Leveraging an Erroneous Treatment

Did We Wake Sleeping Dogs, Reactivate Engagement, or Do Nothing At All?

Predictive Analytics World
San Francisco
April 4th, 2016

Ming Ng, LinkedIn
Jim Porzak, DS4CI.org

V1.1 is as presented but with typo's corrected.

Outline

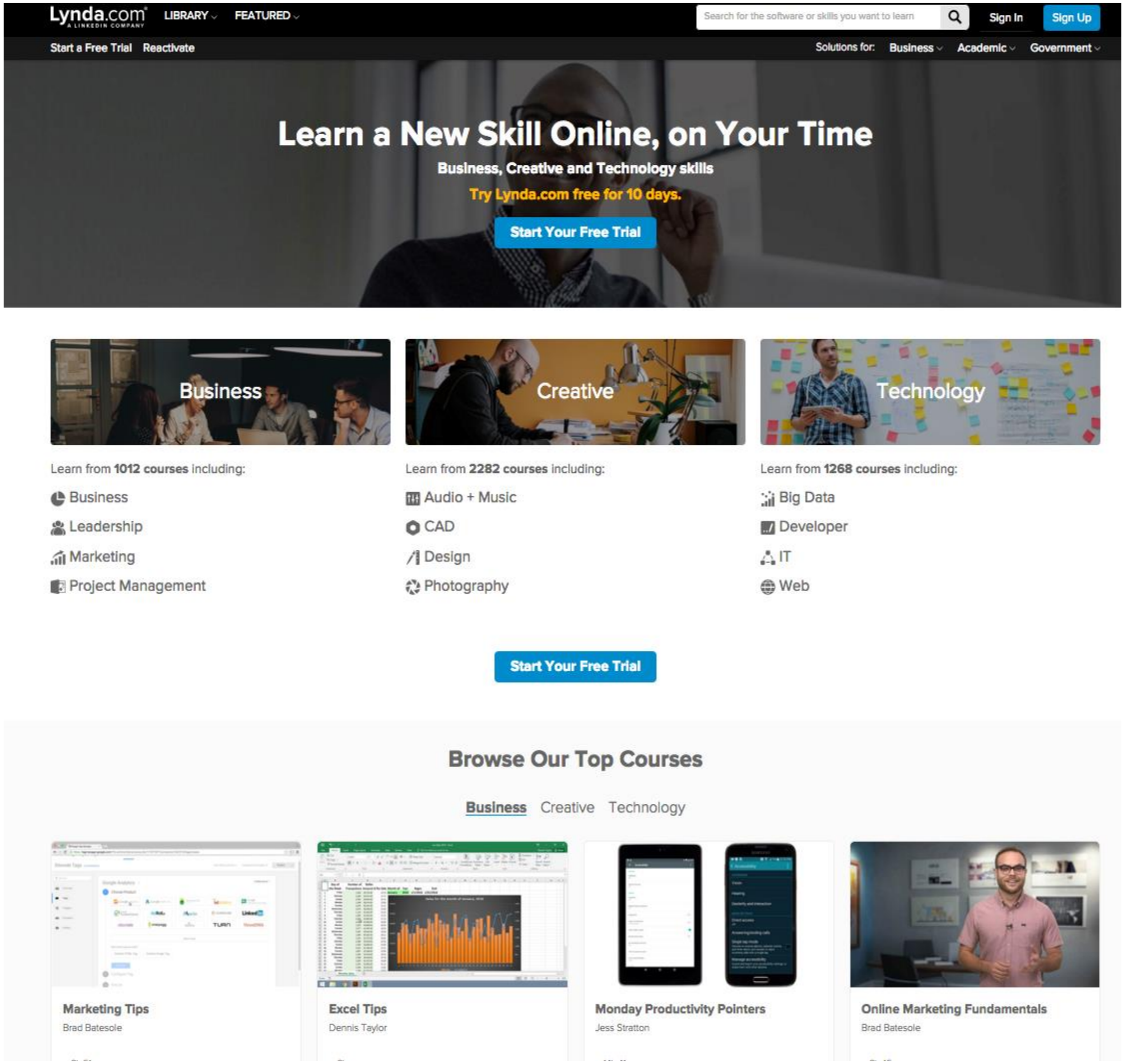
1. Uplift modeling background, we skip this!
   - See talks earlier today and Eric Siegel's book – Chapter 7
2. Lynda.com – What they do. What happened.
3. Analysis – Data, Engagement, & Uplift
4. Conclusions

Appendix has references, tech deep dives, and links to learn more.

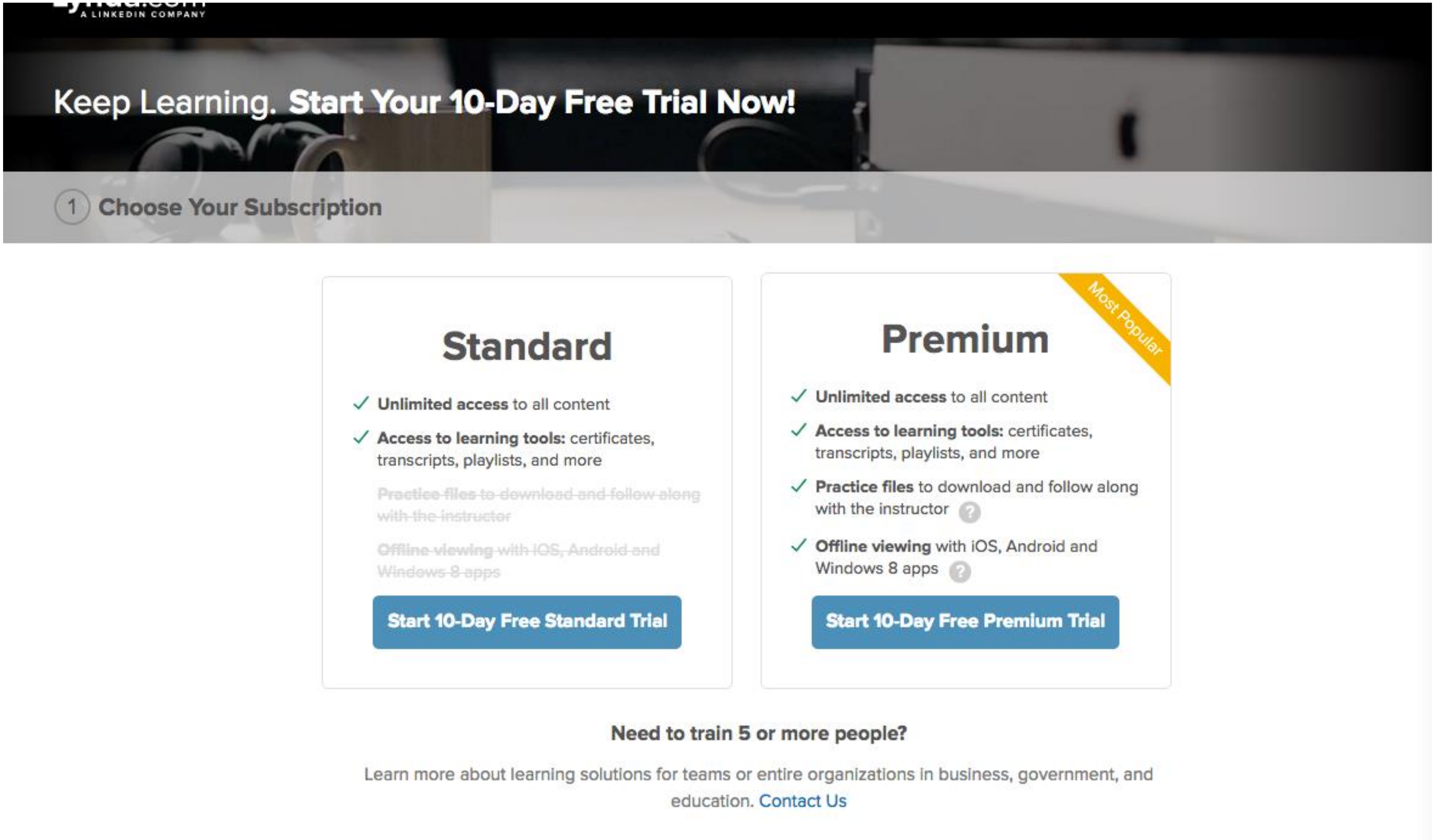
Lynda.com

1. Lynda founded in 1995, bought by Linkedin in 2015.
2. A subscription based e-learning company.
3. Access to 8000+ online courses.
4. Members start with a free trial.
5. At end of the trial, customers can cancel or be billed $24.99 or $34.99 monthly for a subscription.
6. Members are automatically billed unless they cancel.
7. Cancelled members can reactivated their account.

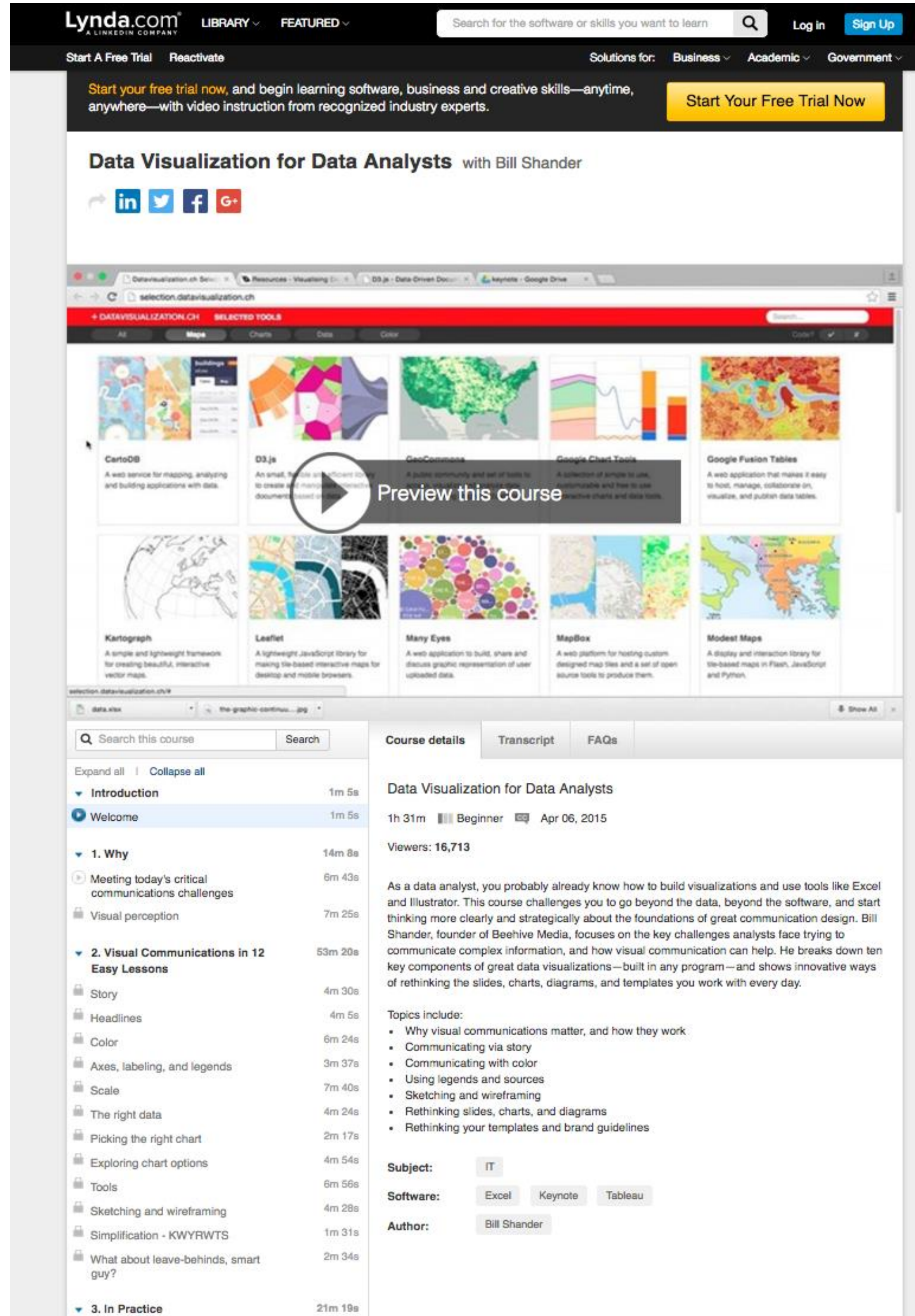
Lynda.com Hompage



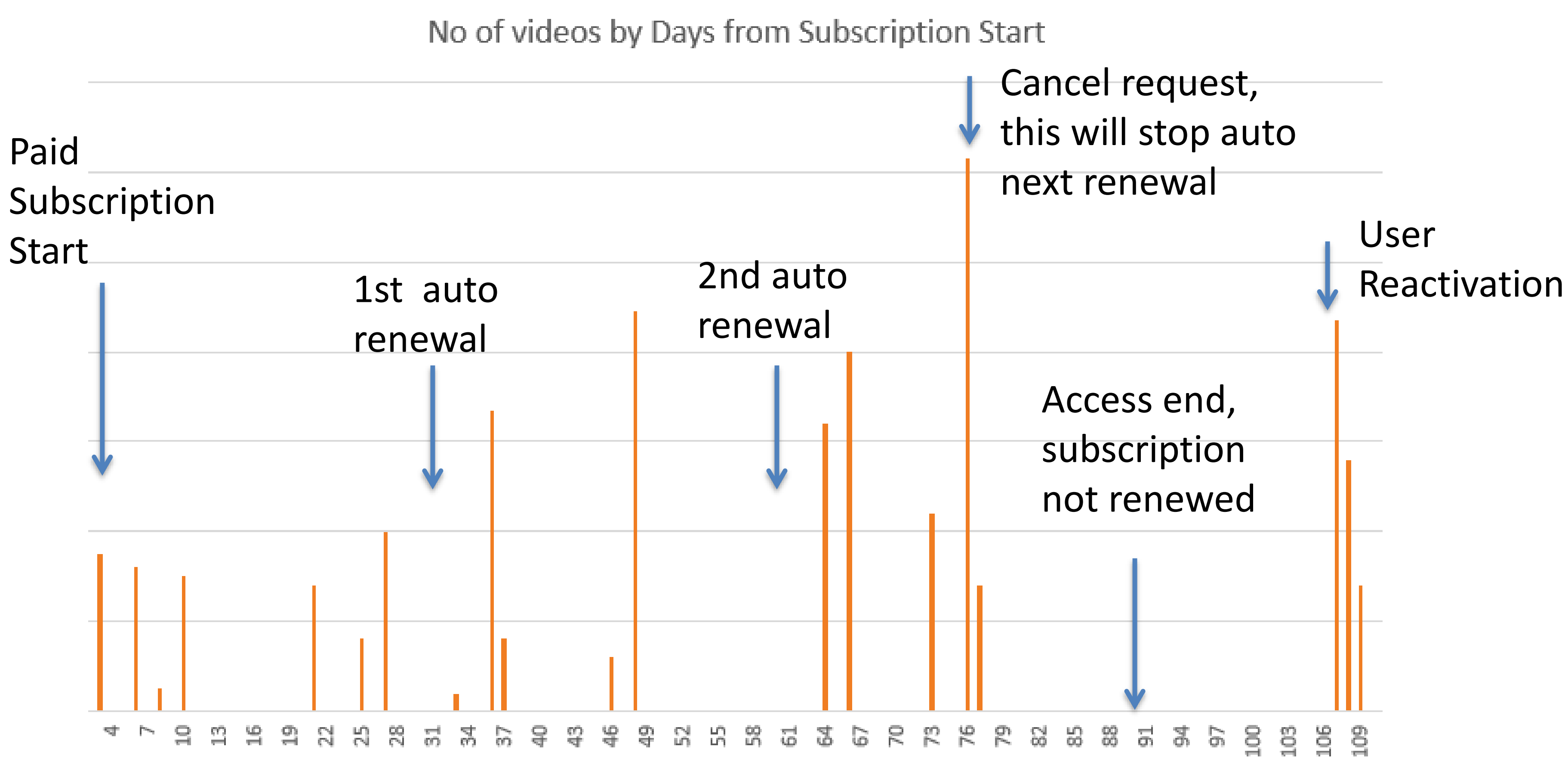
Trial Signup Page



Videos on Lynda.com

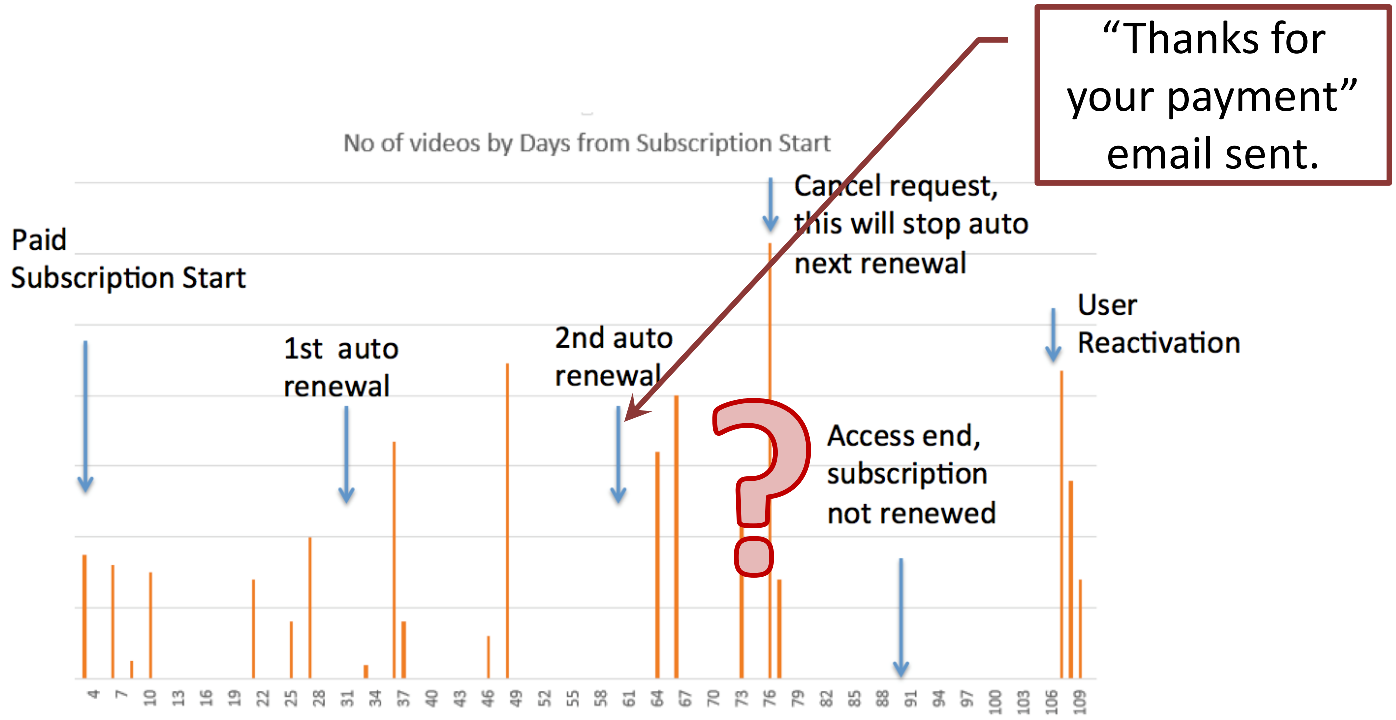


Typical Customer Behavior



Site Error

- When user signs up, they can opt-in to receive a payment receipt via email. The default is to opt-out. Very few opt-in.
- A site error caused users that are opted-out to get an email receipt thanking them for their payment. This error lasted for 8 days in mid August 2015.



The erroneous "Thanks" Email

lynda.com

Dear jim,

Thanks for your membership payment to lynda.com. Your transaction details are below

Payment Information

Payment date	08/27/2015
Order amount	$37.50
Payment method	CreditCard
Card info	
Name on card	

Order Information

Order #	A-S00328225
Order date	08/27/2015
Subscription amount	$37.50
Subscription type	Premium Monthly Fee
Subscription start date	08/27/2014

Billing Information

Full Name	
Email	
Phone	
Address	
City/State/Zip	
Country	United States

CONTACT US

Add info@lynda.com to your address book to ensure delivery.
This message was mailed to jporzak@gmail.com from lynda.com.

VIEW THIS EMAIL ONLINE | MANAGE PREFERENCES

In the Gmail inbox, it looks like this:

lynda.com Inbox Thanks for your payment - Thanks for your payment Dear jim, Thanks for y Aug 27

Which raises the business question:

What did the erroneous “Thanks” email do?

- *Did We Wake Sleeping Dogs?*
  - *Cause a cancel which would not have happened.*
- *Reactivate Engagement?*
  - *Cause an increase in video viewing.*
- *Do Nothing At All?*

The Analysis Workflow

1. Build our “Sleeping Dogs” data set
2. Exploratory data analysis & data profiling
3. Did erroneous “Thanks” email increase engagement?
    1. Look for increase in user video viewing activity.
4. Did they cause cancels?
    1. Use information value (IV), weight of evidence (WOE), and variable clustering to screen & select predictors.
    2. Build the uplift model.
5. Evaluate models & report results.

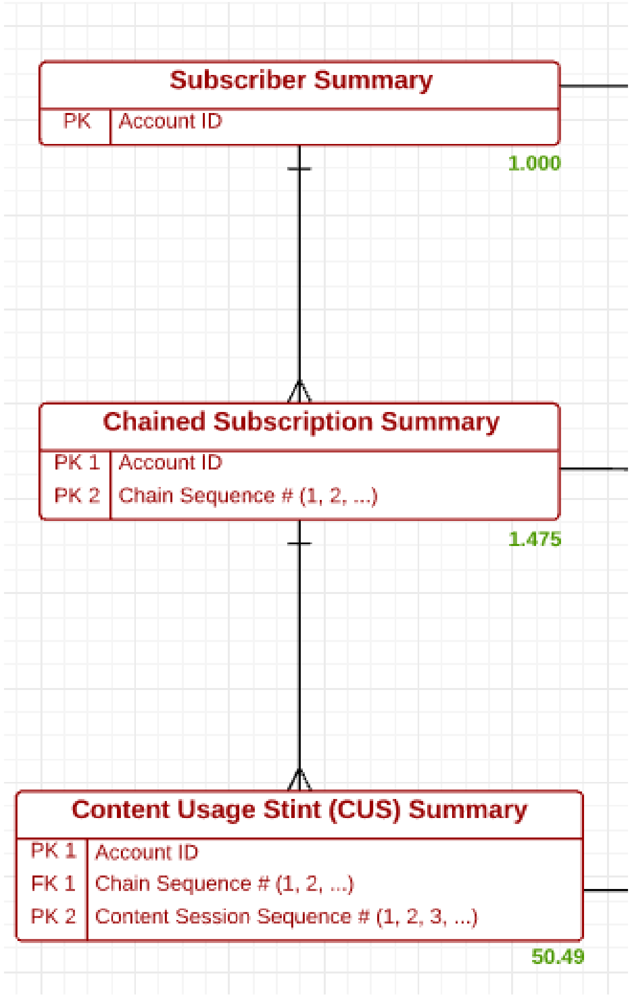
Building Data Sets – 1 of 4

We need two data sets for the two questions:

1. Did erroneous “Thanks” email increase engagement?
   - For non-cancelers, get video engagement metrics for 30 days before and 30 days after email.
2. Can we find an uplift churn effect?
   - Data about subscriber, subscription, video metrics, and if they canceled.
   - Include metrics up to time of email only – *no future looking metrics!*
   - Do 70/30 split into training and validation sub-sets

Building Data Sets – 2 of 4

Secret weapon: Redshift subscriber data mart.



Design:

- Models subscriber behavior
- Wide tables – easy to understand
- High data quality
- Very fast to access

Three levels of abstraction:

1. “Everything” about our subscribers.
2. “Everything” about their <u>subscription chain(s)</u>.
3. “Everything” about their <u>content usage stints</u>.

To get details, just Google:
“site:ds4ci.org structuring data”

Building Data Sets – 3 of 4

For engagement question:

- Select monthly subscribers with a renewal in August, 2015 *who did renew*.
- Record if they got the erroneous “Thanks” email.
- Gather metrics over 30 days prior & 30 days after:
  - Days since last stint; until next stint
  - Number of days active in period
  - Number of stints in period
  - Total minutes in period
  - Top primary topic in period
  - % stints w/ top primary topic
  - Number of topics in period
  - % stints w/ top 3 primary topics

Building Data Sets – 4 of 4

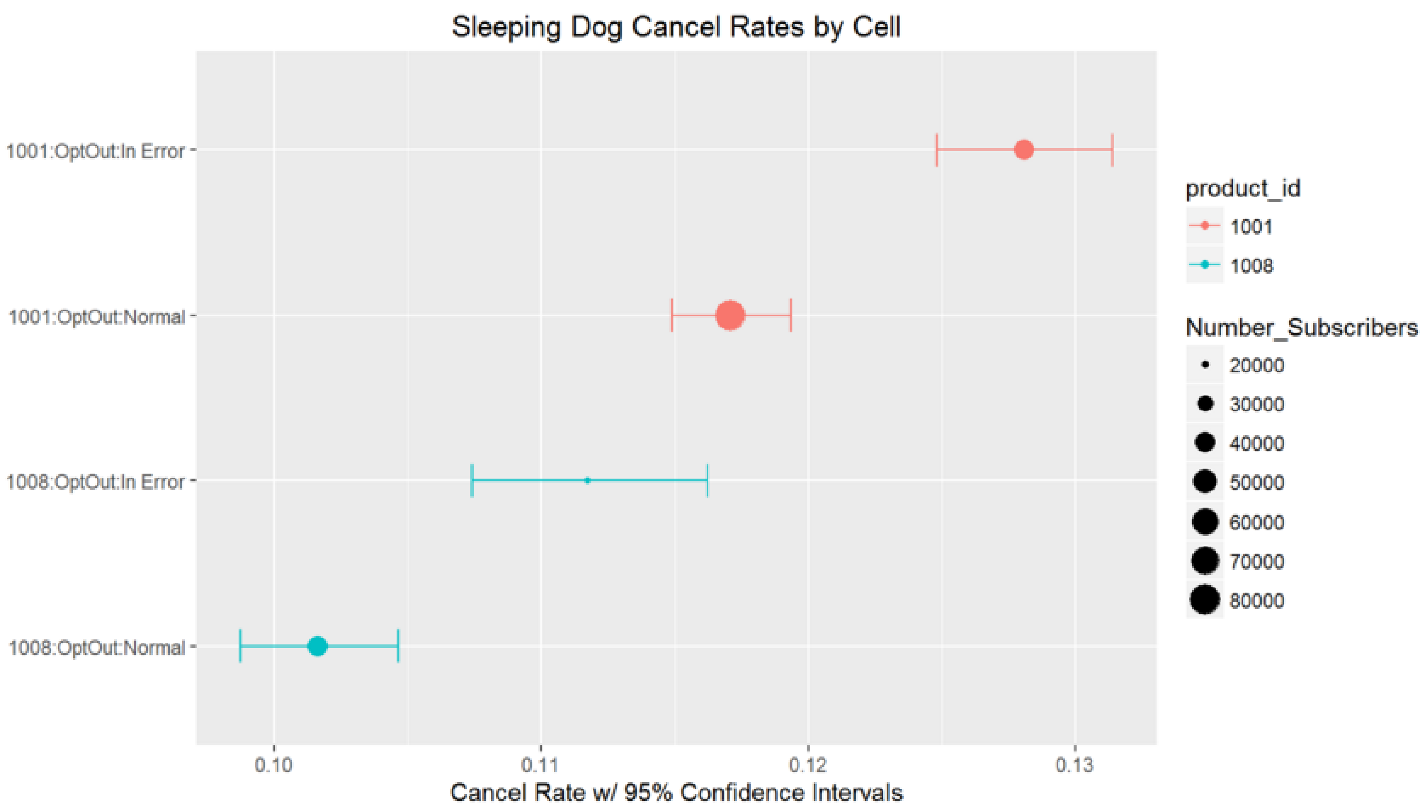
For churn and uplift questions:

- Select monthly subscribers with renewal in August 2015. *Limit metrics to before renewal date!* Random split 70/30.
- Record if they got "Thanks" email & if they canceled
- **Subscriber Metrics**: cohort quarter, initial product & promo & channel, tenure(days), # days a subscriber, % subscribed, local time UTC offset.
- **Subscription Metrics**: current product & promo, # chains, # renewals, # renewals 1st chain, RTD (revenue to date), RTD 1st chain.
- **Usage Metrics**: days since last stint, days active, % active, # stints, #stints/active day, total minutes, # minutes per active day, # top libraries, # top topics, # top software, # top levels, # courses, # video opens, # completed, % completed, % stints in top primary topic, # topics, % top 3 topics, 30 day prior metrics (see prior slide).

Full data set down sampled to 180k Monthly Subscribers with a renewal date in Aug, 2015

Cancel Rates by Cell with 95% CI

product_id	receipt_status	thank_you	Number_Subscribers	Cancel_Rate	Lower_CI	Upper_CI
1001	OptOut	In Error	39682	0.1280681	0.1248038	0.1314048
1001	OptOut	Normal	80442	0.1170906	0.1148808	0.1193371
1008	OptOut	In Error	19867	0.1117431	0.1074123	0.1162248
1008	OptOut	Normal	40416	0.1016429	0.0987224	0.1046394



Products:

1001 is “Standard Monthly” subscription.

1008 is “Premium” with a price point about 40% above the standard subscription.

Engagement Increase? – 1 of 4

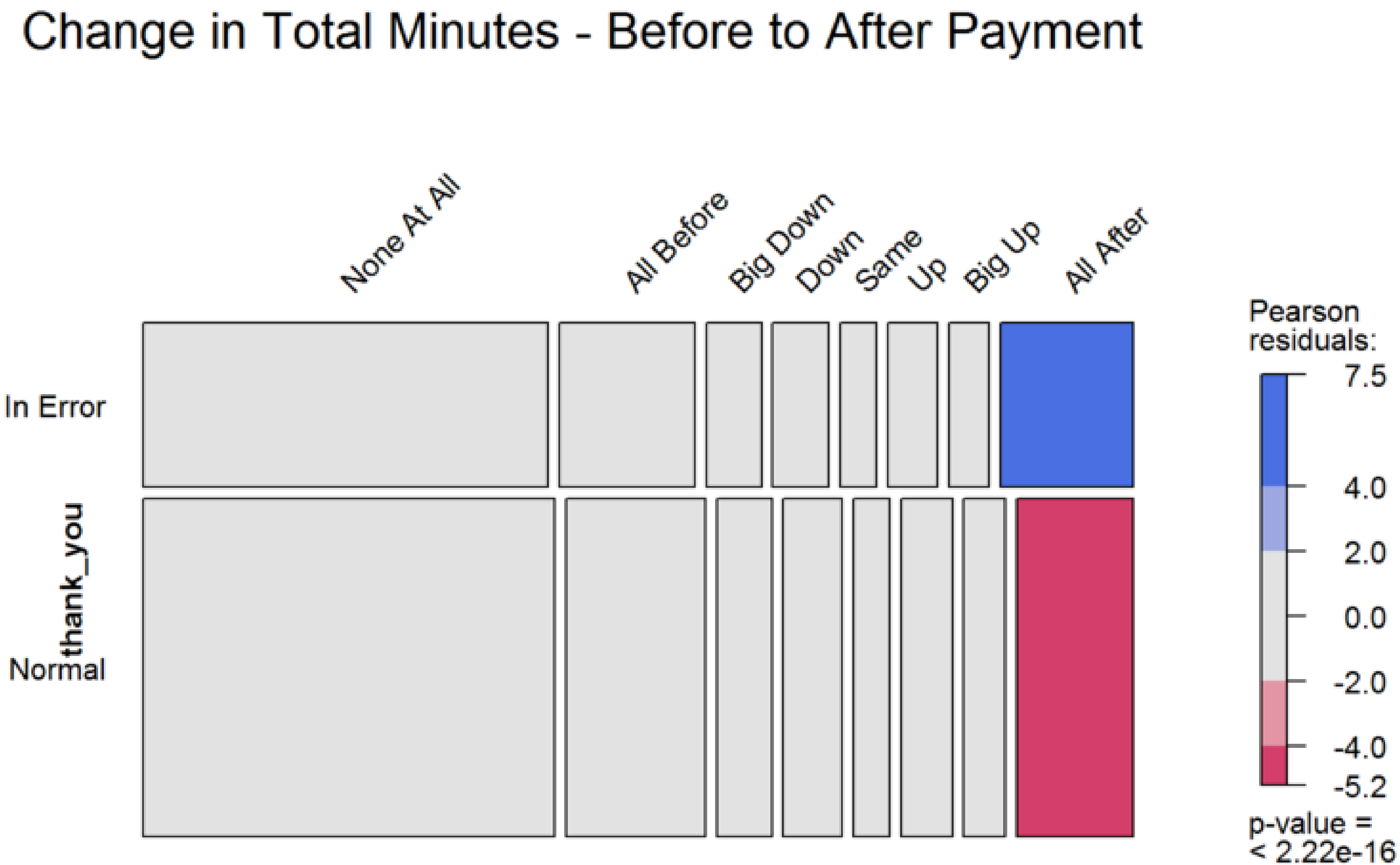
Where we compare video consumption metrics in the 30 days after renewal with the same metric 30 days prior to renewal.

Issue is a large number of subscribers have no usage stints in one or both periods.

Solution: Define “before to after” acceleration levels as a factor these levels:

Acceleration Level	Before	After	Before is NA	After is NA	After / Before Ratio
None at All			Yes	Yes	
All Before				Yes	
Big Down					< 0.25
Down					>= 0.25 & < 0.75
Same					>= 0.75 & < 1.3333
Up					>= 1.3333 & < 4
Big Up					> =4
All After			Yes		

Engagement Increase? – 2 of 4



The other Before/After metrics tell the same story.

Engagement Increase? – 3 of 4

Check the increase in proportion of “All After” counts to total counts for Total Minutes

```
> prop.test(All_After, N)

2-sample test for equality of proportions with continuity correction
data:  All_After out of N
X-squared = 95.4, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01408414 0.02135104

sample estimates:
 In Error    Normal
0.1450992 0.1273817
```

A 1.77 % point increase in subscribers who were inactive before the email and watched videos after the email.

Engagement Increase? – 4 of 4

Business implications

- Even this “frank” reminder that one is subscribed to Lynda.com can reactivate some inactive subscribers!
- Opportunity to test in future:
  1. Positive engagement messaging to inactive subscribers
  2. Uplift modeling to sort out persuadable inactive subscribers

Moving on to Churn Data Set

Three steps in this initial analysis:

1. Look at information value (IV) and weight of evidence (WOE) for binary classification problem. *Initial paring down of candidate predictors.*

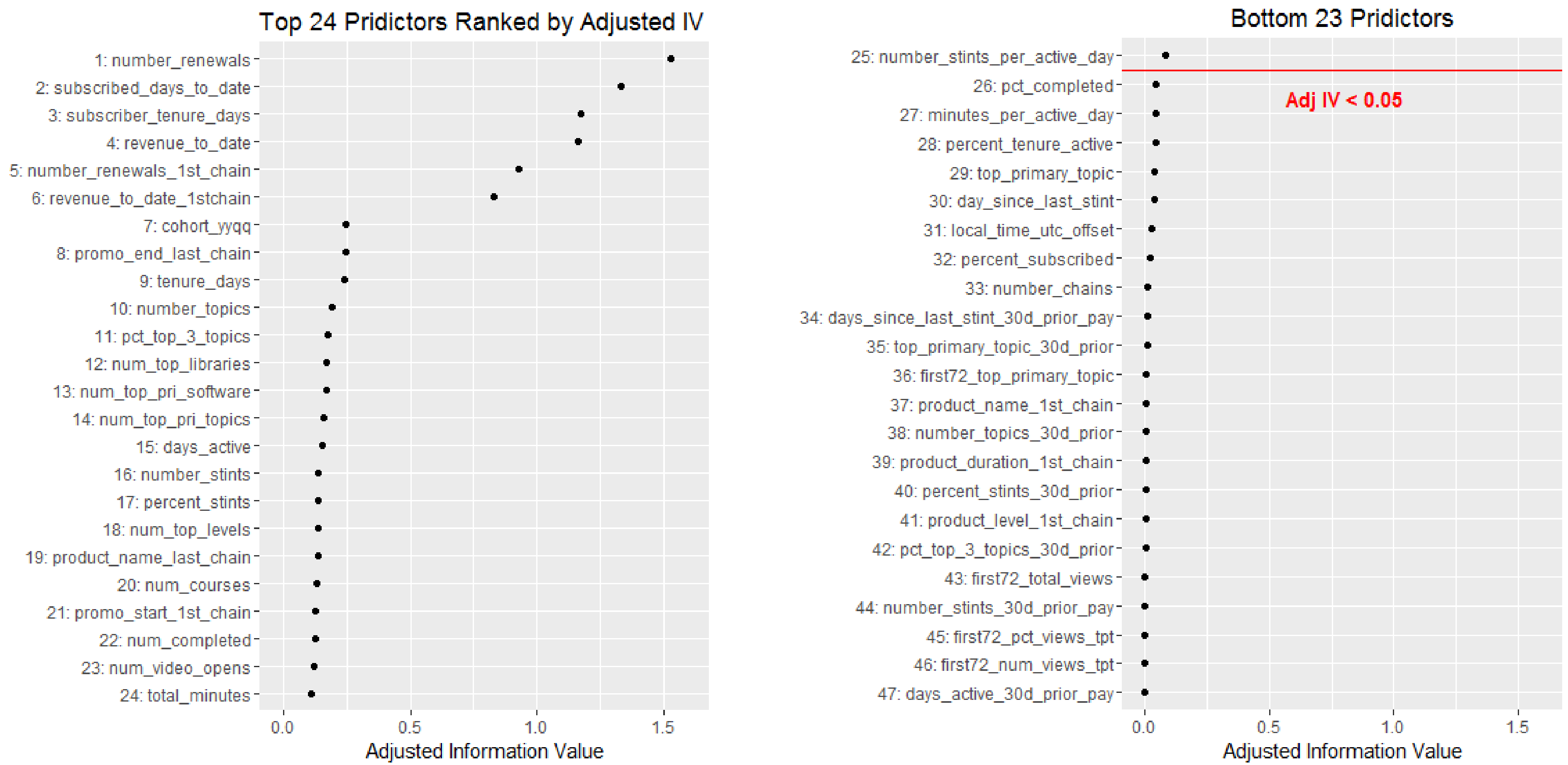
2. Look at net information value (NIV) and net weight of evidence (NWOE) for uplift problem. *Exploratory look at reasonableness of uplift effort.*

3. Do variable clustering and selection based on NIV. *To get final set of candidate predictors for uplift.*

Basically, we are following the example in Kim Larsen's [Information Package Vignette]

Churn – Info Value & WOE – 1 of 3

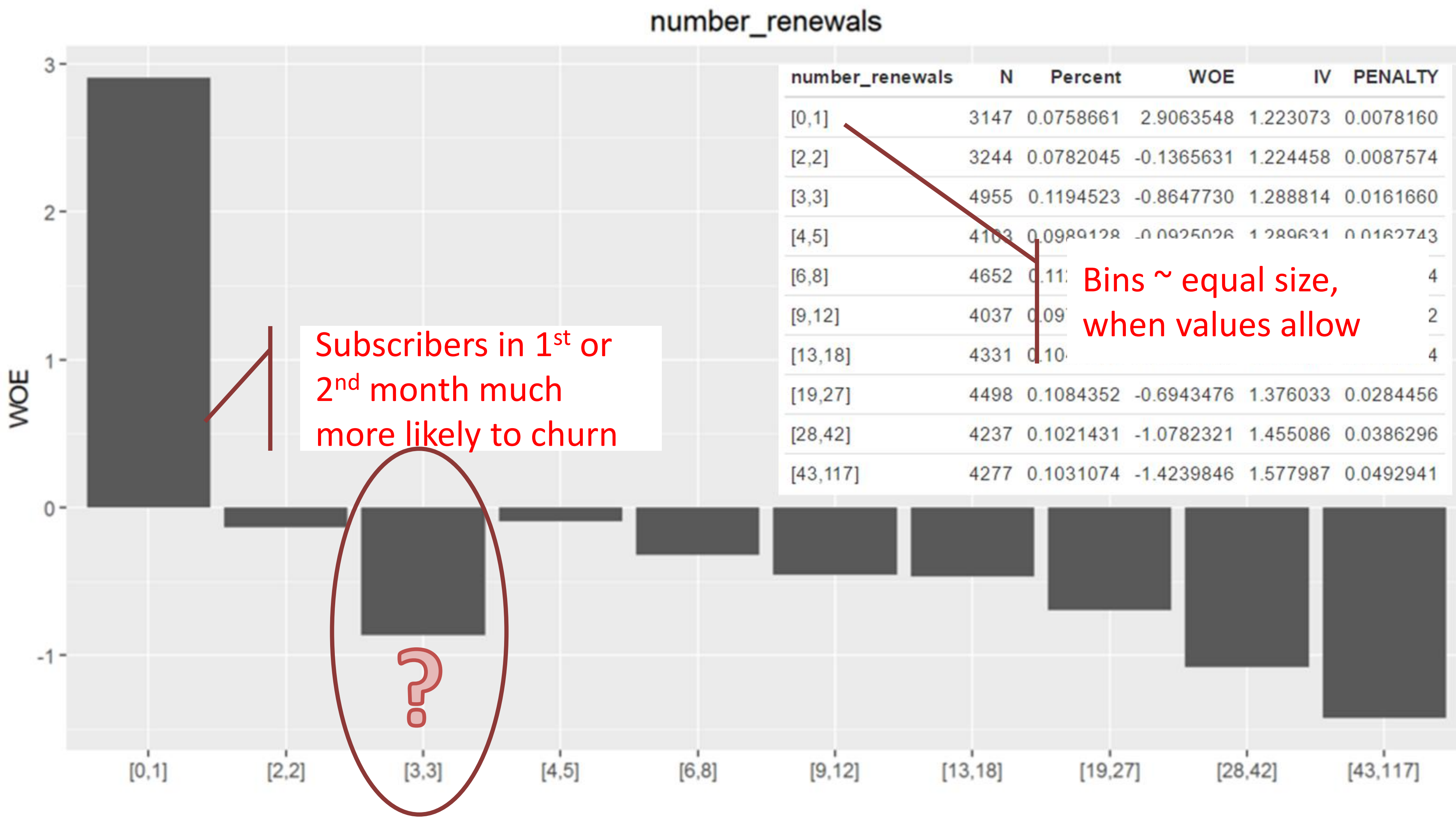
Initially we look at the binary classifier problem – did subscribers cancel?



Using Kim Larsen's *information* package in R for IV, WOE, and Variable Selection.

Churn – Info Value & WOE – 2 of 3

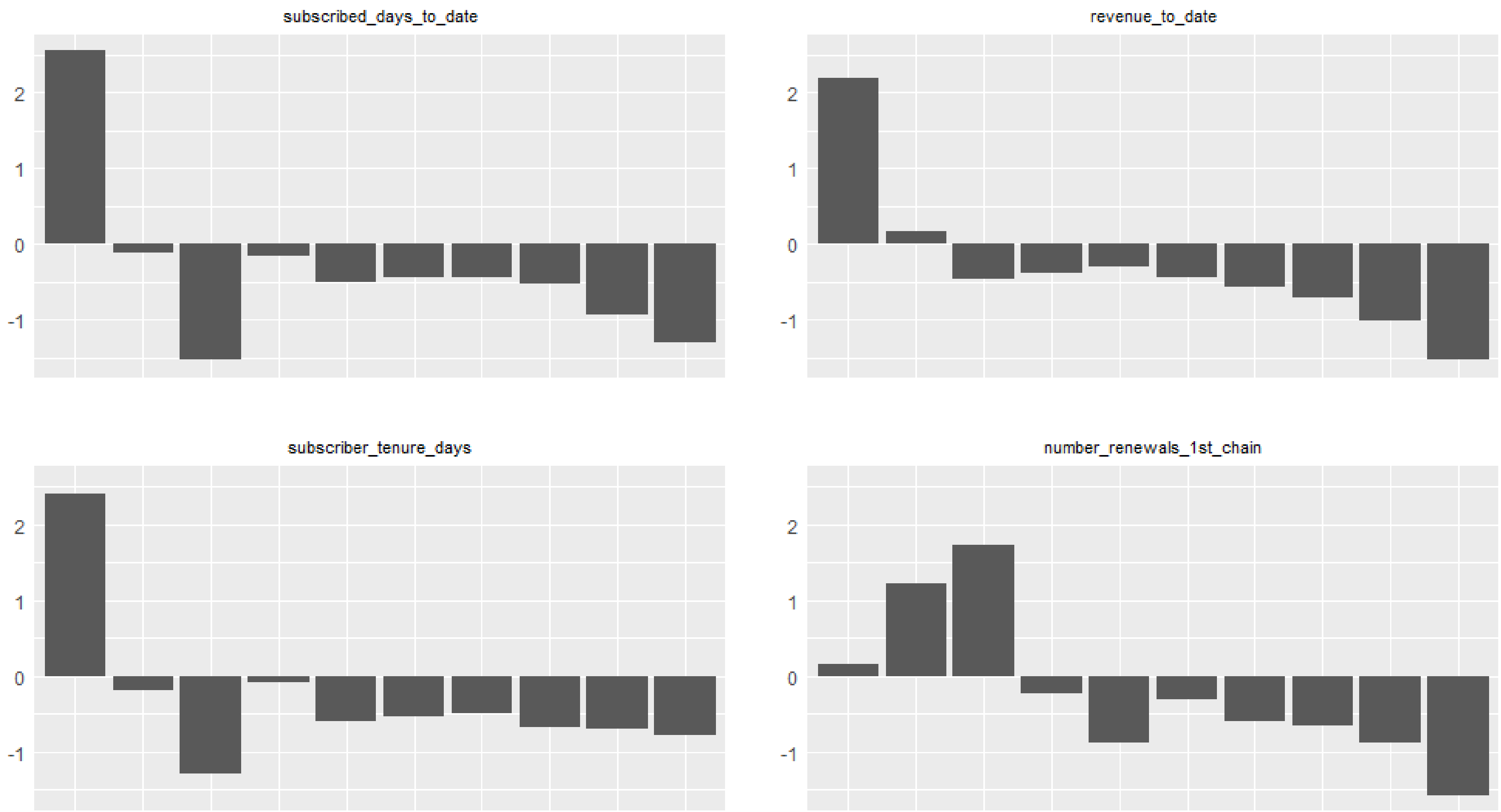
Weight of Evidence for top ranked predictor



number_renewals	N	Percent	WOE	IV	PENALTY
[0,1]	3147	0.0758661	2.9063548	1.223073	0.0078160
[2,2]	3244	0.0782045	-0.1365631	1.224458	0.0087574
[3,3]	4955	0.1194523	-0.8647730	1.288814	0.0161660
[4,5]	4103	0.0989128	-0.0925026	1.289631	0.0162743
[6,8]	4652	[illegible]	[illegible]	[illegible]	[illegible]4
[9,12]	4037	[illegible]	[illegible]	[illegible]	[illegible]2
[13,18]	4331	[illegible]	[illegible]	[illegible]	[illegible]4
[19,27]	4498	0.1084352	-0.6943476	1.376033	0.0284456
[28,42]	4237	0.1021431	-1.0782321	1.455086	0.0386296
[43,117]	4277	0.1031074	-1.4239846	1.577987	0.0492941

Churn – Info Value & WOE – 3 of 3

WOE for Predictors ranked 2nd through 5th



Uplift – Net Info Value & NWOE – 1 of 3

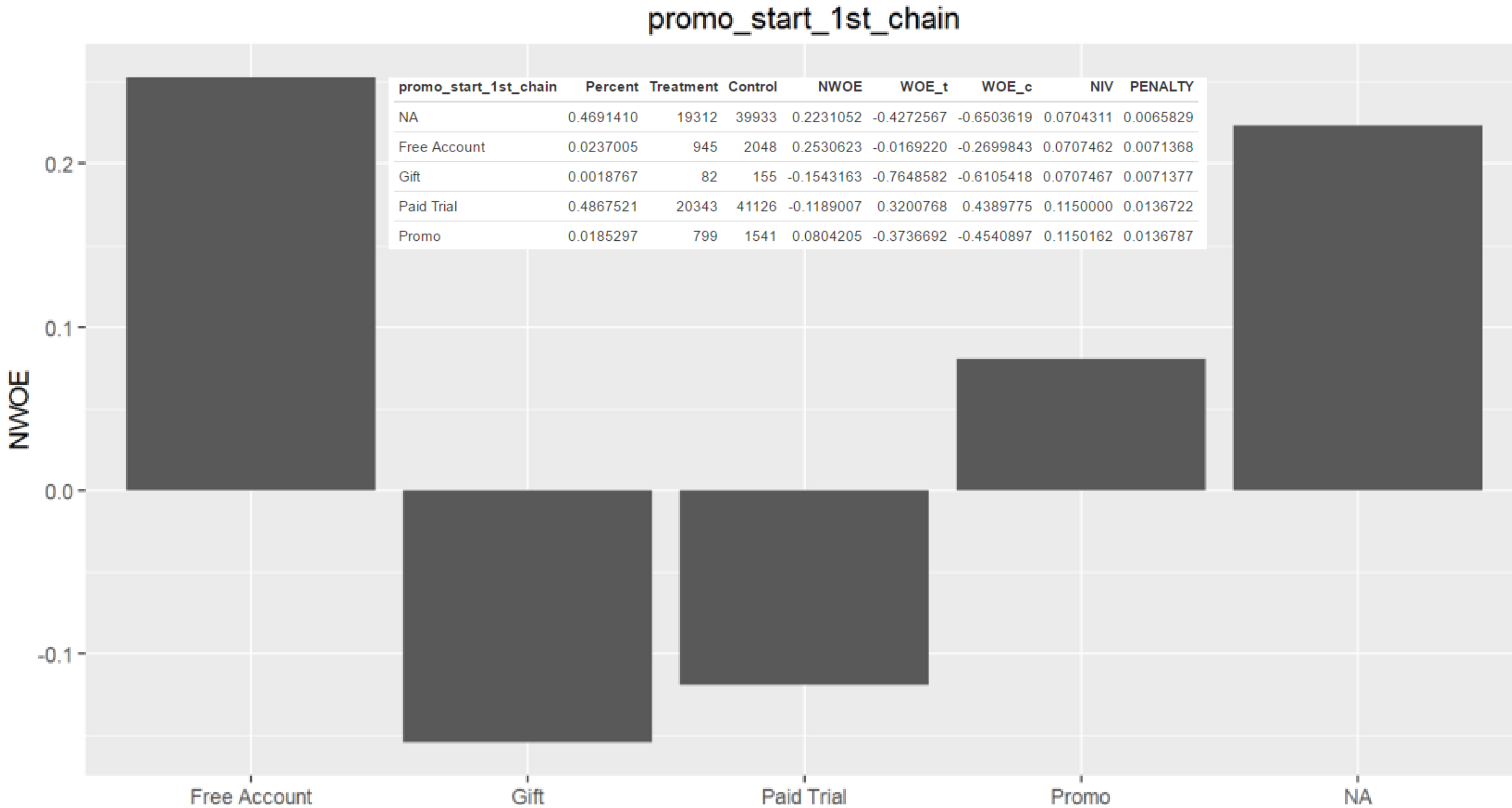
Now we look at influence of the treatment – "Thanks" email.

NIV Summary - Uplift on CANCEL with IN_ERROR

Rank	Variable	NIV	PENALTY	AdjNIV
1	promo_start_1st_chain	0.1150162	0.0136787	0.1013375
2	promo_end_last_chain	0.0867594	0.0086961	0.0780634
3	revenue_to_date	0.0597068	0.0114268	0.0482800
4	number_renewals_1st_chain	0.0509918	0.0116935	0.0392983
5	number_renewals	0.0473835	0.0083988	0.0389847
6	cohort_yyqq	0.0450361	0.0062491	0.0387870
7	subscribed_days_to_date	0.0415793	0.0076231	0.0339562
8	subscriber_tenure_days	0.0408931	0.0085478	0.0323453
9	revenue_to_date_1stchain	0.0285060	0.0060453	0.0224607
10	num_top_libraries	0.0072613	0.0019881	0.0052733
11	num_top_levels	0.0055762	0.0020531	0.0035230
12	number_stints_per_active_day	0.0053177	0.0023751	0.0029427
13	num_video_opens	0.0093161	0.0067694	0.0025467
14	days_active	0.0086517	0.0063290	0.0023226
15	num_courses	0.0077882	0.0060121	0.0017760

Uplift – Net Info Value & NWOE – 2 of 3

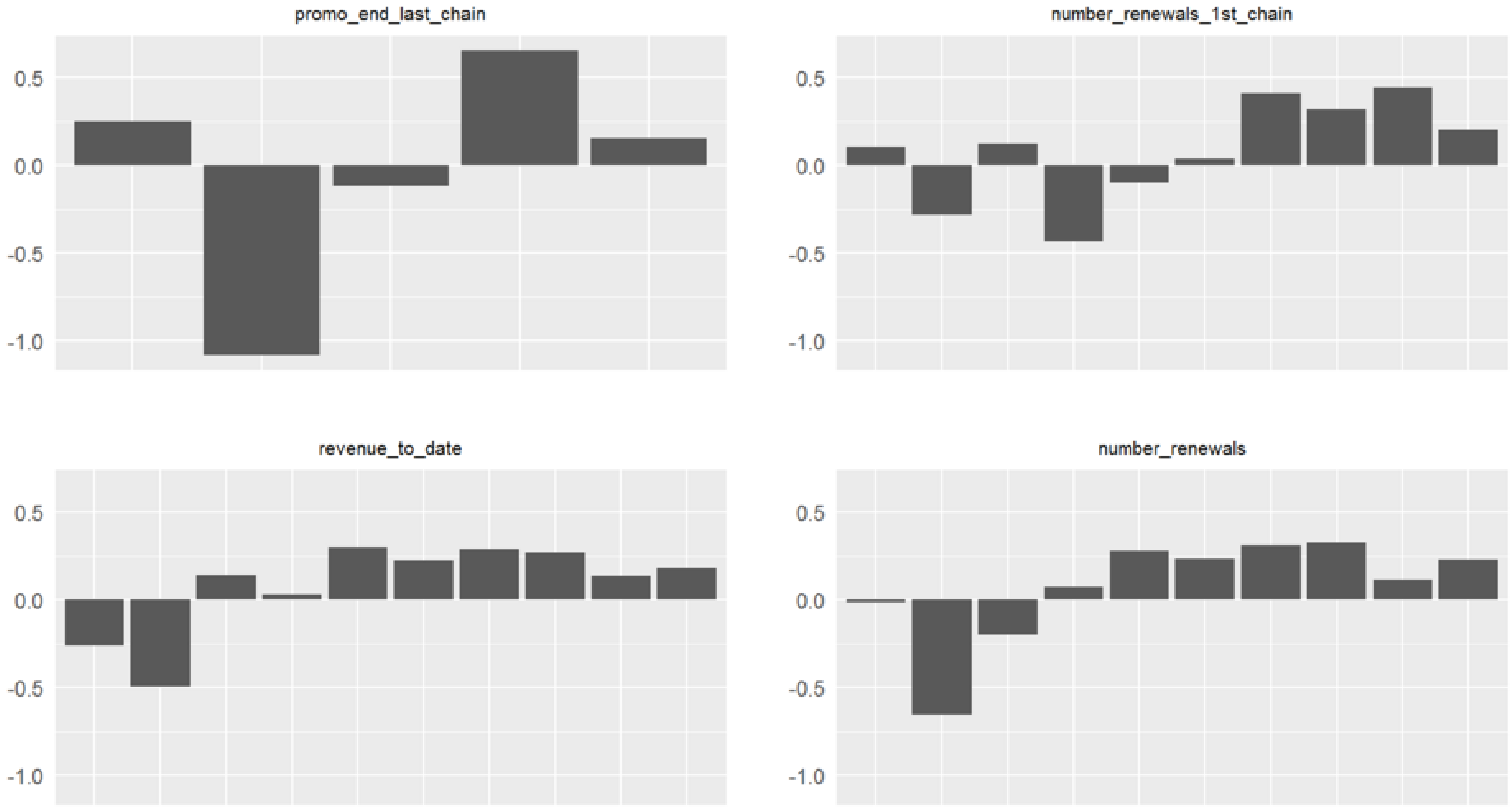
Net WOE for top ranked predictor



promo_start_1st_chain	Percent	Treatment	Control	NWOE	WOE_t	WOE_c	NIV	PENALTY
NA	0.4691410	19312	39933	0.2231052	-0.4272567	-0.6503619	0.0704311	0.0065829
Free Account	0.0237005	945	2048	0.2530623	-0.0169220	-0.2699843	0.0707462	0.0071368
Gift	0.0018767	82	155	-0.1543163	-0.7648582	-0.6105418	0.0707467	0.0071377
Paid Trial	0.4867521	20343	41126	-0.1189007	0.3200768	0.4389775	0.1150000	0.0136722
Promo	0.0185297	799	1541	0.0804205	-0.3736692	-0.4540897	0.1150162	0.0136787

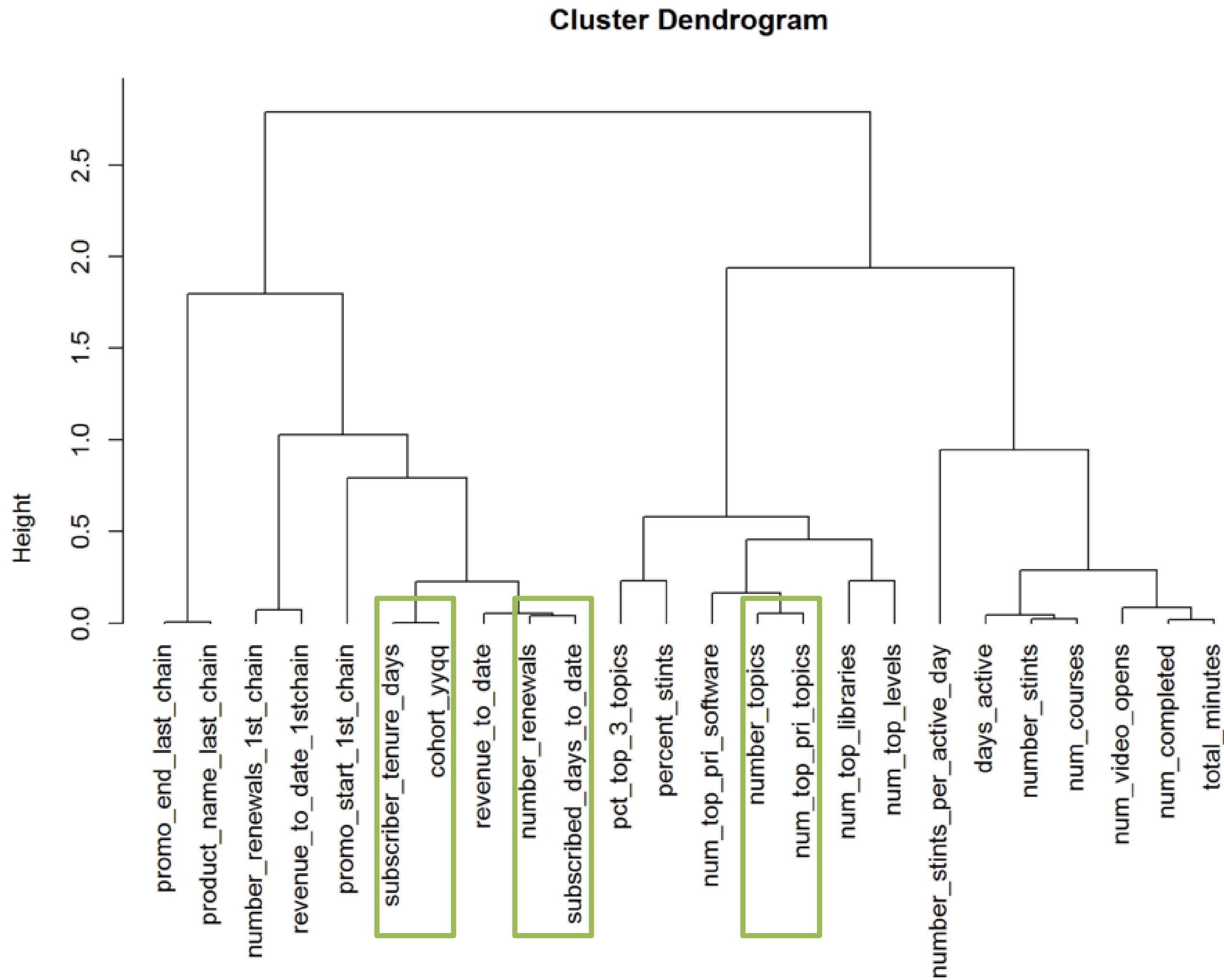
Uplift – Net Info Value & NWOE – 3 of 3

Net WOE for 2nd through 5th ranked predictors



Net IV Variable Clustering – 1 of 3

From each cluster, pick variable with highest Net IV



Net IV Variable Clustering – 2 of 3

From each cluster, pick variable with highest Net IV.
Here are first six clusters:

Cluster	Variable	IV	PENALTY	AdjIV	Rank
1	number_renewals	1.5779868	0.0492941	1.5286927	1
1	subscribed_days_to_date	1.3633481	0.0303146	1.3330335	2
2	subscriber_tenure_days	1.2020536	0.0292712	1.1727824	1
2	cohort_yyqq	0.2793005	0.0329926	0.2463080	2
3	revenue_to_date	1.2030511	0.0420015	1.1610496	1
4	number_renewals_1st_chain	0.9578077	0.0275698	0.9302379	1
5	revenue_to_date_1stchain	0.8736754	0.0461355	0.8275399	1
6	number_topics	0.2213035	0.0305025	0.1908010	1
6	num_top_pri_topics	0.1870420	0.0279144	0.1591275	2

Net IV Variable Clustering – 3 of 3

The final 17 predictors to be passed to uplift modeling.

Cluster	Variable	IV	PENALTY	AdjIV
1	number_renewals	1.5779868	0.0492941	1.5286927
2	subscriber_tenure_days	1.2020536	0.0292712	1.1727824
3	revenue_to_date	1.2030511	0.0420015	1.1610496
4	number_renewals_1st_chain	0.9578077	0.0275698	0.9302379
5	revenue_to_date_1stchain	0.8736754	0.0461355	0.8275399
16	promo_end_last_chain	0.2521735	0.0077314	0.2444421
6	number_topics	0.2213035	0.0305025	0.1908010
7	pct_top_3_topics	0.2010102	0.0268427	0.1741675
8	num_top_libraries	0.1810161	0.0100911	0.1709250
9	num_top_pri_software	0.1892514	0.0205555	0.1686959
10	days_active	0.1899305	0.0334434	0.1564871
11	percent_stints	0.1537172	0.0149828	0.1387344
12	num_top_levels	0.1525107	0.0162759	0.1362347
17	promo_start_1st_chain	0.1320440	0.0037101	0.1283339
13	num_completed	0.1564453	0.0301728	0.1262724
14	num_video_opens	0.1440695	0.0248437	0.1192258
15	number_stints_per_active_day	0.0980979	0.0160791	0.0820187

Now we are ready for *uplift*

Analysis steps:

1. Check train/validate split did not introduce bias
2. Run both upliftRF & ccif methods – pick best for deep dive
3. Get NIV via *uplift*. Compare with what we got from *Information.*
4. Plot relative importance of candidate predictors.
5. Profile resulting model: Predicted uplift & predictors
6. Business implications. What did we learn? Next steps?

Basically, following Chapters 10 & 11 in Leo Guelman's PhD Thesis: *Optimal personalized treatment learning models with insurance applications*.

Uplift Modeling – 1 of 8

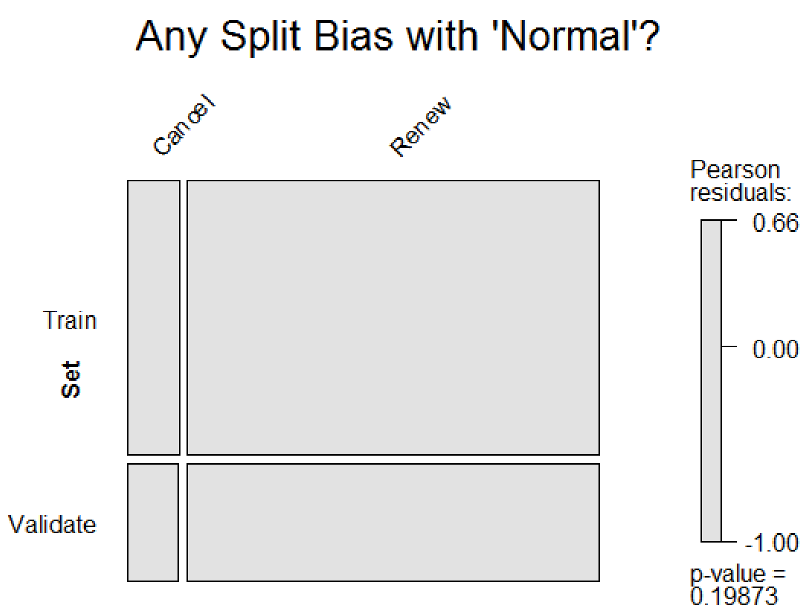
First check that training & validation set splits ~ same

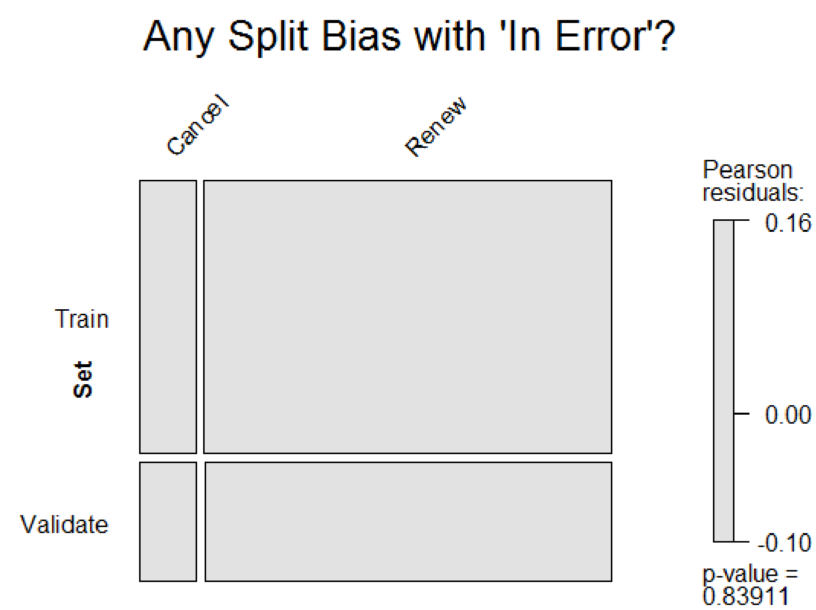
Training Set Splits

	Normal	In Error
Renew	0.8873153	0.8775584
Cancel	0.1126847	0.1224416

Validation Set Splits

	Normal	In Error
Renew	0.8898627	0.8769648
Cancel	0.1101373	0.1230352





Uplift Modeling – 2 of 8

The two modeling methods we looked at.

Uplift Random Forest (aka upliftRF) → Used This → **Causal Conditional Inference Forests (aka ccif)**

Uplift Random Forest (aka upliftRF)

- A variation of the classic Leo Breiman method.
- Well known
- Pretty fast
- Issues – see ccif

Causal Conditional Inference Forests (aka ccif)

- Fixes issues with upliftRF

1. Overfitting
2. Selection bias toward covariates with many possible splits

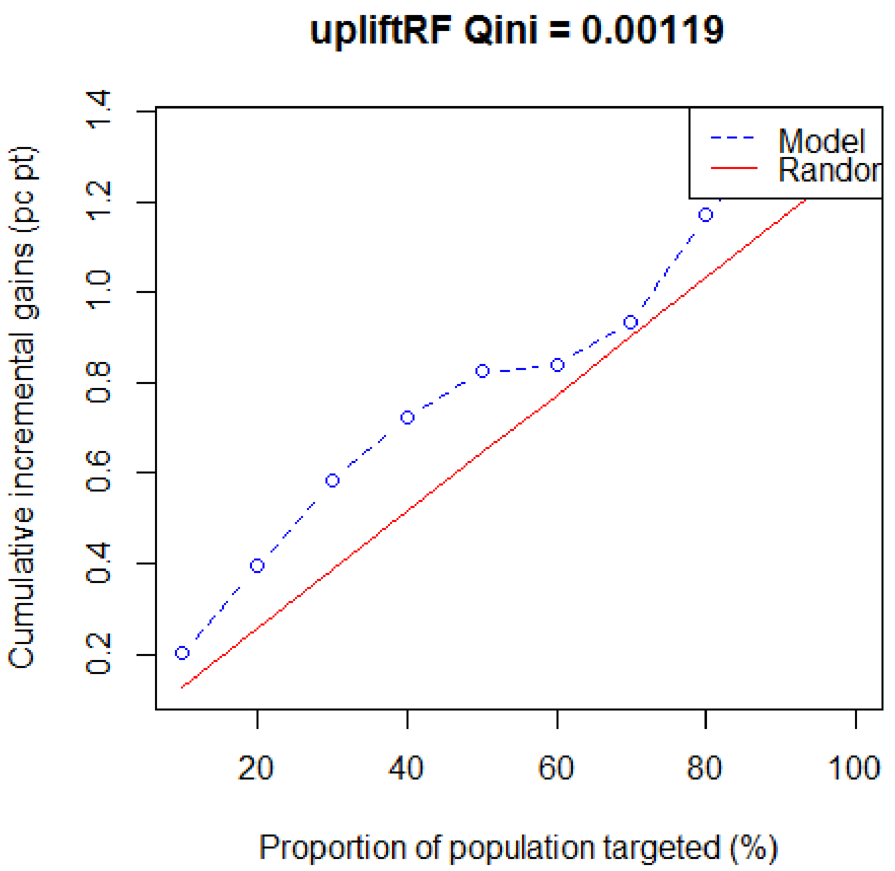
- Slow (*but Leo working on update*)
- Better lift & better story

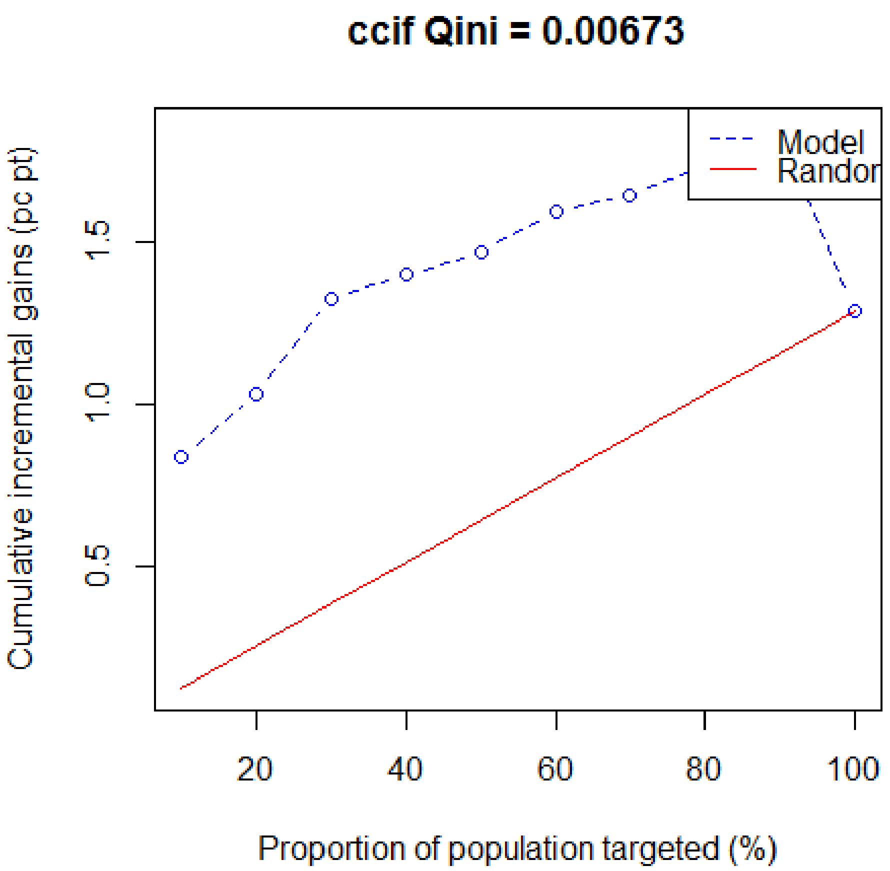
We are using the R package “uplift” by Leo Guelman.

See Appendix for pseudo code of each method & references.

Uplift Modeling – 3 of 8

Qini Curves & Coefficient for Each Method





Uplift Modeling – 4 of 8

Net IV from *uplift*

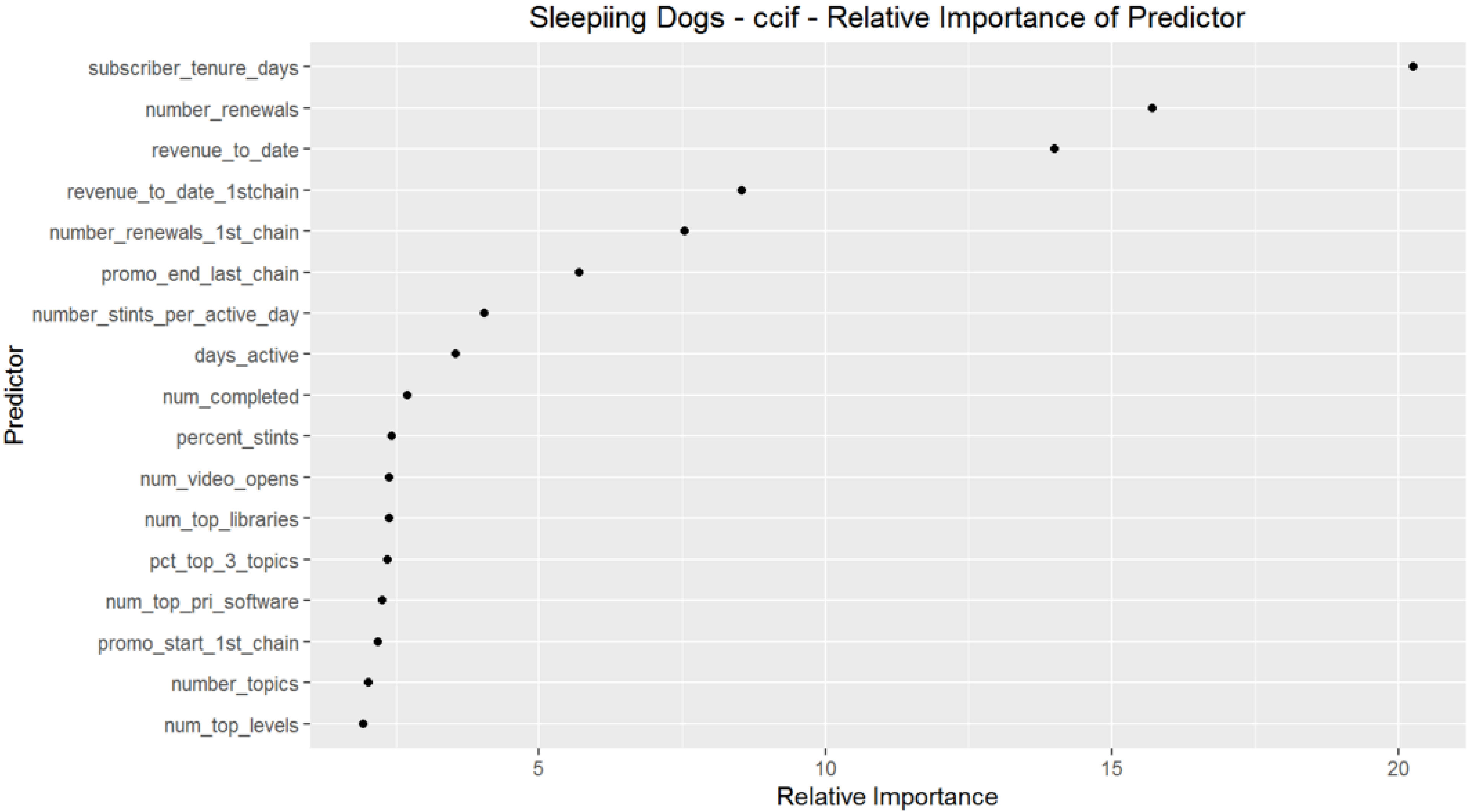
	niv	penalty	adj_niv
number_renewals	1.710	0.0617	1.6483
subscriber_tenure_days	1.380	0.0261	1.3539
promo_start_1st_chain	1.157	0.0347	1.1223
revenue_to_date	1.010	0.0361	0.9739
promo_end_last_chain	0.916	0.0586	0.8574
number_renewals_1st_chain	0.851	0.0609	0.7901
revenue_to_date_1stchain	0.374	0.0328	0.3412
num_completed	0.145	0.0111	0.1339
num_video_opens	0.129	0.0129	0.1161
num_top_pri_software	0.090	0.0109	0.0791
number_topics	0.091	0.0147	0.0763
days_active	0.079	0.0082	0.0708
number_stints_per_active_day	0.085	0.0142	0.0708
percent_stints	0.070	0.0101	0.0599
pct_top_3_topics	0.065	0.0120	0.0530
num_top_levels	0.062	0.0101	0.0519
num_top_libraries	0.061	0.0168	0.0442

Net IV from *information*

	Cluster	Variable	IV	PENALTY	AdjIV
	1	number_renewals	1.5779868	0.0492941	1.5286927
	2	subscriber_tenure_days	1.2020536	0.0292712	1.1727824
▲1	3	revenue_to_date	1.2030511	0.0420015	1.1610496
▲2	4	number_renewals_1st_chain	0.9578077	0.0275698	0.9302379
▲2	5	revenue_to_date_1stchain	0.8736754	0.0461355	0.8275399
▼1	16	promo_end_last_chain	0.2521735	0.0077314	0.2444421
▲4	6	number_topics	0.2213035	0.0305025	0.1908010
▲7	7	pct_top_3_topics	0.2010102	0.0268427	0.1741675
▲8	8	num_top_libraries	0.1810161	0.0100911	0.1709250
	9	num_top_pri_software	0.1892514	0.0205555	0.1686959
▲1	10	days_active	0.1899305	0.0334434	0.1564871
▲1	11	percent_stints	0.1537172	0.0149828	0.1387344
▲3	12	num_top_levels	0.1525107	0.0162759	0.1362347
▼11	17	promo_start_1st_chain	0.1320440	0.0037101	0.1283339
▼7	13	num_completed	0.1564453	0.0301728	0.1262724
▼7	14	num_video_opens	0.1440695	0.0248437	0.1192258
▼4	15	number_stints_per_active_day	0.0980979	0.0160791	0.0820187

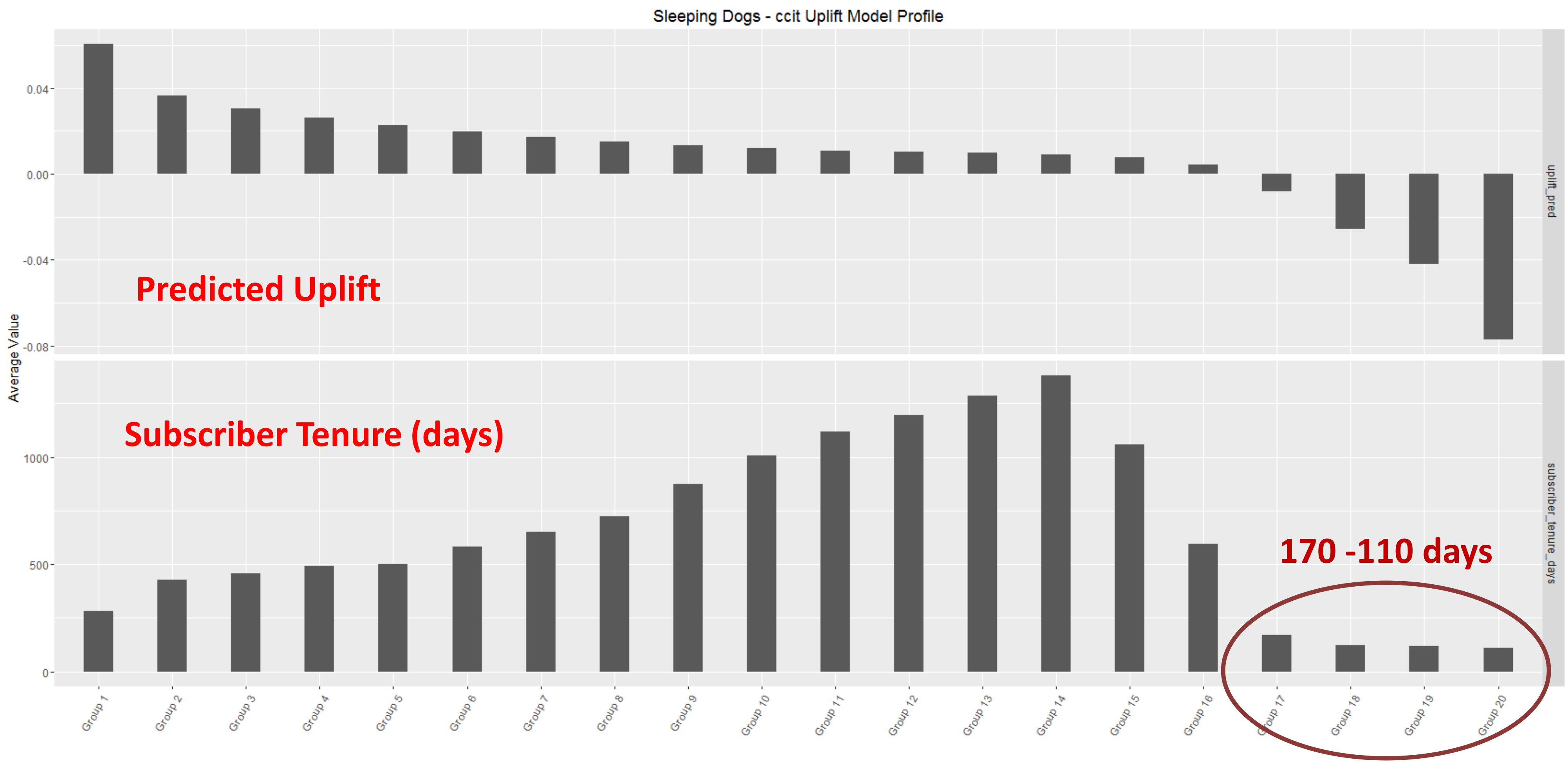
Uplift Modeling – 5 of 8

ccif Relative Importance (in separating positive & negative outcomes)



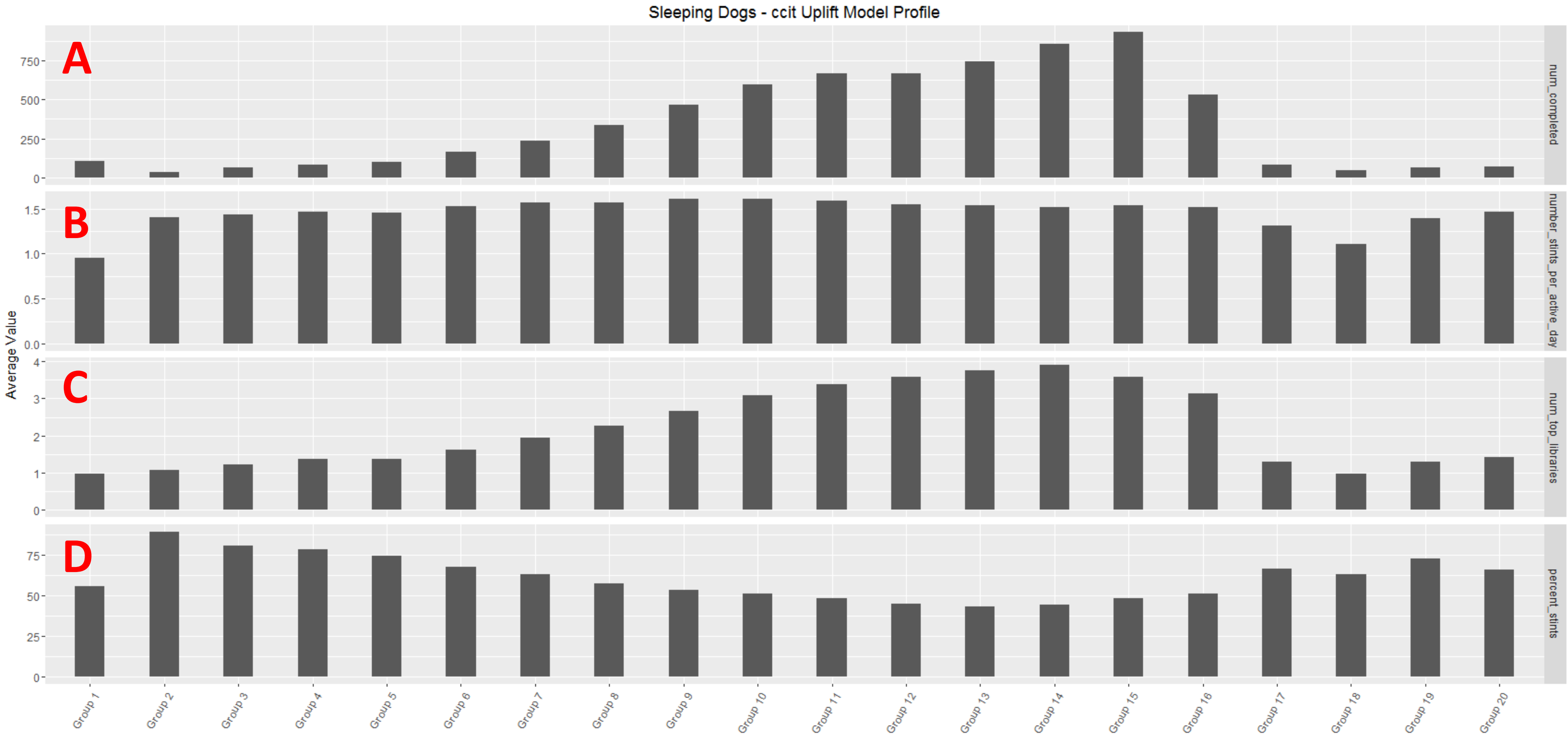
Uplift Modeling – 6 of 8

ccif Predicted Uplift & Top Predictor – Subscriber Tenure (days)
Grouped in vigintiles by predicted uplift ranking.



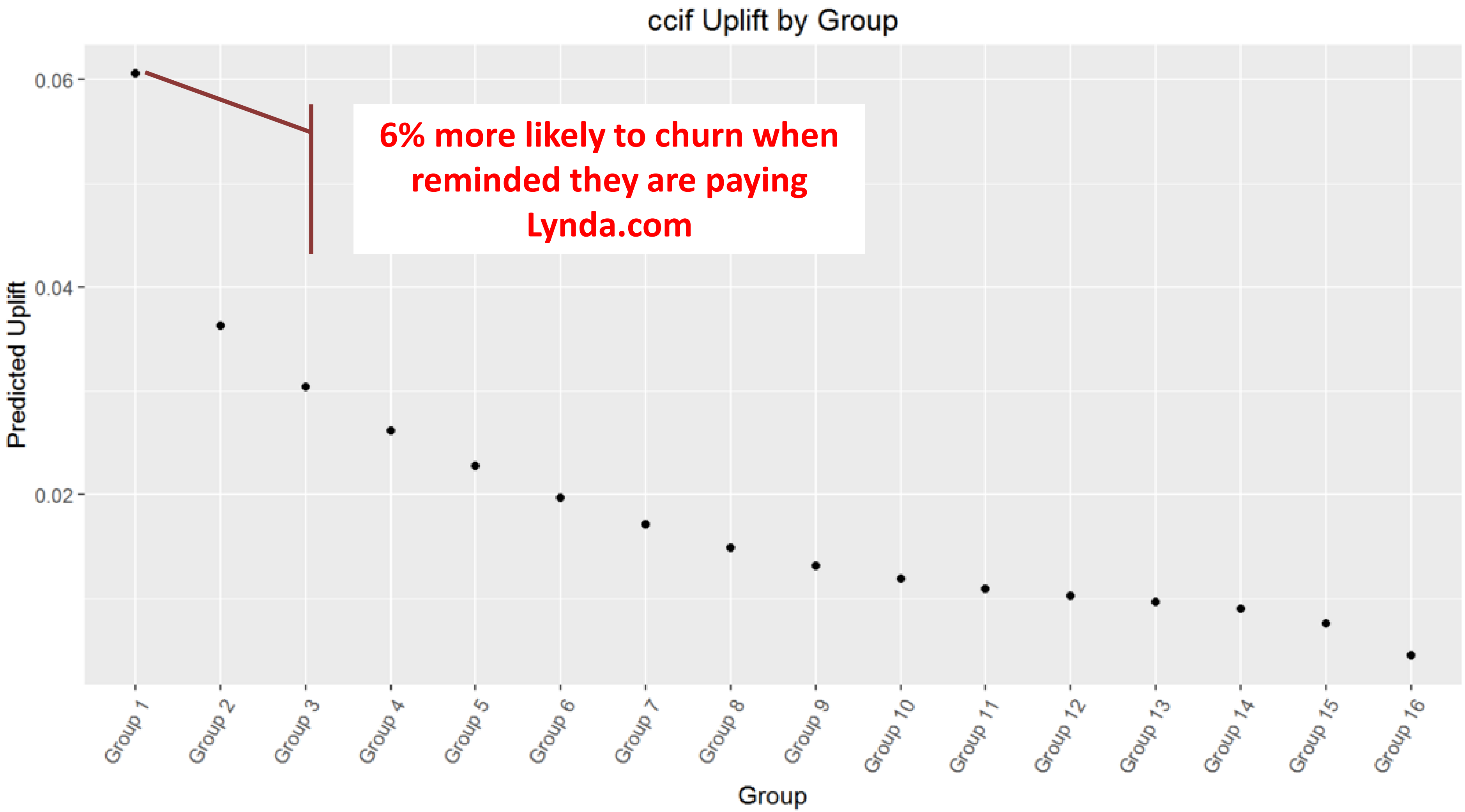
Uplift Modeling – 7 of 8

Four interesting predictors: A) # Videos Completed, B) # Stints / Active Day, C) # Top Libraries, D) % Stints w/ Top Primary Topic



Uplift Modeling – 8 of 8

Groups with positive uplift



Next Steps

Experiments to:

1. Determine best uplift groups to increase engagement of inactive subscribers.
2. Try a positive “You are a member” message rather than a negative “Thanks for you payment” message.

Data deep dive:

1. Figure out who are the Group 17-20 subscribers (those with negative uplift) who’s first subscription was 110 – 170 days before August, 2015 billing.

Remember the business question?

What did the erroneous “Thanks” email do?

✓ • *Did We Wake Sleeping Dogs?*
  – *Cause a cancel which would not have happened.*

✓ • *Reactivate Engagement?*
  – *Cause an increase in video viewing.*

✓ • *Do Nothing At All?*

Subscribers, being people, responded differently to the “Thanks” email. Some negatively, some positively and, for most, it had no effect. Isn’t marketing fun!

What We Covered

- Lynda.com – What they do & what happened
- Building data sets
- Apply *Information* package for IV & WOE
- Variable selection for uplift model
- Apply *uplift* package
- Business conclusions

Contacts:
MingNg.LinkedIn.com
Jim@DS4CI.org

Questions? Comments?
Now is the time!

APPENDIX

1. Uplift algorithm pseudo code for upliftRF & ccit.
2. R environment used in this analysis.
3. Learning More – Where to Start?

upliftRF Uplift Pseudocode

Algorithm 1 Uplift random forest

for $b = 1$ to B **do**
  Sample a fraction ν of the training observations L without replacement
  Grow an uplift decision tree UT_b to the sampled data:
  for each terminal node **do**
    repeat
      Select n covariates at random from the p covariates
      Select the best variable/split-point among the n covariates based on KL_{ratio}
      Split the node into two branches
    until a minimum node size l_{min} is reached
  end for
end for
Output the ensemble of uplift trees UT_b; $b = \{1, \ldots, B\}$
The predicted personalized treatment effect for a new data point $\mathbf{x}$, is obtained by averaging the predictions of the individual trees in the ensemble: $\hat{\tau}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} UT_b(\mathbf{x})$

Where,
B is # trees to grow,

KL is Kulback-Leiber distance

From: Guelman, L., Guillen, M. and Perez-Marin, A.M. (2014) “Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study.”, UB Riskcenter Working Papers Series 2014-06

ccif Uplift Pseudocode

Algorithm 2 Causal conditional inference forests

for $b = 1$ to B **do**
  Draw a sample with replacement from the training observations L such that P(A=1) = P(A=0) = 1/2
  Grow a conditional causal inference tree $CCIT_b$ to the sampled data:
  for each terminal node **do**
    repeat
      Select n covariates at random from the p covariates
      Test the global null hypothesis of no interaction effect between the treatment A and any of the n covariates (i.e., $H_0 = \cap_{j=1}^{n} H_0^j$, where $H_0^j : E[W|X_j] = E[W]$) at a level of significance α based on a permutation test
      if the null hypothesis H_0 cannot be rejected **then**
        Stop
      else
        Select the j^*th covariate X_{j*} with the strongest interaction effect (i.e., the one with the smallest adjusted P value)
        Choose a partition Ω^* of the covariate X_{j*} in two disjoint sets $\mathcal{M} \subset X_{j*}$ and $X_{j*} \setminus \mathcal{M}$ based on the $G^2(\Omega)$ split criterion
      end if
    until a minimum node size l_{min} is reached
  end for
end for
Output the ensemble of causal conditional inference trees $CCIT_b$; $b = \{1, \ldots, B\}$
The predicted personalized treatment effect for a new data point $\mathbf{x}$, is obtained by averaging the predictions of the individual trees in the ensemble: $\hat{\tau}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} CCIT_b(\mathbf{x})$

Where,
B is # trees to grow,
A is treatment flag,

G^2 is split criteria proposed by Su *et al.* See Guelman *et al*, equation (19)

From: Guelman, L., Guillen, M. and Perez-Marin, A.M. (2014) “Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study.”, UB Riskcenter Working Papers Series 2014-06

R Environment Used

- R download https://cran.rstudio.com/index.html
- RStudio download https://www.rstudio.com/
- R Packages (https://cran.rstudio.com/web/packages/) :
- Hadley Wickham: *ggplot2, dplyr, tidyr, readr, stringr*
- Yihui Xie: *knitr*
- David Meyer, et al: *vcd*
- Michael Friendly: *vcdExtra*
- Kim Larsen: *Information*
- Marie Chavent, et al: *ClustOfVar*
- Leo Guelman: *uplift*

Learning More – Where to Start?

- Jim's Archives www.ds4ci.org/archives
  - Structuring Data for Customer Insights for more about the Redshift CI datamart.
- Uplift Modeling
  - Eric Siegel, *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die – Revised and Updated.* (2016) Chapter 7.
  - PAW SF 2016 Sessions:
    - Eric Sigel, Case Study: U.S. Bank; Uplift Modeling: Optimize for Influence and Persuade by the Numbers
    - Patrick Surry, Case Study: Telenor; Applying Next Generation Uplift Modeling to Optimize Customer Retention Programs
  - Leo Guelman, et al (the author of the R package *uplift):*
    - Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study
    - Optimal personalized treatment learning models with insurance applications. (PhD Thesis)
  - Michal Soltys et al, Ensemble methods for uplift modeling
- Information Value & Weight of Evidence
  - Kim Larsen – stichfix blog post or Information package vignette
- Visualizing categorical data
  - Vignettes in *vcd* and *vcdextra* packages
  - *Discrete Data Analysis with R*, Friendly & Meyer, CRC Press (2015)