SEM Breakeven Bid Case Study

Team HELLO DATA

Ziang (Allen) Lu
Jiajia (Deborah) Yin
Yi Zhang
Yilun (Tom) Zhang

Overview

I. **Background**

II. **Problem Formalization**

III. **Statistical Methods**

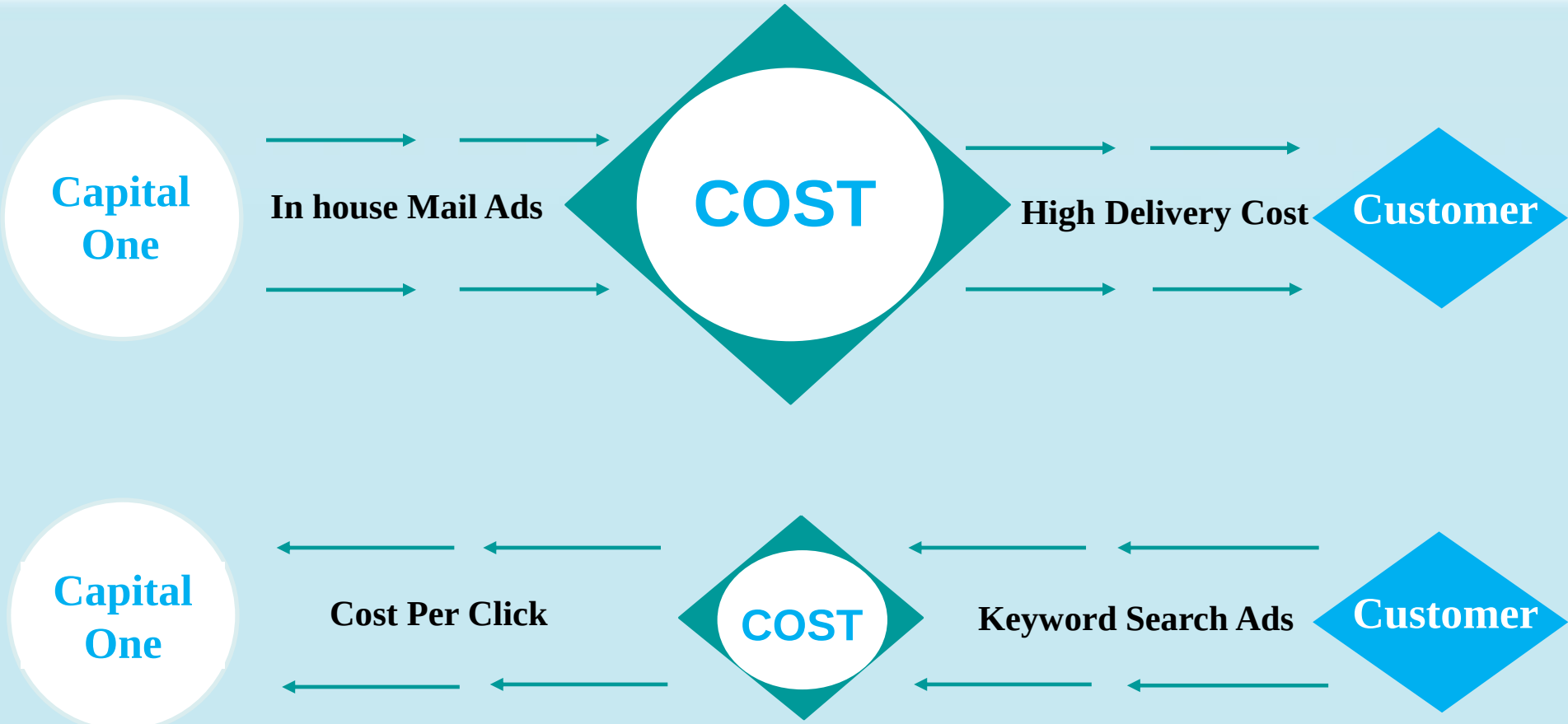
i. Data Preprocessing and Transformation

ii. CR, AR, PR Modelling

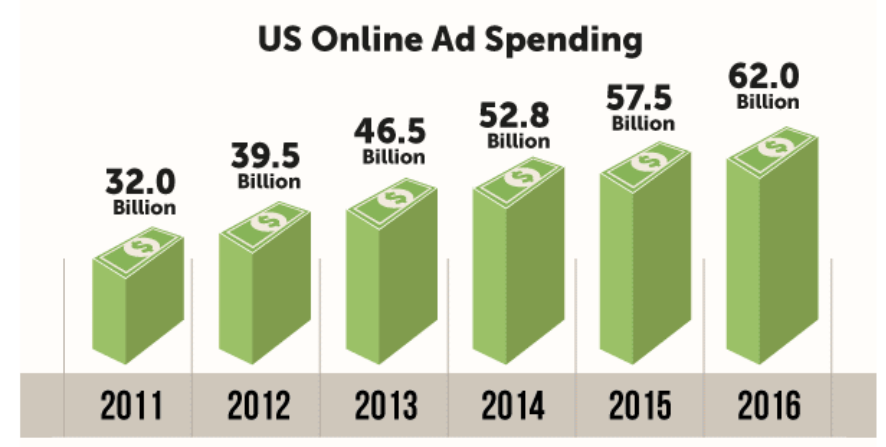
iii. Validation Set Prediction

IV. **Business Insights**

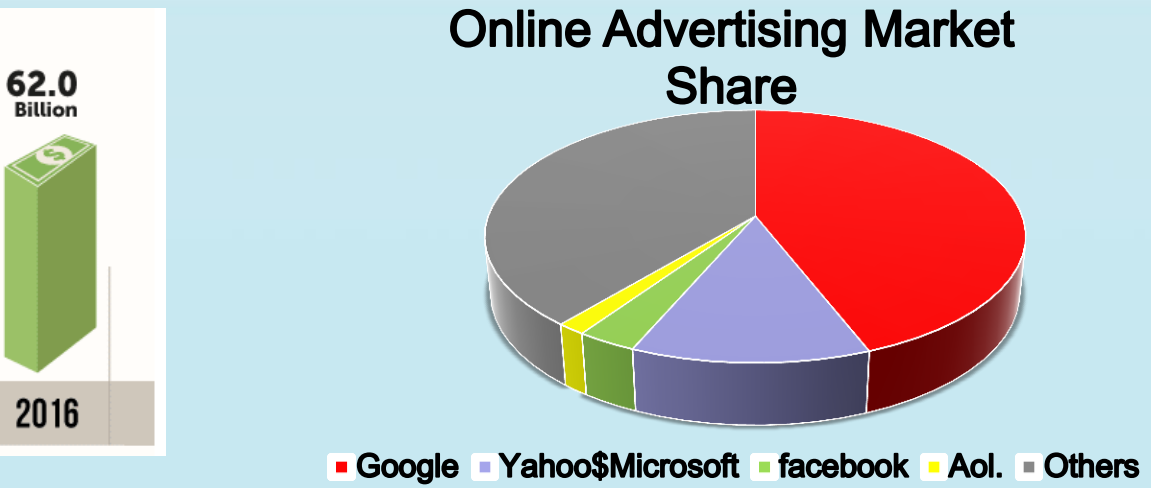
Background -- Capital One



Background – Online Ad Industry Overview

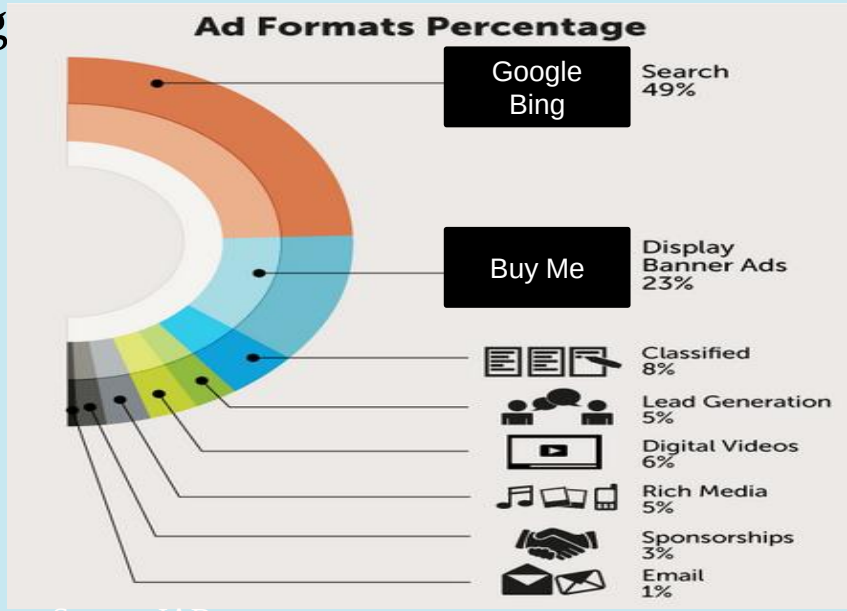


Source: IAB



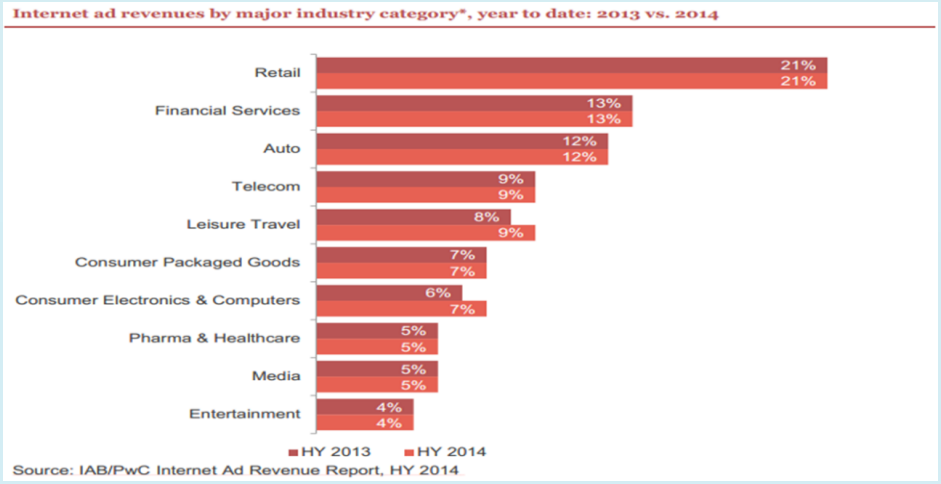
Source: IAB

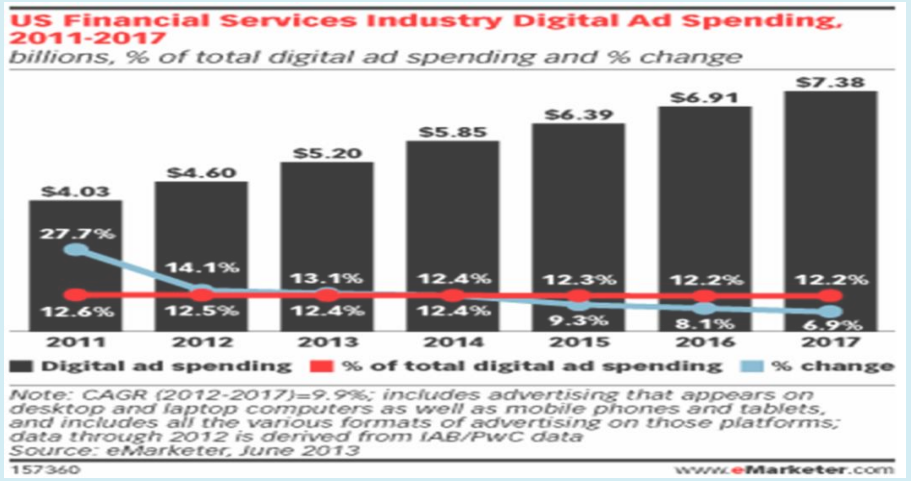
- Increased growth rate at online ads. spending each year
- Keyword search occupies half of the market shares
- Cost-efficient investments to boost business



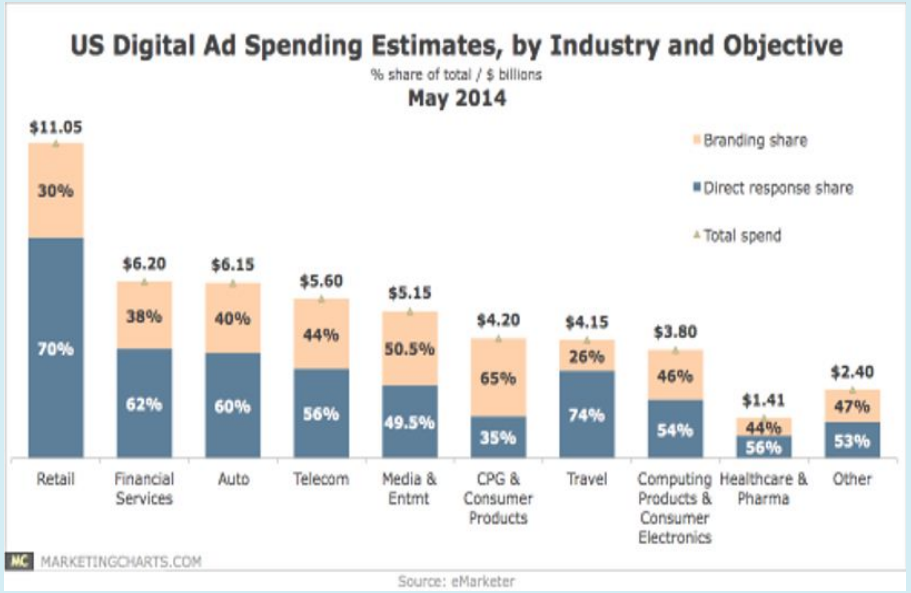
Source: IAB

Background -- Finance Industry Outlook

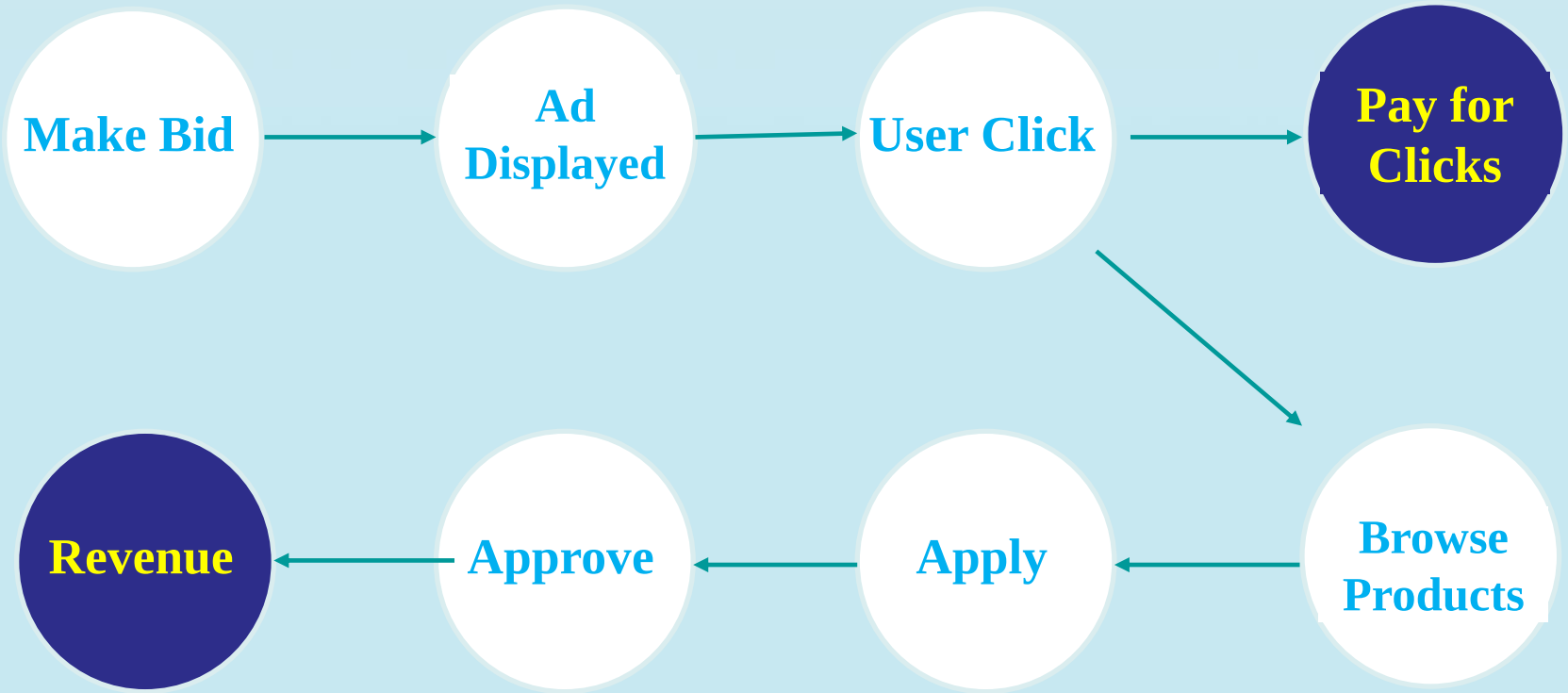




- Second largest industry spent in online ads market
- Spending amount grows dramatically each year
- Proper bidding strategy is essential in attracting more customers than peers



Problem Formalization



Problem Formalization – Formulas

Total Revenue → **Equality** ← *Total Cost*

Product Revenue
× Approved Number of Applications

Cost Per Click
× Number of Clicks

$$Breakeven\ Bid = Product\ Revenue \times Conversion\ Rate \times Approval\ Rate$$

Problem Formalization

For each bid

- What's its expected Conversion Rate?
- What's its expected Approval Rate?
- What's its expected Revenue generated?

Ultimate

What is the maximum amount we are willing to pay for each click to breakeven (i.e. max bid)?

Statistical Methods – Preprocessing & Transformation

- Selected significant features:
  - ENGN_ID, LANG_ID, DVIC_ID, KEYWD_TXT, MTCH_TYPE_ID, etc.
- **Separated keyword combinations to single keywords**
- Grouped all information, using the combination of above features and *SINGLE* keywords as the primary key
- Calculated Conversion Rate for existing data
- Substituted NaN/NA with appropriate value (mostly 0)

	ENGN_ID	LANG_ID	DVIC_ID	CMPGN_NM	KEYWD_TXT	MTCH_TYPE_ID	IMPRESSIONS	CONDITIONAL_IMPRESSIONS	TOTAL_QUALITY_SCORE	IMPRESSION_TOTAL_RANK	VISITS	CONV_RATE
1	G	E	D	CMPGN9	KW1	B	4.025000	4.025000	16.895000	5.710500	0.135000000	0.000000000
2	G	E	D	CMPGN6	KW10	E	3.176471	3.176471	3.176471	14.364706	0.000000000	0.000000000
3	G	E	D	CMPGN9	KW100	B	3.109677	3.109677	7.225806	10.270323	0.006451613	0.000000000
4	G	E	D	CMPGN1	KW101	B	6.500000	6.500000	19.500000	21.600000	0.000000000	0.000000000
5	G	E	D	CMPGN2	KW102	B	12.422535	12.422535	99.711268	16.935211	0.739436620	0.009523810
6	G	E	D	CMPGN2	KW102	E	1.000000	1.000000	9.000000	1.000000	0.000000000	0.000000000

Statistical Methods -- Conversion Rate Modelling

- Use classification and regression tree (CART) method to model Conversion Rates for biddings containing a *SINGLE* keyword
- Predict Conversion Rates of biddings with multiple keywords by weighing each *SINGLE* keyword based on their relationships with the bidding's corresponding CMPGN_NM (Ad_ID)

(See Appendix A&B)

- E.g. $p = proportion\ of\ keywords\ that\ are\ frequent\ in\ CMPGN$

$$CONV\_RATE(KEYWD\_TXT) = \sum_{\{KW_i \in CMPGN_j\}} CONV\_RATE(KW_i) \times \sqrt{p} + \sum_{\{KW_i \notin CMPGN_j\}} CONV\_RATE(KW_i) \times (1 - \sqrt{p})$$

Statistical Methods -- Conversion Rate Modelling & Evaluation

Model Name	Root Mean Squared Error
CART	0.0594
Decision Table	0.0729
KStar	0.0784
Decision Stump	0.0785
M5Rules	0.0808
Linear Regression	0.0808
ZeroR	0.0811
Regression By Discretization	0.0811
Multi Scheme	0.0811
REP Tree	0.0812
M5P	0.0828
IBk	0.1051

- CART is a decision tree based method that calculates the average of known data according to some of their common features to make predictions by placing new data into the appropriate branch of the decision tree
- Comparing to other models run in Weka, CART has the best root mean squared error (from 10-fold-cross-validation)

* For Weka Software, see Appendix C

Statistical Methods -- Approval Rate Modelling

- $AR(Prod\ i) =$
  $Approved\ Prod\ i\ Applications\ /\ Total\ Prod\ i\ Applications$
- For each keyword, extract the probability distribution over the 6 products it potentially leads to
- Obtain the Approval Rate corresponding to each keyword as the expected AR over the products
- $AR(KW) = \sum_i \Pr(Prod\ i \mid KW) \cdot AR(Prod\ i)$
- Approval Rate of bids with multiple keywords are averaged over its single keywords

Statistical Methods -- Product Revenue Modelling

- $PR(Prod\ i) =$
  $Prod\ i\ Total\ Revenue\ /\ Prod\ i\ Approved\ Applications$
- For each keyword, extract the probability distribution over the 6 products it potentially leads to
- Obtain the Product Revenue corresponding to each keyword as the expected PR over the products
- $PR(KW) = \sum_i \Pr(Prod\ i \mid KW) \cdot PR(Prod\ i)$
- Again, Product Revenue of bids with multiple keywords are averaged over its single keywords

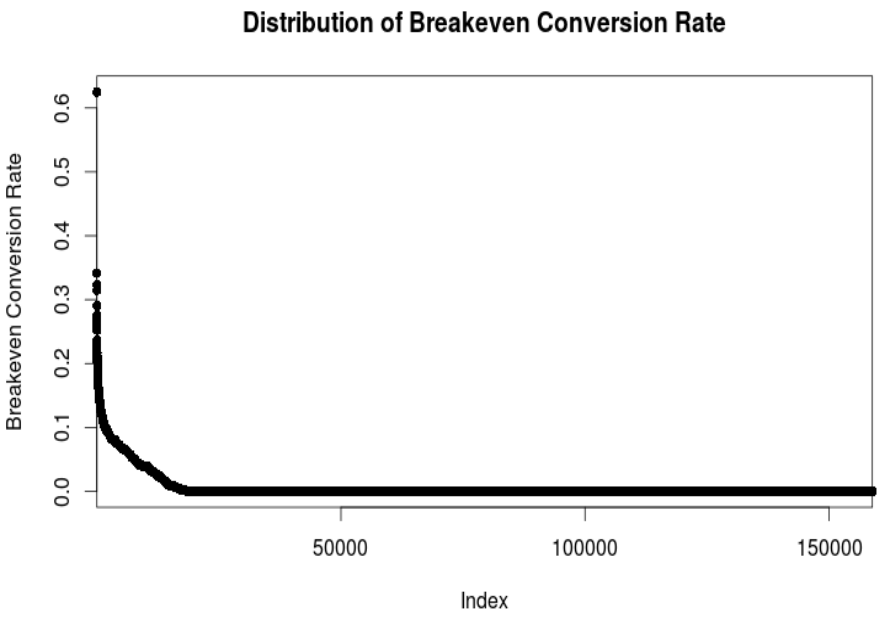
Statistical Methods -- Validation Set Prediction

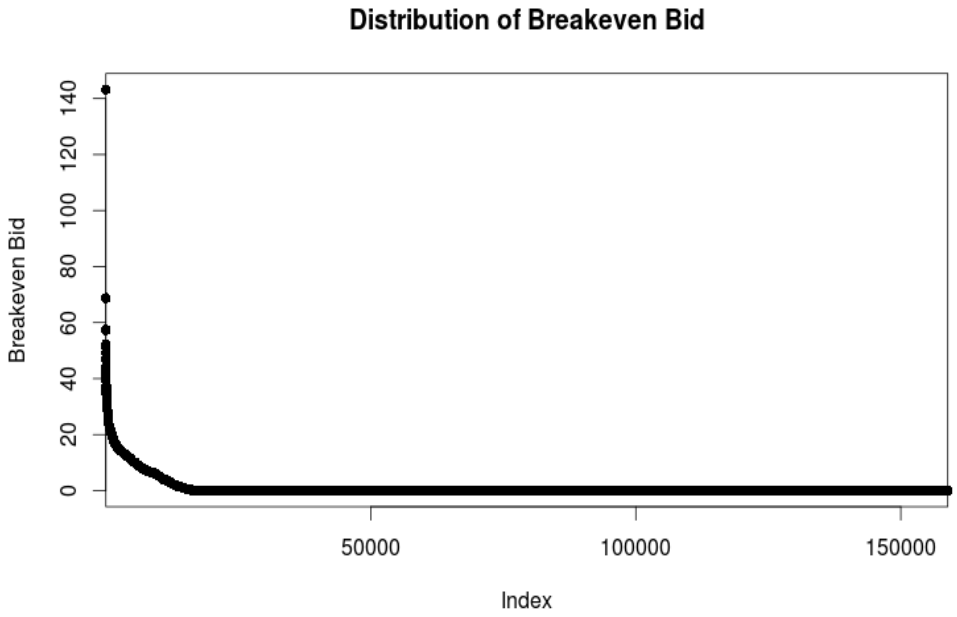
- Obtained Conversion Rate, Approval Rate and Product Revenue for all combinations of SINGLE keyword and selected features
- Estimated the above three values for each bidding in the validation set
- Thus, determined the breakeven bid (max bid) for each bidding

- Note that there are **22** new *SINGLE* keywords and **286** new keyword combinations in the validation set, thus using the *SINGLE* keywords in modelling reduces the amount of unknown information

  - Performance of prediction will be improved when more data containing these keywords are gathered

Statistical Methods -- Result Visualization



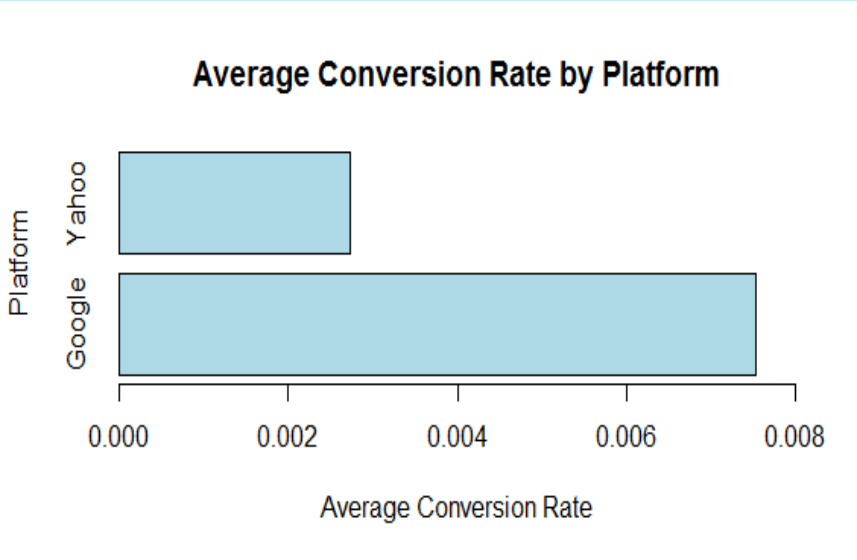


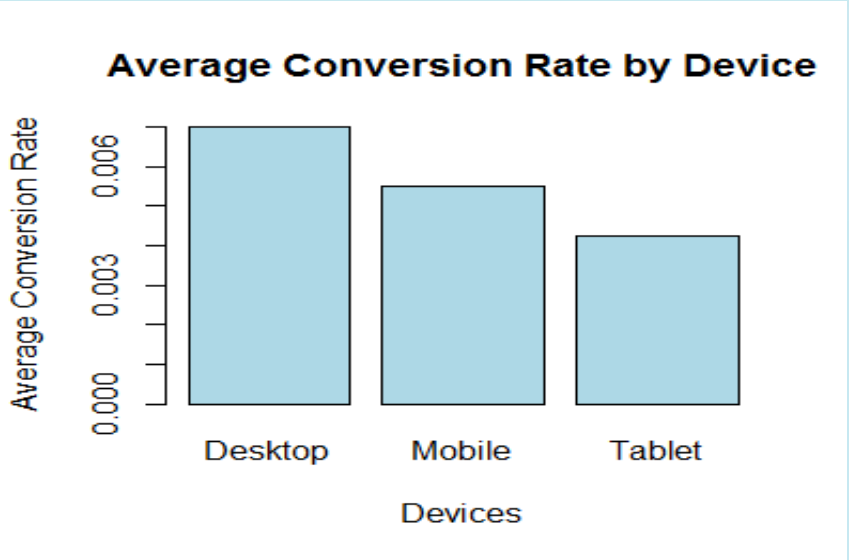
- The Conversion Rate and the Breakeven Bid have similar distributions
- 137466 (86.53%) zeros in predicted Conversion Rate
- 138143 (86.95%) zeros in predicted Breakeven Bid

Statistical Methods -- Discussion

- Due the variety in KEYWD_TXT, some models including linear regression and logistic regression are either very ineffective or unable to run in reasonable time
- Due to the nature of this problem which results in many conversion rates being zero, our model tends to be positively biased and overestimate the Conversion Rate when it is actually low (i.e. 0)
- In the future, meta algorithms such as bagging or boosting can be used to improve the prediction accuracy
- Programmatic feature selections (PCA, forward/ backward/ stepwise selection) can be conducted to reduce the dimensionality of the problem as well as potentially improve the accuracy

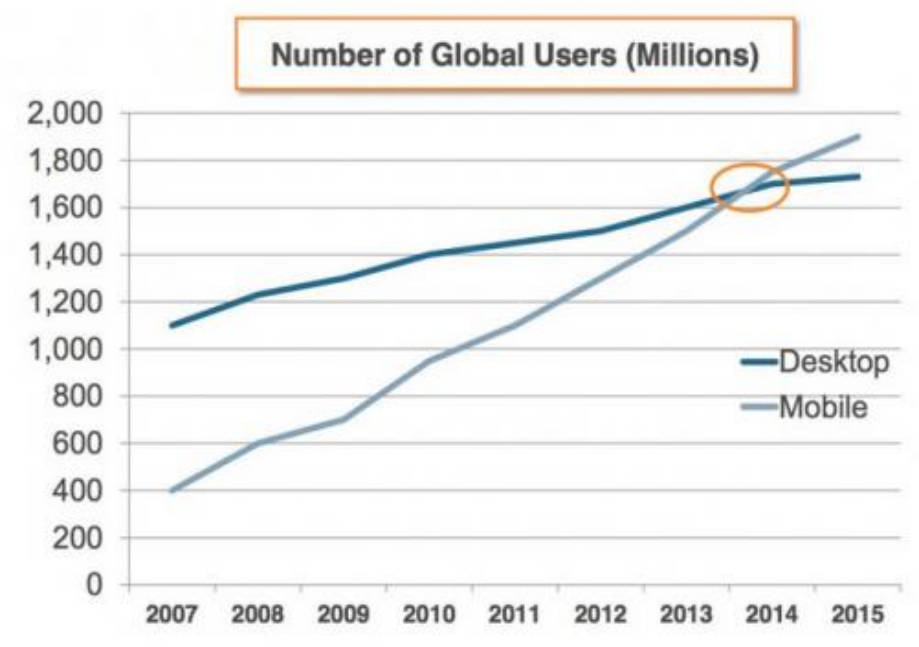
Business Insights





- Google is a more profitable search platform to investment in comparing to Yahoo!
- Desktop is still a significant leader in average number of product applications per user click among digital devices
- To boost up Conversion Rates of other devices, we could potentially improve the user experiences on them or introduce promotions for these devices

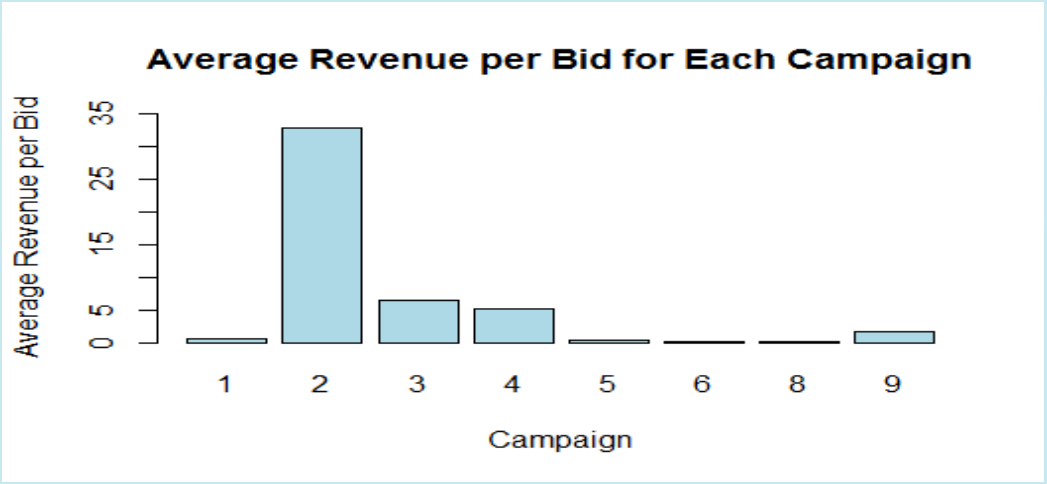
Business Insights

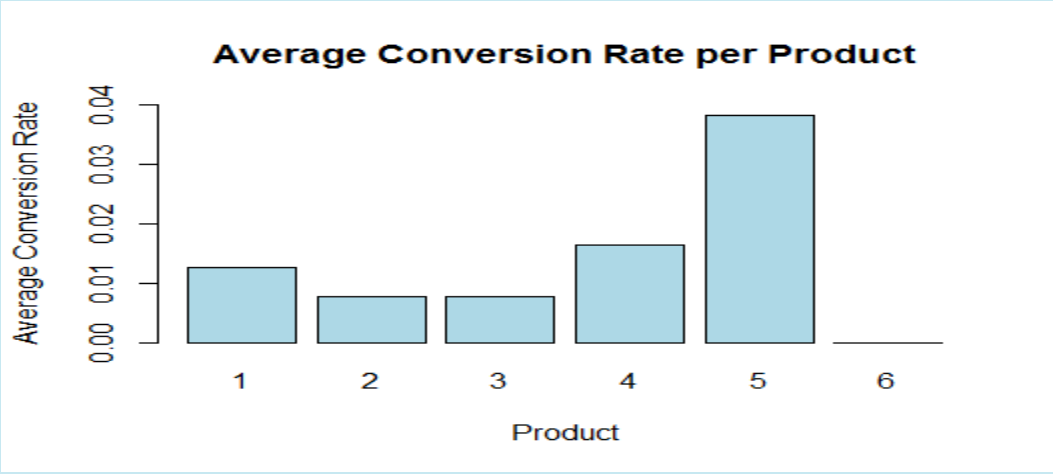


- The trend in mobile usage beats the trend in desktop starting from 2014
- Mobile searching will play a significant role in future search engine market

Business Insights

- Campaign 2 has been the most successful in terms of revenue generated per bid we place, where future lesson could be learned
- Product 5 has been most popular product in terms of its average conversion rate, and other ones need to catch up
- Most frequent key words from the most successful Campaigns 2 are: KW104, KW382





Key Findings Summary

- Determining the maximum bidding strategy is absolutely crucial
- From the data set, we made separate quantitative models
  - CART for the Conversion Rates
  - Weighted Average for Approval Rates and
  - Weighted Average for Product Revenue
  - to calculate the maximum bid
- We made deeper observations on the data for future directions

Q&A

THANK YOU

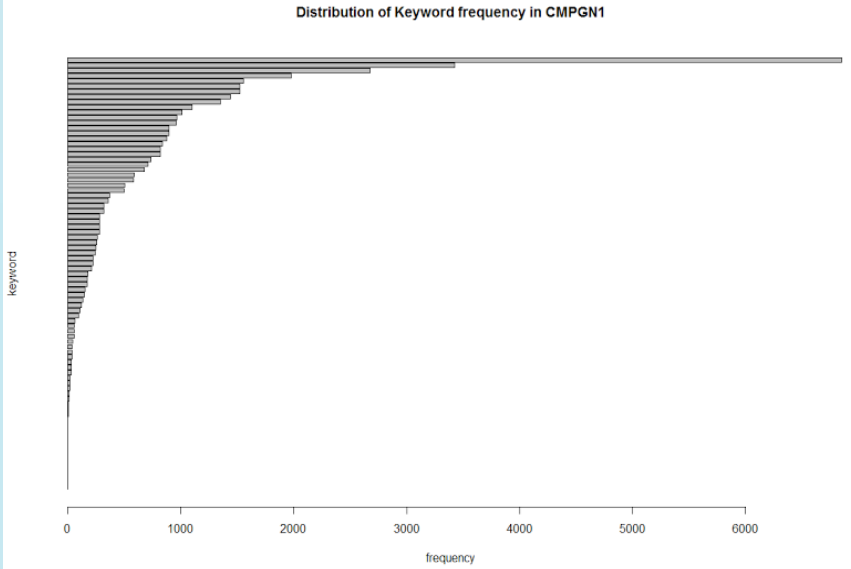
Appendix A

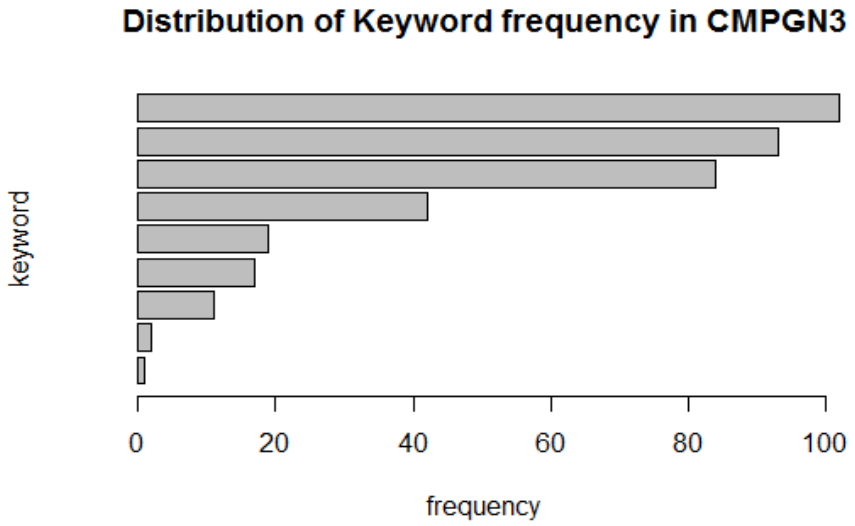
-- Table for Keyword Frequency in Campaign (Ads Group)

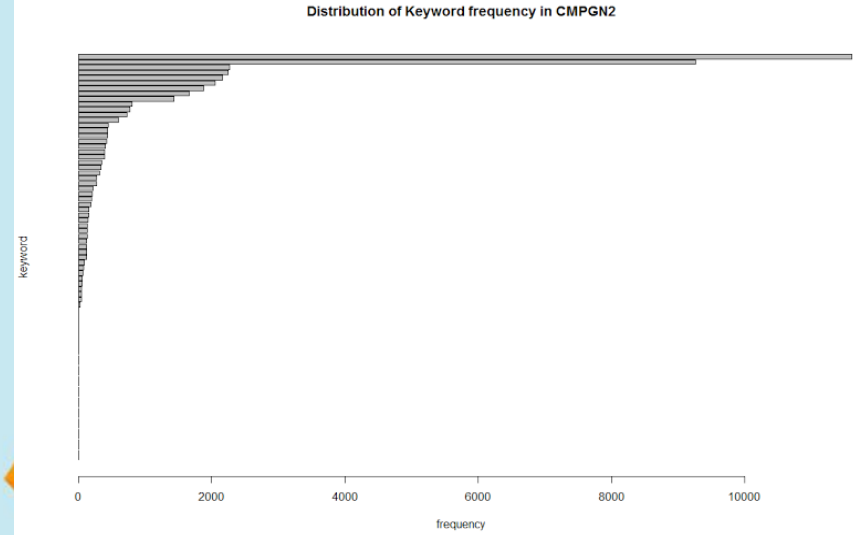
CMGPN1		CMPGN2		CMPGN3		CMPGN4		CMPGN5		CMPGN6		CMPGN8		CMPGN9	
KW	Freq	KW	Freq	KW	Freq	KW	Freq	KW	Freq	KW	Freq	KW	Freq	KW	Freq
KW178	6852	KW104	11601	KW104	102	KW178	9183	KW178	5142	KW97	8697	KW195	1200	KW178	19206
KW121	3424	KW382	9262	KW382	93	KW426	4999	KW587	4029	KW121	7158	KW25	607	KW121	10865
KW587	2677	KW178	2266	KW124	84	KW121	4471	KW121	3856	KW178	7138	KW121	495	KW587	9105
KW306	1977	KW97	2242	KW178	42	KW587	4284	KW340	3333	KW92	6877	KW178	487	KW340	6795
KW97	1556	KW92	2159	KW340	19	KW97	2376	KW563	1709	KW288	6379	KW97	233	KW97	4867
KW324	1523	KW121	2053	KW196	17	KW62	2267	KW125	1585	KW340	4104	KW515	229	KW124	4544
KW560	1522	KW340	1881	KW92	11	KW340	1765	KW53	1510	KW456	3476	KW340	209	KW567	4228

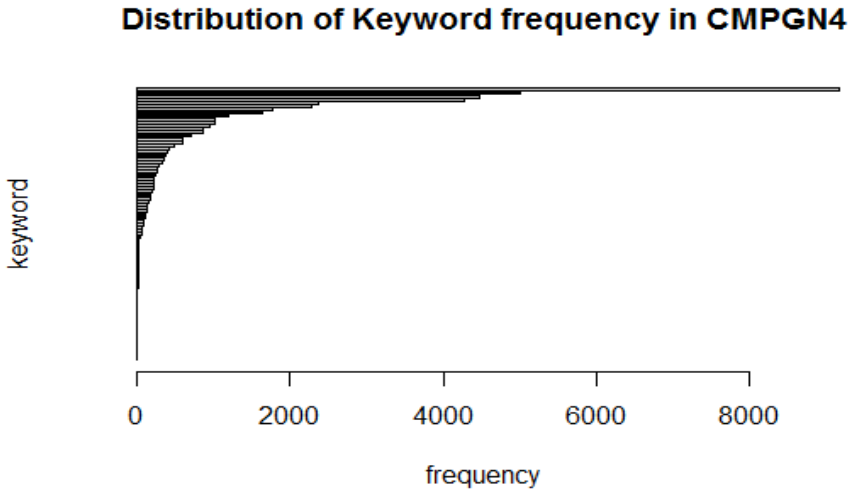
Appendix B

-- Plots for Keyword Frequency in Campaign (Ads Group)



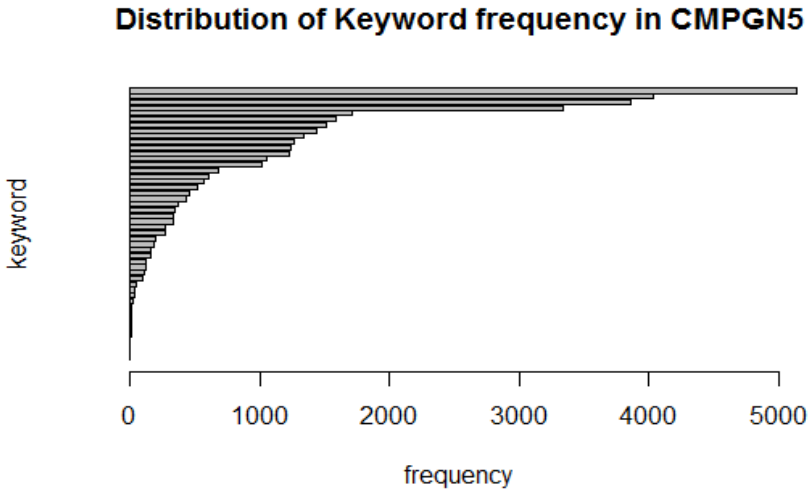


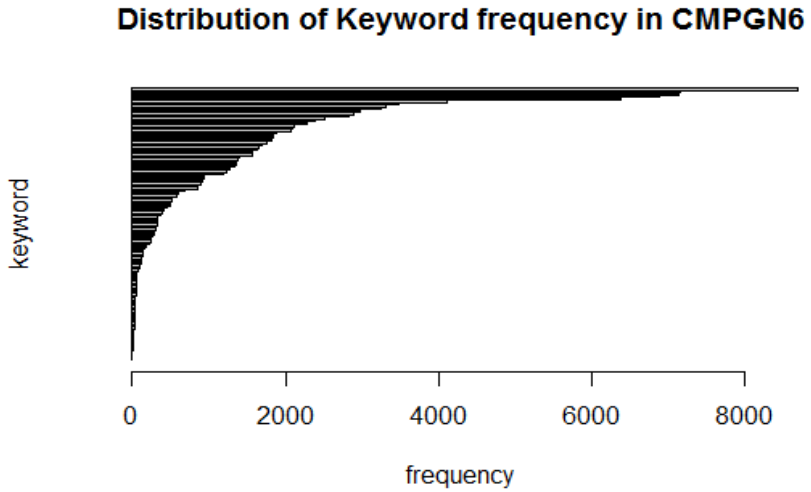


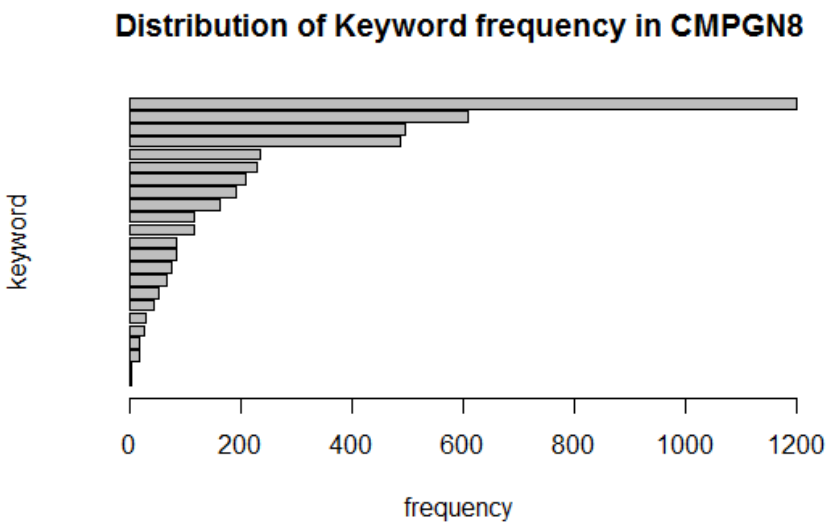


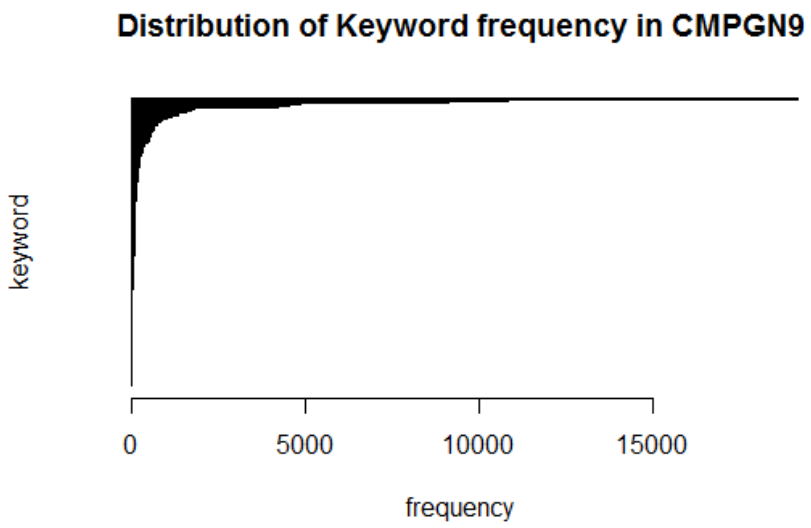
Appendix B

-- Plots for Keyword Frequency in Campaign (Ads Group)









Appendix C
-- Weka Software

- Please see attached Word document for Weka Software instruction