Gender Prediction

Analysis done by: Atieh Kermany

Executive Summary

Business challenge

Customer segmentation based on gender for better targeting and strategic planning

Opportunity

Current transactions data provides us with the opportunity to identify the gender probability based on purchased products and customers behaviour

Objective of this document

- To predict the customers gender based on other available features such as purchased items (female or male related)

Initial findings from analysis

- % of purchased items and the gender they are intended for can be leveraged to find a gender probability for each customers
- Only 4.2% of customers have similar % of female and male purchased items or have only purchased unisex items

Stages and findings

THE ICONIC

Stage 1

- Method: unzipped a database per customer transactions file and performed SQL queries to find followings:

Revenue by card	Customers % who purchased female items by card	Avg. revenue for IOS, Android or Desktop customers
$ 50,372,282	65.48 %	$ 1493.0

  - List of customers to target for an email campaign promoting a new men's luxury brand

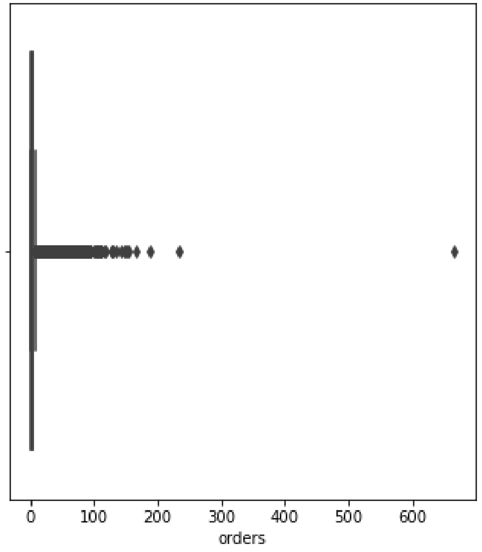
Stage 1

- List of customers to target for an email campaign promoting a new men's luxury brand
  - To do this task we first need to define the business KPI, duration and budget allocated for the campaign
    - KPI examples: brand lift (awareness)? conversion lift (revenue generation)?
  - RCT test required parameters based on KPI and for effective sample size calculation:
    - Expected conversion rate, uplift, confidence interval, and power
    - Based on above, required sample size can be calculated and then refined based on the allocated budget and test duration
    - By adding one or all the below conditions accordingly:
      - Adding avg spend per item to filter to those high spenders per item as this is a luxury brand
        - This can be done through a monetary segmentation as well and refined based on the new brand prices
      - Adding minimum number of orders
      - % of male items/items
      - Adding recency segmentation based on days since last transaction to exclude inactive customers

Required sample size	number of customers who had at least ordered one male item	number of customers with spend per item > $50, male items orders > 50%, orders > 3 and
1,769	17,106	1,909

Stage 2 – data cleaning

- Data: is in json format → converted to pandas dataframe (shape: 46,030)
- Row duplicates removal (new data shape: 46,279)
- Replacing 'is newsletter subscriber' flags from 'Y' and 'N' to 1 and 0
- Changing 'days since last order' from hours to days (/24)
- Missing values: only in 'coupon discount applied' column (count: 10,204)→ replaced by median (0)
- Removing when 'revenue'=0 if none of 'cancels', 'returns', 'vouchers', and 'discounts' are 0 (shape: 45,144)
- Removing rows when no payment method recorded (new data shape: 45,059)
- Checking that the total items match the female, male and unisex items, and similarly the orders match their respective breakdown (device type and delivery point)
- I noticed that the female and male items don't match the sub-category counts
  - More info on data collection is required to further investigate this problem
- The distribution analysis shows few customers as outlier but not necessarily as error

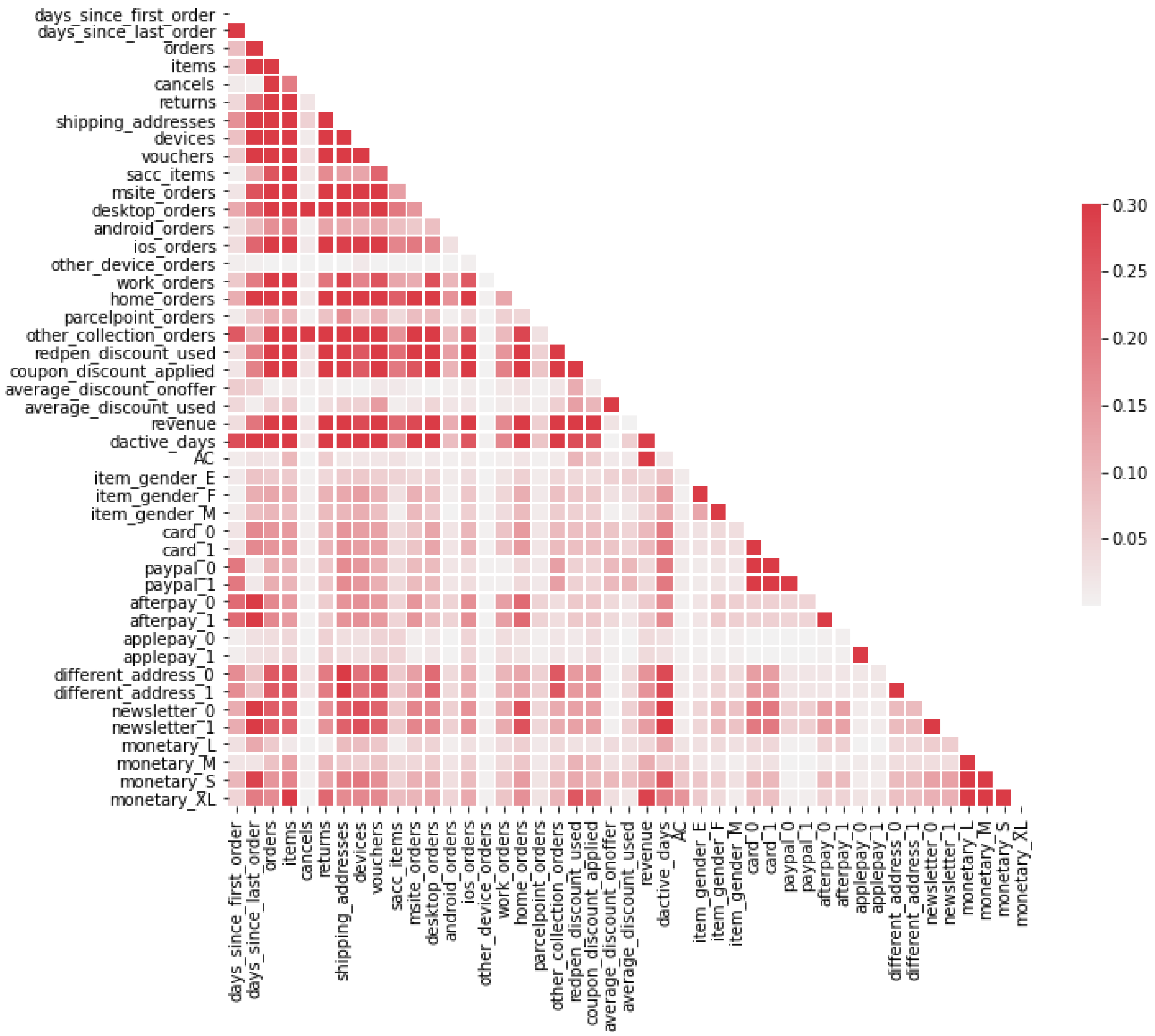


Stage 3 – Feature Engineering

- Creating female, male, and unisex item purchase rates from ‘female items’, ‘male items’ and ‘unisex items’
  - For those with unisex purchases larger than female and male purchases:
    - Creating female and male purchase rates from sub-category items and as portion of the ‘unisex item rate’
    - Summing the two rates to create a final female and male item purchase rates
  - Normalization: ensuring the summation of the rates (female and male) will be 1
    - to avoid divided by 0 error, a filter is applied on when (female + male items rate = 0)
- Creating a categorical item gender flag based on items rate (‘item gender’)
  - 'M' if male items rate > female items rate, 'F' if female items rate > male items rate, and 'U' if the two are equal
  - % of customers who have equal female and male purchased items rate: 4.2%
- Creating number of active days feature from ‘days since first order’ and ‘days since last order’
- Creating average cheque (AC) feature and monetary segment
  - AC = from revenue/(orders – cancels)
    - AC = 0 when (orders – cancels = 0)
  - Monetary segment is defined based on AC percentile into 4 groups of ‘S’, ‘M’, ‘L’, and ‘XL’
- One hot encoding on some of the categorical/binary columns: ‘item gender’, ‘cc payment’, ‘paypal payment’, ‘afterpay payment’, ‘apple payment’, ‘different addresses’, ‘is newsletter subscriber’, and ‘monetary segment’
- More segmentation and one hot coded features can also be created and investigated

Stage 3 – Correlation and Feature Importance Analysis

- Important features picked from the correlation analysis:
  - ‘item gender M’
  - ‘item gender F’
  - ‘item gender E’
  - ‘active days’
  - ‘devices



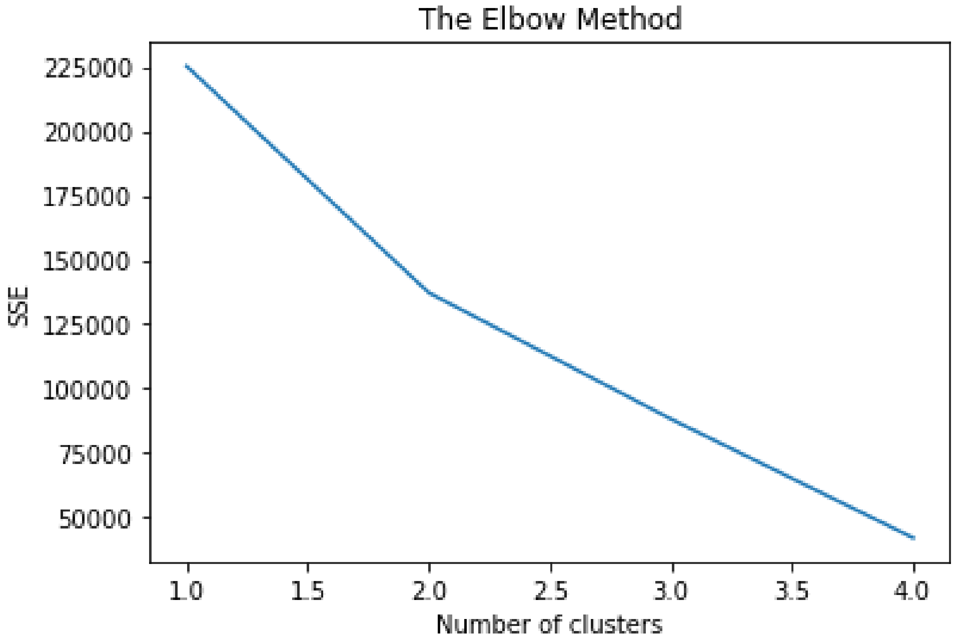
Next steps:

- Features into a numpy array
- Standardisation

Stage 3 – k-means modelling

- Applying k-means algorithm to predict gender through clustering (2 clusters: female and male)
  - The features are refined for elbow method to show '2' as the number of clusters (male and female here) with minimum error
- Saving the predicted values as a new column: 'male' (1 if male and 0 if female)
- Comparing the 'male column' with the 'item gender M', and 'item gender F' columns to measure accuracy
  - Females are predicted correctly, males and all unknows (4.2% of total population) are predicted as male
  - The 'male' column results can be compared with the labeled data (if known) to measure the exact accuracy

~Actual females	Predicted females	~Actual males	Predicted males	~Actual unknown
31,495	31,495	11,662	13,564	1,902



Thank You