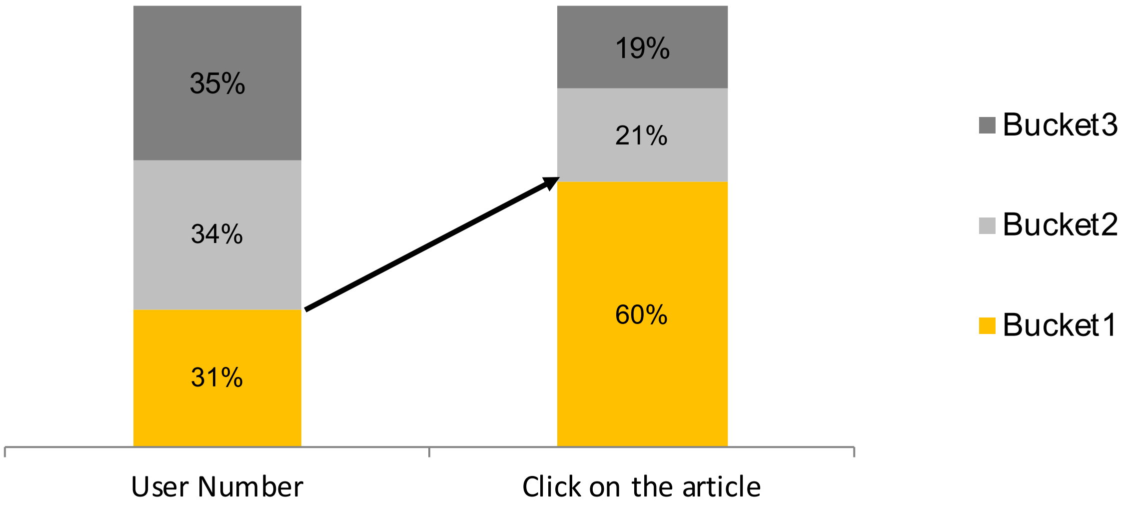
- **What is the aim of the analysis?**

This analysis tells who are these people and how to reach them using the web activity data.

- Database used in this section is score database and web_actibity_173.

# 3.1 Persona Analysis on Total Most Valuable Customers

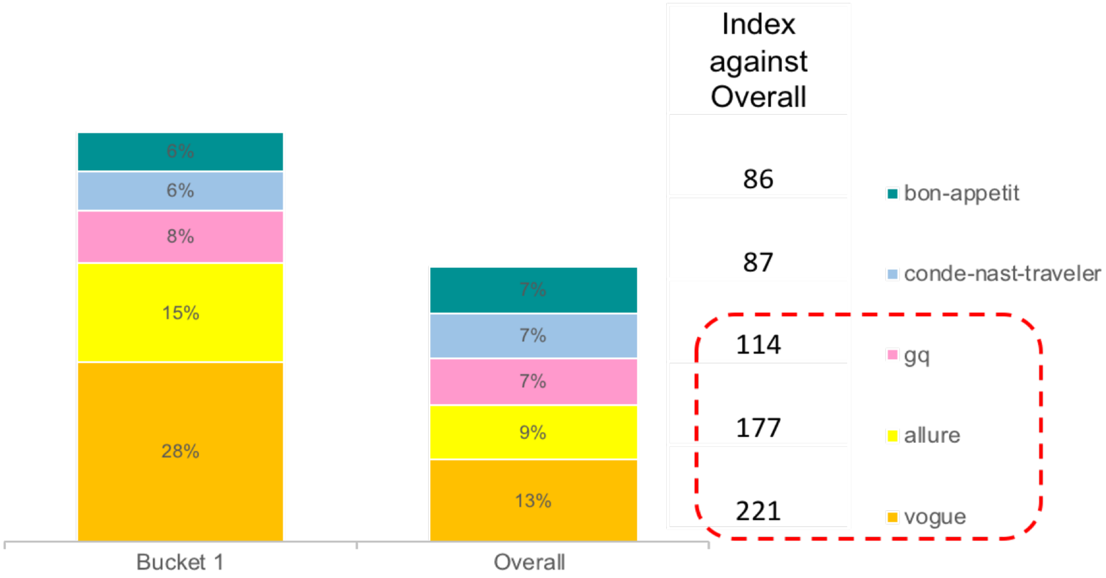
**Clicks on the article** After combining the data with web activities, Bucket 1 contained 31% of the total users, who contributed to over 60% of the clicks on the article.




Note: Click on the article refers to the records, whose scroll_depth is not NA.

# 3.1 Persona Analysis on Total Most Valuable Customers

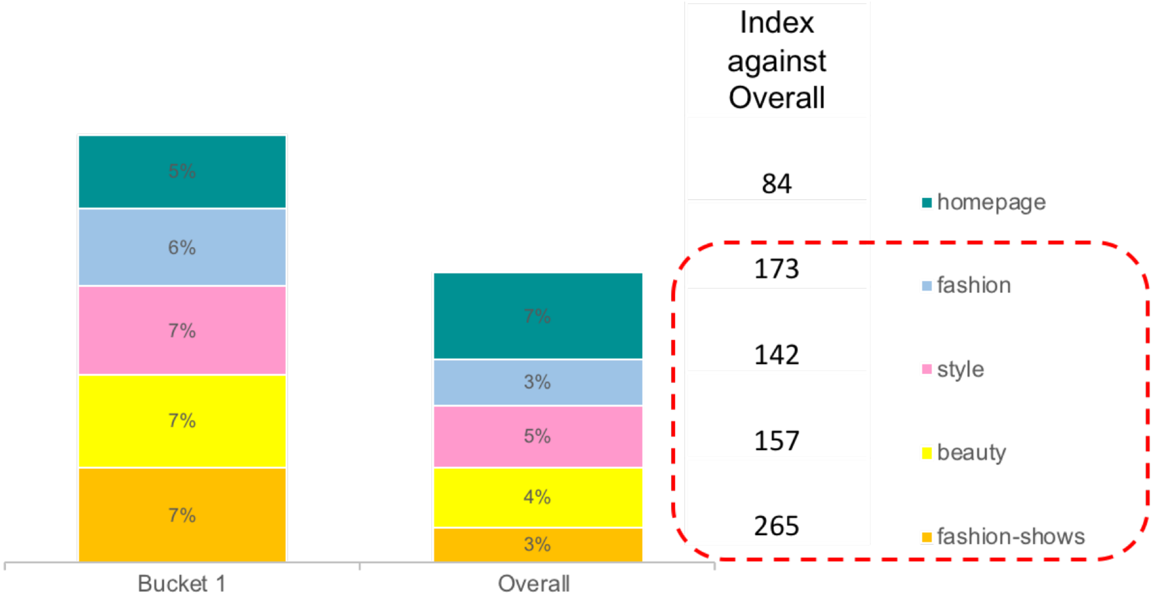
**Top 5 Brand%** Vogue, Allure, GQ are the top 3 brands for bucket 1, favored by over 50% of the customers. Moreover, bucket 1 even more skewed to the top 3 brands comparing with the overall customers, especially for vogue.




Note: Index = percentage in Bucket1 / percentage in Overall *100

# 3.1 Persona Analysis on Total Most Valuable Customers

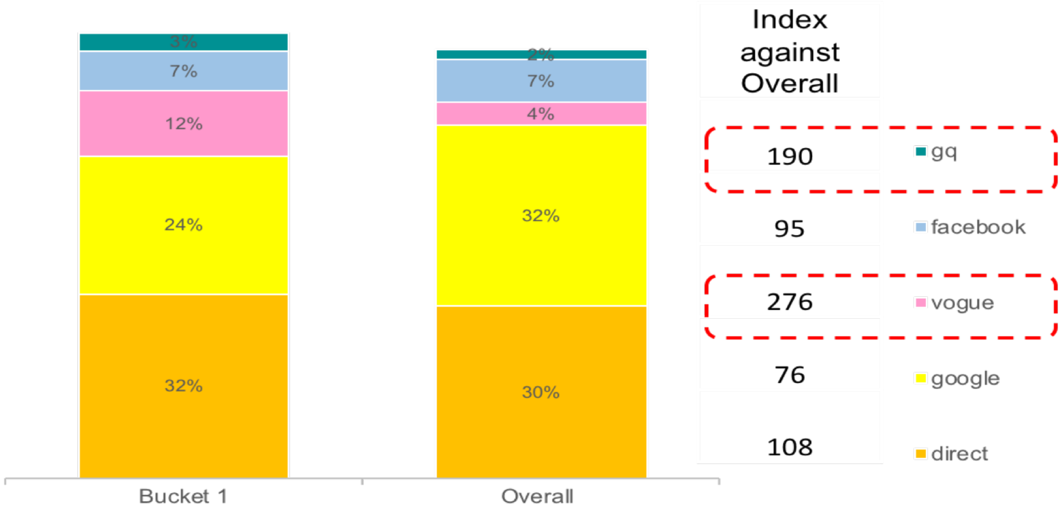
**<u>Top 5 Channel%</u>** Channel is sparser than brands. Bucket 1 shows more preference in beauty & fashion related channels.




Note: Index = percentage in Bucket1 / percentage in Overall *100

# 3.1 Persona Analysis on Total Most Valuable Customers

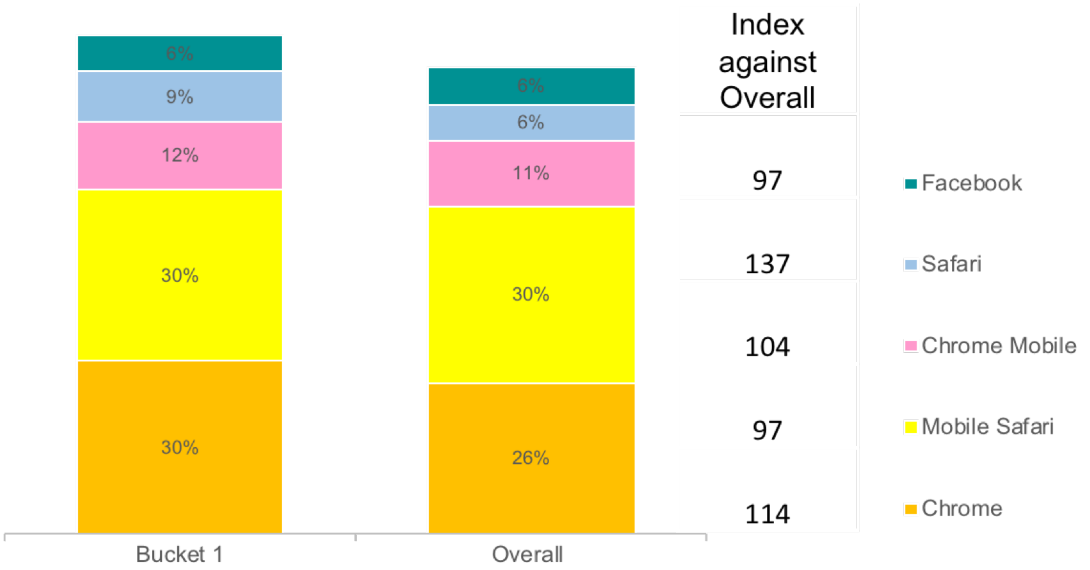
**<u>Top 5 Referrer%</u>** 1) Direct is the largest referrer for bucket1 customers.
2) Bucket 1 also skewed to fashion&style-related referrers, such as Vogue and GQ, which is aligned with the channel information.




Note: 1. Index = percentage in Bucket1 / percentage in Overall *100
2. Referrer is not in the original dataset. It was generated from the ReferrerURL.

# 3.1 Persona Analysis on Total Most Valuable Customers

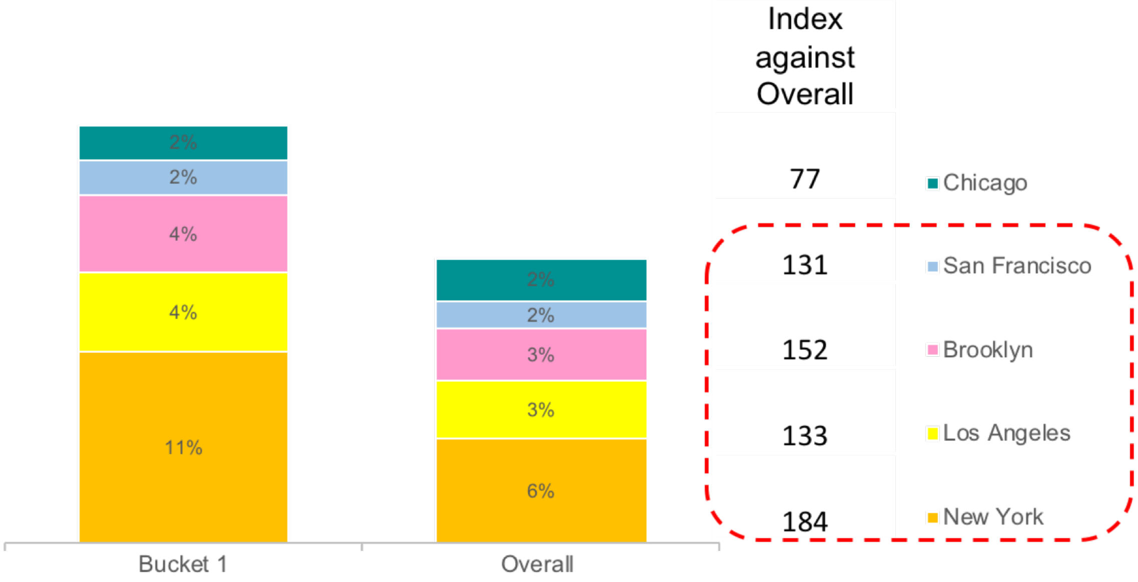
**Top 5 Browser%** Around 60% of the access is through Chrome and mobile Safari. Bucket 1 customers show slight preference on Chrome browser.




Note: Index = percentage in Bucket1 / percentage in Overall *100

# 3.1 Persona Analysis on Total Most Valuable Customers

**Country& Top 5 City** 1) Country wise, for both bucket 1 and overall customers , around 98% of the customers are from United States. 2) Within the US, New York, Brooklyn and Los Angeles are the top 3 cities.




Note: 1. Index = percentage in Bucket1 / percentage in Overall *100
2. Country is filtered to US only when calculating city%.

# 3.1 Persona Analysis on Total Most Valuable Customers

## Potentional opportunity: Poland

% of Total SUM(Numbe...
31.00% 98.73%

### Most Valuable Users* Ratio by Country

| Buck.. | Sweden | Poland | Israel | China | Italy | Australia | Mexico | Puerto Rico | Singapore | United States | United Kingdom | Hong Kong | Canada | France | Germany |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 98.73% | 95.73% | 81.53% | 77.36% | 76.87% | 73.99% | 63.64% | 62.51% | 60.37% | 54.99% | 54.51% | 54.31% | 53.72% | 53.36% | 33.84% |

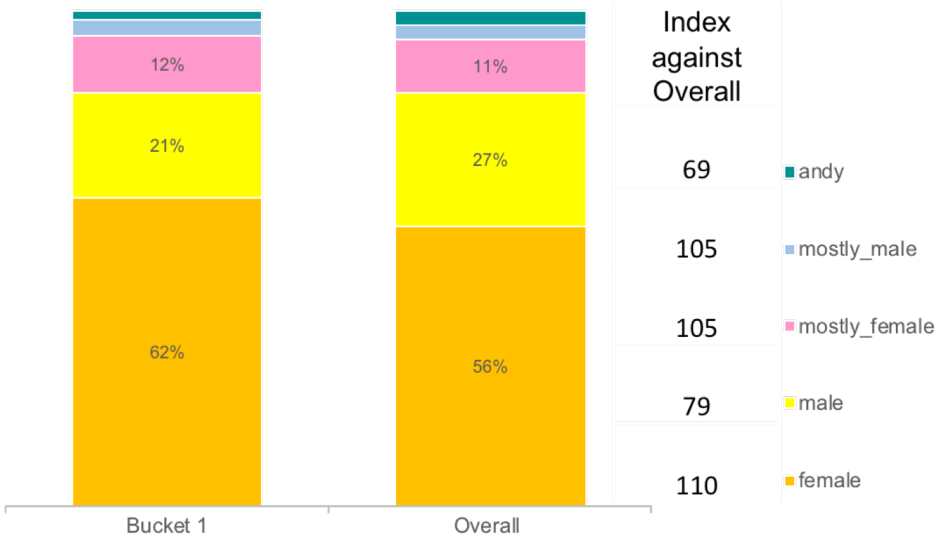
### Market Share by Country

| Bucket | Sweden | Poland | Israel | China | Italy | Australia | Mexico | Puerto Rico | Singapore | United States | United Kingdom | Hong Kong | Canada | France | Germany |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.04% | 0.25% | 0.12% | 0.12% | 0.06% | 0.07% | 0.12% | 0.11% | 0.06% | 98.27% | 0.24% | 0.04% | 0.34% | 0.10% | 0.07% |

Note: 1. Most Valuable users refer to users in Bucket 1
2. Market share less than 0.04% is excluded

# 3.1 Persona Analysis on Total Most Valuable Customers

**Author Gender%** Female authors were favored by both bucket 1 and overall customers.




Top 5 favorable authors for Bucket 1

**1. Christian Allaire**
(Fashion and Style Writer at Vogue)

**2. The Epicurious editors**

**3. the editors of GQ**

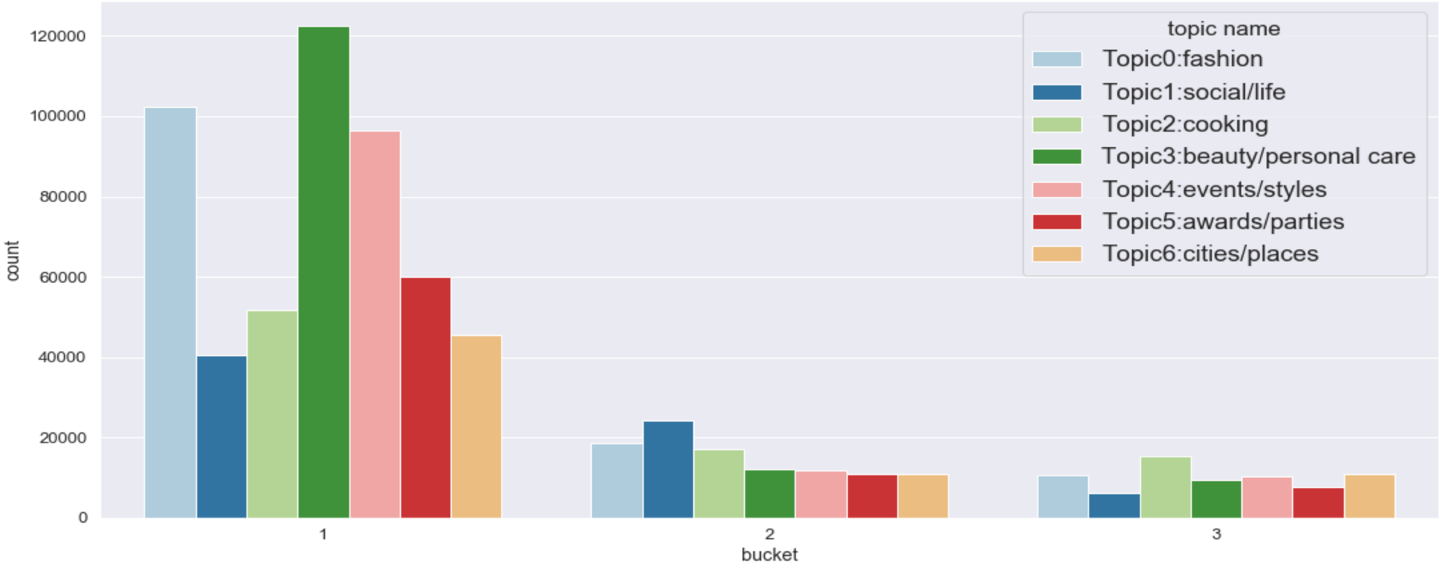
**4. Cnt editors**

**5. Megan Gustashaw**
(Freelance Writer)

Note: 1. Index = percentage in Bucket1 / percentage in Overall *100
2. Gender is not in the original dataset. It was generated from the name of the author vis gender-guesser 0.4.0.
3. Andy is when the name is valid equally for both male and female.

# 3.1 Persona Analysis on Total Most Valuable Customers

**<u>Topic distribution across bucket</u>** Bucket 1 customers are most interested in articles about beauty, fashion and events/styles.



# Persona Analysis on Total Most Valuable Customers

## One-slide summary

Bucket 1 customers

**Click on the article**
More likely to click on the article

**Referer**
- Direct is the largest referrer
- Skew to fashion and style-related referers, such as Vogue and GQ

**Brand**
More preference on Vogue, Allure, GQ

**Browser**
Mostly through Chrome and mobile Safari

**Channel**
Shows more preference on beauty and fashion related channels.

**Author**
- Read more articles from female authors.
- Top 1 author is Christian Allaire.

**Topic**
More interested in articles about beauty, fashion and events/styles.

**Location**
- More resources on top 3 cities: New York, Brooklyn and Los Angeles.
- Poland would be the first to consider, if the company would like to expand overseas market.

# Persona Analysis on Most Valuable Customers in Each Model

90%+ of all bucket 1 customers have high intention of purchasing only one categories among all 5 different categories.

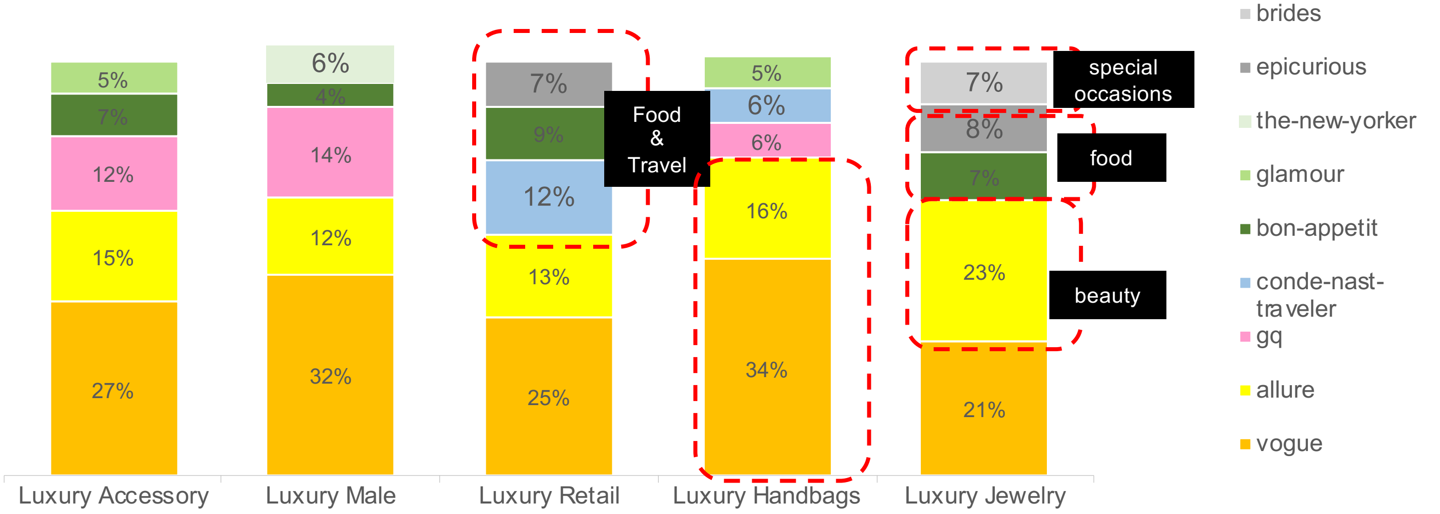
Note: Five models here refers to models for luxury accessories, luxury male, luxury retail, luxury handbags, luxury jewelry.

# 3.2 Persona Analysis on Most Valuable Customers in Each Model

**Distribution**: Model luxury male has the highest ratio of the most valuable  users while model luxury jewelry has the lowest.

# 3.2 Persona Analysis on Most Valuable Customers in Each Model

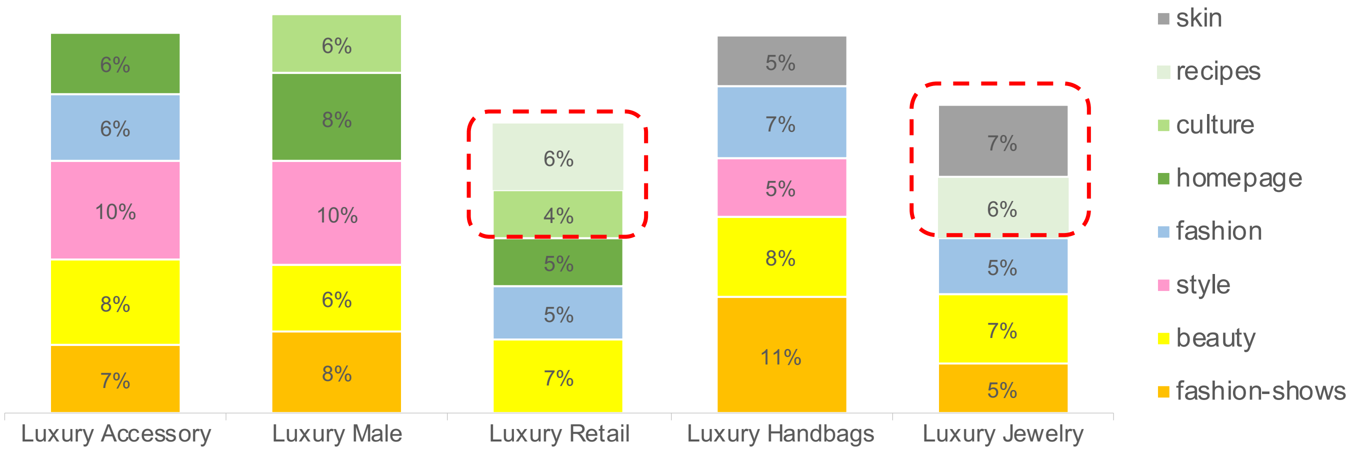
**<u>Top 5 Brand%</u>** 1) Vogue is the largest brand for all models, except for Luxury jewelry.
2) Luxury retail customers cares more about the food & travel, and less about fashion and style comparing with others. 3) Customers who are likely to purchase luxury jewelry show their interest in beauty, food, and special occasions.




Note: For each model, only top5 brands are listed.

# 3.2 Persona Analysis on Most Valuable Customers in Each Model

**Top 5 Channel%** 1) Channel for luxury retail customers are sparser , top 5 only account for 27% of total channels. New channel, like culture and recipes, pop up.
2) Luxury jewelry potential buyers show interest in skin and recipes channels, which indicates that they might have a delicate lifestyle.




Note: For each model, only top5 channels are listed.

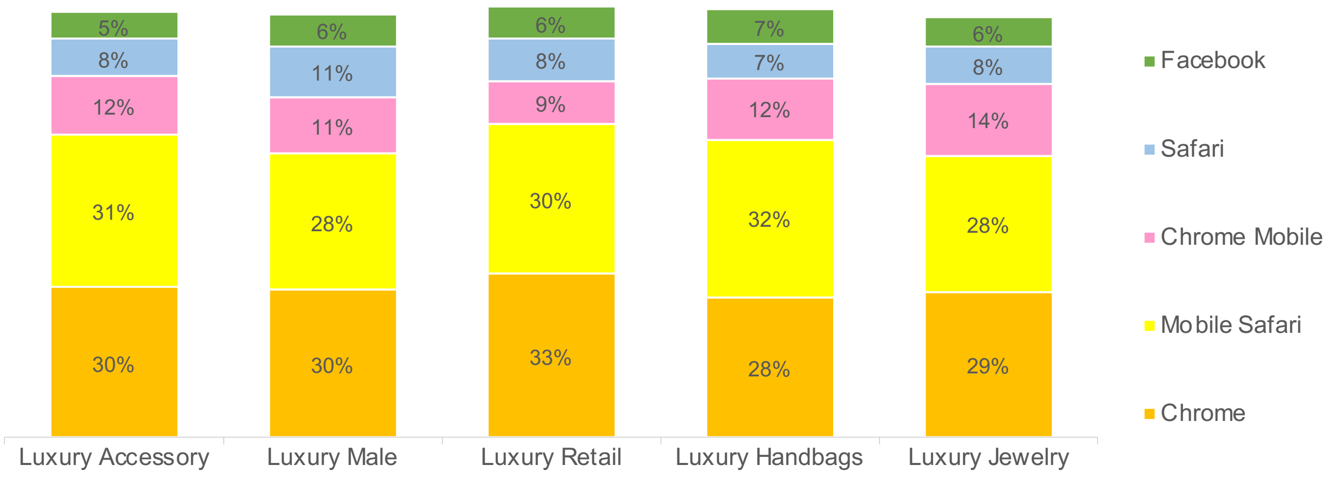
# 3.2 Persona Analysis on Most Valuable Customers in Each Model

**<u>Top 5 Referer%</u>** Luxury jewelry potential buyers shows relatively higher interest in google, and Luxury retail potential buyers shows relatively higher interest in direct.


Note: For each model, only top5 referers are listed.

# 3.2 Persona Analysis on Most Valuable Customers in Each Model

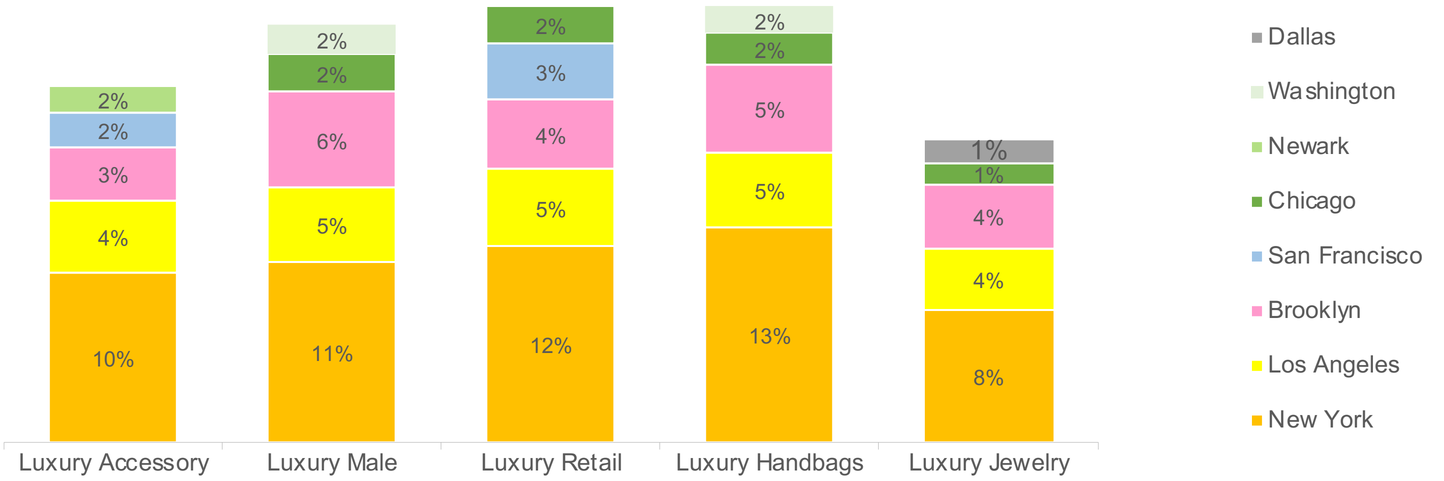
**Top 5 Browser%** Around 60% of the access is through Chrome and mobile Safari. Similar pattern can be found across all models.




Note: For each model, only top5 browsers are listed.

# 3.2 Persona Analysis on Most Valuable Customers in Each Model

**Top 5 City%** Country wise, around 98% of the customers are from United States for all models. Within the US, New York, Brooklyn and Los Angeles are the top 3 cities.




Note: 1. For each model, only top5 cities are listed.
2. Country is filtered to US only when calculating city%.

# 3.2 Persona Analysis on Most Valuable Customers in Each Model

**Author gender%** Female writers are more favored by luxury jewelry protentional buyers, and male writers more by luxury male buyers.


Note: 1. Gender is not in the original dataset. It was generated from the name of the author vis gender-guesser 0.4.0.
2. Andy is when the name is valid equally for both male and female.

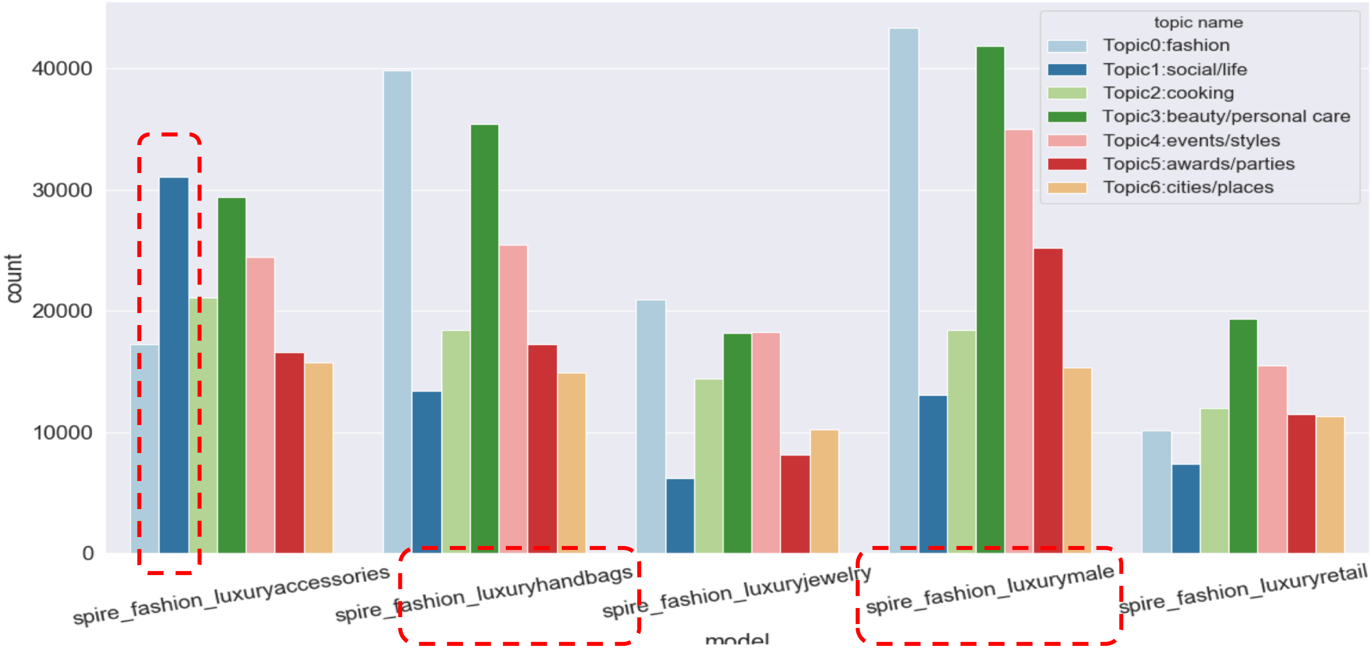
# 3.2 Persona Analysis on Most Valuable Customers in Each Model

**Top5 Authors for each model**

| Luxury Accessory | Luxury Male | Luxury Retail | Luxury Handbags | Luxury Jewelry |
|---|---|---|---|---|
| **Megan Gustashaw** (Freelance Writer) | **the editors of GQ** | **CNT editors** | **Christian Allaire** (Fashion and Style Writer at Vogue) | **the Epicurious editors** |
| **Christian Allaire** (Fashion and Style Writer at Vogue) | **Megan Gustashaw** (Freelance Writer) | **the Epicurious editors** | **Nicole Phelps** (Director, Vogue) | **Liana Schaffner** (associate editor at allure magazine) |
| **the editors of GQ** | **Nicole Phelps** (Director, Vogue) | **Caitlin Morton** (Digital Editor at CNT) | **Jenna Rennert** (Beauty Editor at Vogue) | **Christian Allaire** (Fashion and Style Writer at Vogue) |
| **Jenna Rennert** (Beauty Editor at Vogue) | **Christian Allaire** (Fashion and Style Writer at Vogue) | **Jenna Rennert** (Beauty Editor at Vogue) | **Sarah Mower** (Chief critics at Vogue) | **Melanie Rud Chadwick** (Writer & Beauty expert) |
| **Nicole Phelps** (Director, Vogue) | **Jenna Rennert** (Beauty Editor at Vogue) | **Christian Allaire** (Fashion and Style Writer at Vogue) | **CNT editors** | **Jenna Rennert** (Beauty Editor at Vogue) |

# 3.2 Persona Analysis on Most Valuable Customers in Each Model

**Topic distribution across models**. 1) customers in luxury accessories model are more interested in articles about social/life, while this topic is the least popular in the luxury jewelry model.
2) The distribution for luxury handbags and luxury male are quite similar.



3.2

# Persona Analysis on Most Valuable Customers in Each Model

## Word Cloud
(title and topic)

### Luxury retail model


### Luxury male model


### Luxury accessories model


### Luxury handbags model


### Luxury jewel mode


Note: Most Valuable users refer to users in Bucket 1

# 3.2 Persona Analysis on Most Valuable Customers in Each Model

**To wrap up**

- Luxury jewelry customers show more interest in skin& beauty.
- Luxury retail customers are activities-enthusiast, who show interest in food and traveling.
- Topic distribution for luxury male and luxury handbags are similar. Both of them like fashion and wear-related contents, except for the male wear.

Note: Most Valuable users refer to users in Bucket 1

4

# Models Built to Play With web_activity Data

- **Logistic model**
- **Random Forest**
- Neural Network
- Linear regression

# 4 Workflow

- Data cleaning
- Feature engineering
- EDA
- Build models
- Model insights

Check NA; Fill in na; Transform to lowercase;

Transform scrolldepth to binary classification; Extract the first author; Extract root url; Add gender column; Text feature

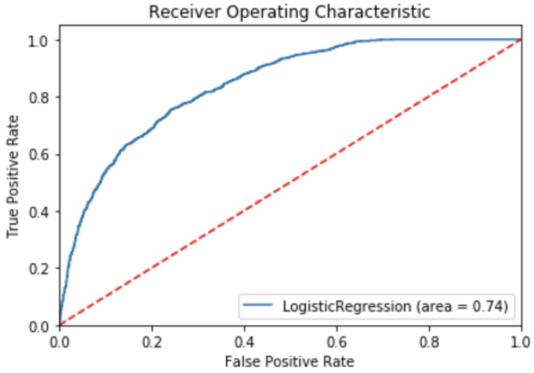
Check number of total user, number of active user, scroll depth, and conversion rates from different dimensions

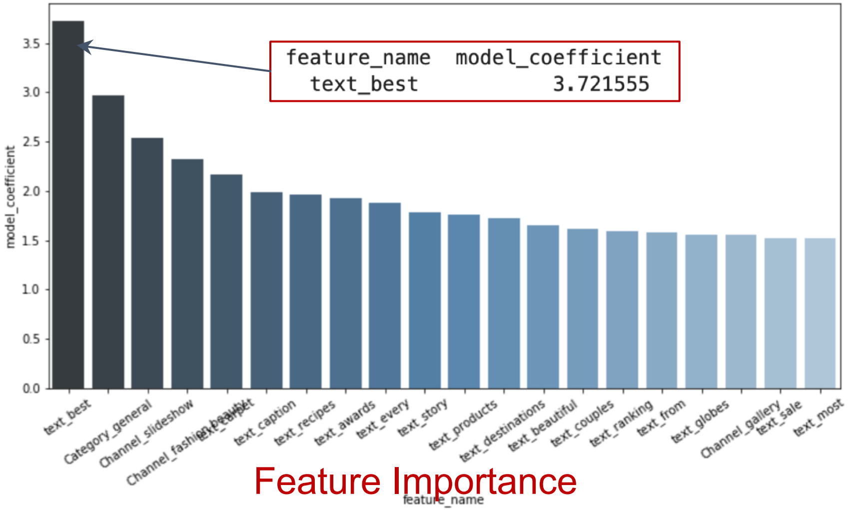
Train/test split, Train model, Fit model, Evaluate model

# 4.1 Logistic Regression

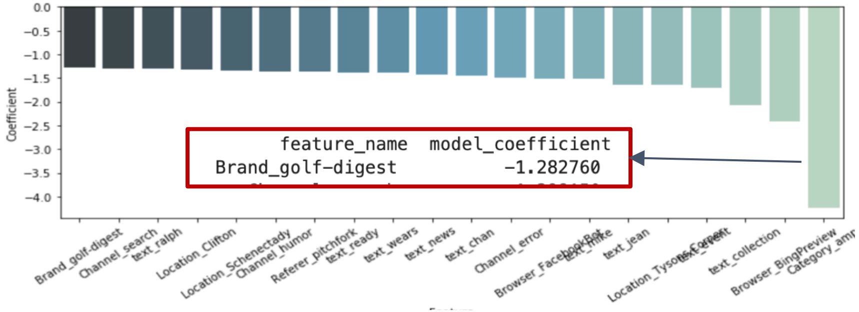
**Predictors**: features extracted from text mining(TF-IDF), referrer, brand, channel, category, location, author_gender)
**Target**: scroll_depth_0_1

**Accuracy** of logistic regression classifier on test set: 0.77
**ROC_AUC Score** of logistic regression classifier on test set: 0.74




Feature Importance



# 4.2 Random Forest Model

## H2O Random Forest for Variables Pre-Selection

- Important Variables (fig.1 in Appendix)
- Select Top 6 variables for building predictive model

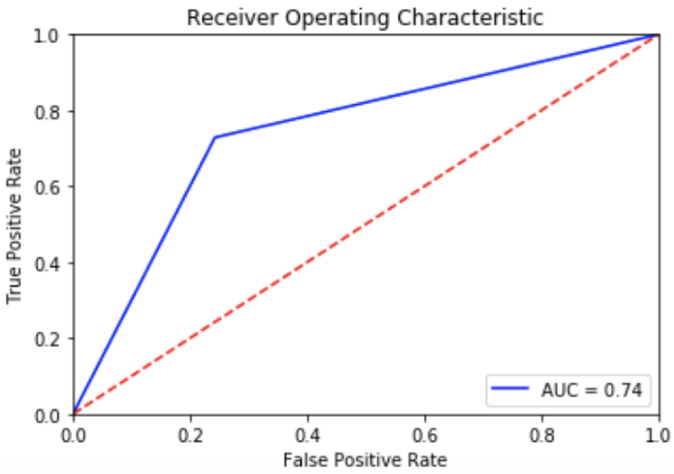
## Build Random Forest Model (Scikit-Learn)

- RandomForestClassifier(n_estimators = 650,max_depth = 50)
- Target: ‘ScrollDepth_0_1’

  Predictors: 'Brand', 'Channel', 'Category', 'Referrer', 'Author1','gender'

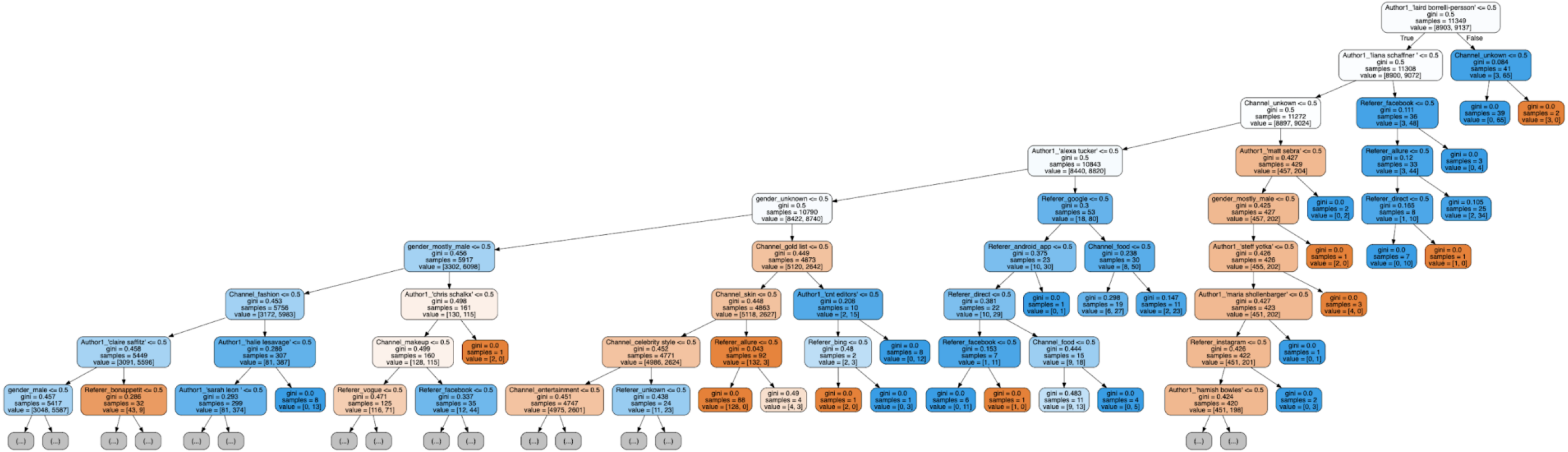
# 4.2 Result and Insights

```
        0     1
0    2835   907
1     613  1645
              precision    recall  f1-score   support

           0       0.82      0.76      0.79      3742
           1       0.64      0.73      0.68      2258

    accuracy                           0.75      6000
   macro avg       0.73      0.74      0.74      6000
weighted avg       0.76      0.75      0.75      6000
```




| | importance |
|---|---|
| Category_general | 0.103533 |
| Category_amp | 0.098866 |
| gender_unknown | 0.059094 |
| gender_female | 0.032511 |
| Brand_conde-nast-traveler | 0.023192 |
| Referer_google | 0.021983 |
| Referer_direct | 0.021708 |
| Brand_vogue | 0.021150 |
| Referer_facebook | 0.013095 |
| Channel_unkown | 0.012873 |
| Author1_'cnt editors' | 0.011440 |
| Channel_the bests | 0.010055 |
| gender_mostly_female | 0.009900 |
| Channel_[object object] | 0.009663 |
| gender_male | 0.009102 |
| Author1_'vanity fair magazine' | 0.008872 |
| Brand_vanity-fair | 0.008082 |
| Channel_magazine | 0.007506 |
| Channel_the places | 0.006954 |
| Channel_style | 0.006486 |

# 4.2 Result and Insights

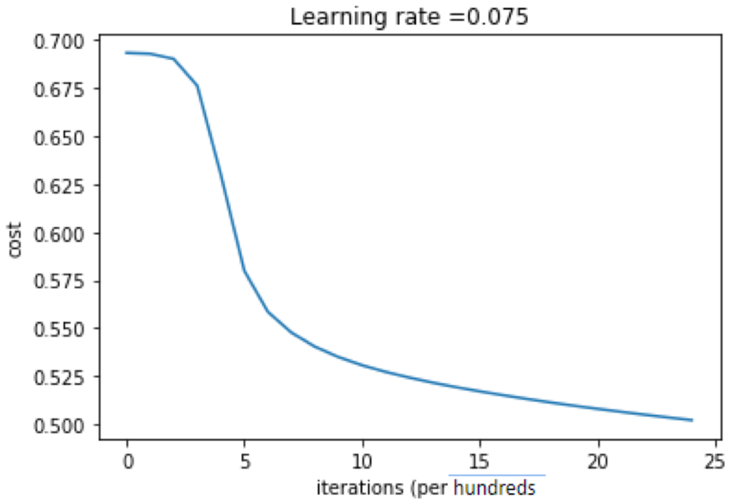


# 4.3 Neural Network

## 2 Layer Neural Network

Feature:'Brand','Author1','Browser','Category','Country' (2019 dummy variables)

Layer 1: 20 Nodes
Layer 2(output layer): 1 Node




```
predictions_train = predict(X_trainset, y_trainset, parameters)
```

Accuracy: 0.7356250000000002
AUC: 0.8219854679176607

```
predictions_test = predict(X_testset, y_testset, parameters)
```

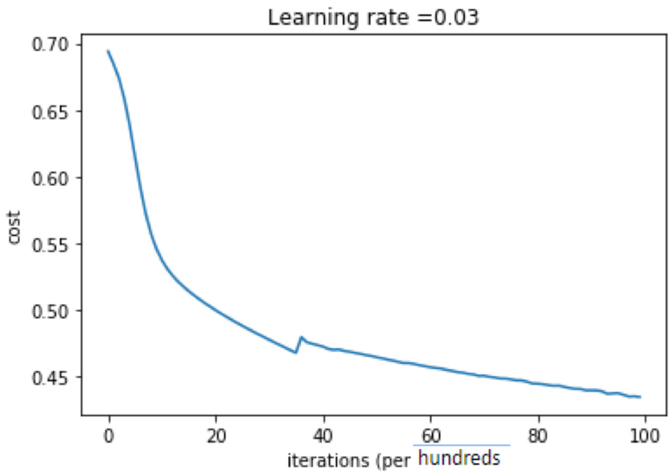
Accuracy: 0.7310000000000002
AUC: 0.81249895306611

# 4.3 Neural Network

## 4 Layer Neural Network

Feature:'Brand','Author1','Browser','Category','Country'
(2019 dummy variables)

Layer 1: 100 Nodes
Layer 2: 50 Node
Layer 3: 10 Node
Layer 4(output layer): 1 Node




```
pred_train = predict(X_trainset, y_trainset, parameters)
```

```
Accuracy: 0.7917500000000003
AUC: 0.8868533209687064
```

```
pred_test = predict(X_testset, y_testset, parameters)
```

```
Accuracy: 0.7415000000000003
AUC: 0.8265284922276878
```

# 4.4 Linear Regression p-value

Feature:'Brand','Author1','Browser','Category','Country' (2019 dummy variables)
Use p value to select significant variables. ($p < 0.05$)
Result: Out of 2019 variables, there are **310 significant variables**.

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Brand_allure** | 0.0047 | 0.021 | 0.229 | 0.819 | -0.036 | 0.045 |
| **Brand_architectural-digest** | 0.1373 | 0.030 | 4.652 | 0.000 | 0.079 | 0.195 |
| **Brand_ars-technica** | -0.0551 | 0.020 | -2.746 | 0.006 | -0.094 | -0.016 |
| **Brand_bon-appetit** | 0.0029 | 0.024 | 0.117 | 0.907 | -0.045 | 0.051 |
| **Brand_brides** | -0.0157 | 0.035 | -0.443 | 0.658 | -0.085 | 0.054 |

2019 variables
*Linear model*
R-square: 0.436

*Logistic model using same variables:*
Accuracy: 0.744
AUC: 0.8158

*RF model using same variables:*
Accuracy: 0.735
AUC: 0.8043

310 variables
*Linear model*
R-square: 0.676

*Logistic model using same variables:*
Accuracy: 0.723
AUC: 0.8129

*RF model using same variables:*
Accuracy: 0.723
AUC: 0.8095

Result: Model can maintain similar performance with fewer variables.

# 5 Recommendations

**Model performance:**

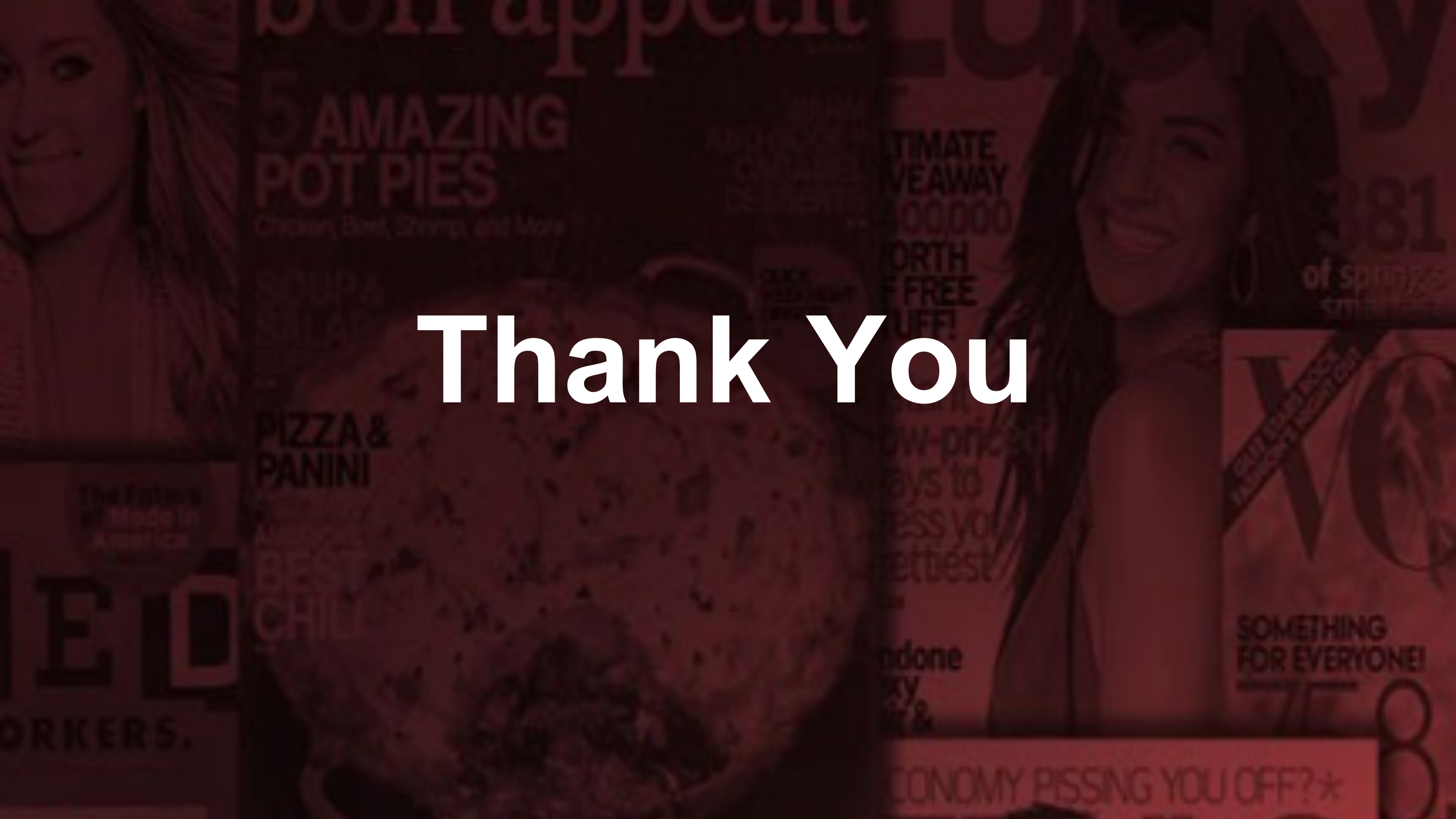
- Do variables pre-selection before building model
- Reduce dimension using PCA before building model

**Business strategies: :**

- Poland can be the next  market expansion opportunity after U.S. - it has 95.73% of the most valuable customers and is among the top 10  market sources.
- Invest more on new user acquisition programs and Search Engines Optimization as our analysis show internal and search referrals have  higher impact on propensity than referred externally and 60% of users are using Google Chrome(Google Search).
- Invest more on male writers, who are more favored by male customers (luxury_male_model)

Advertisement business:

- Luxury retail customers also show their interests in gourmet and travelling
- Luxury Jewelry customers also show their interests in beauty


AMAZING
POT PIES
Chicken, Beef, Shrimp, and More
PIZZA &
PANINI
SOMETHING
FOR EVERYONE!

# Thank You

# Appendix

Variable Pre-selection using H2O Random Forest

| | variable | relative_importance | scaled_importance | percentage |
|---|---|---|---|---|
| **0** | Channel | 59233.898438 | 1.000000 | 0.286599 |
| **1** | Category | 51064.945312 | 0.862090 | 0.247074 |
| **2** | Author1 | 33900.750000 | 0.572320 | 0.164027 |
| **3** | Brand | 18139.576172 | 0.306236 | 0.087767 |
| **4** | gender | 12643.190430 | 0.213445 | 0.061173 |
| **5** | Referer | 12106.101562 | 0.204378 | 0.058575 |
| **6** | clean_keyword | 7788.810547 | 0.131492 | 0.037686 |
| **7** | Browser | 5560.371582 | 0.093871 | 0.026903 |
| **8** | clean_title | 5226.294434 | 0.088231 | 0.025287 |
| **9** | Location | 1014.477905 | 0.017127 | 0.004908 |

Fig.1: variable importance

# Appendix

310 selected variables using Linear regression p value

'Brand_architectural-digest',
'Brand_ars-technica',
'Brand_golf-digest',
'Brand_gq',
'Brand_pitchfork',
'Brand_the-new-yorker',
'Brand_vogue',
'Browser_Apple Mail',
'Browser_BingPreview',
'Browser_Chrome',
'Browser_Chrome Mobile',
'Browser_Chromium',
'Browser_Edge',
'Browser_Facebook',
'Browser_Flipboard',
'Browser_HeadlessChrome',
'Browser_Instagram',
'Browser_Mobile Safari',
'Browser_Pinterest',
'Browser_Samsung Internet',
'Category_cmSubscribe',
'Category_general',
'Country_Bangladesh',
'Country_Canada',
‘Author ….