Marketing Analytics Q1 2016 – Kaggle Project

Samantha Vo - Lorenza Lin - Hana Xu
Peter Bergen - Carlos Escalante Trevino

COMPANY PROFILE

Allstate is an insurance company. Its car insurance product package consists of 7 insurance options, each has several coverage levels.

Some option examples:

Collision (damage caused by collision with other vehicle or car rolling over)

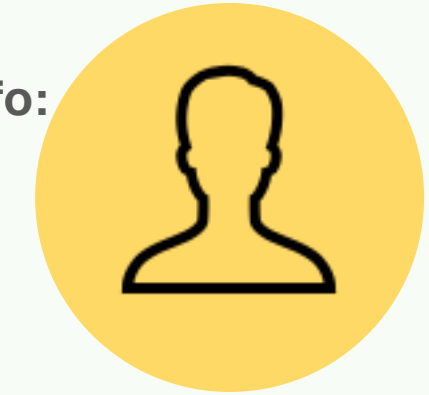
Bodily Injury Liability (damages resulting from injury of another person when you are at fault)

Uninsured Motorist Property Damage (when an uninsured driver can't afford to pay)

QUOTING PROCESS

After filling out initial info, and getting a suggested quotes, users can fiddle with the features to hit their target.

1.User provide info:
- Demographic
- Car Information



2. System gives quote

3. User edits features

4. If satisfied, purchase!

PROJECT GOALS

After filling out initial info, and getting a suggested quotes, users can fiddle with the features to hit their target. Allstate believes that longer it takes for the customer to find their ideal quote, the more likely they are to go to different provider. It ran a **Kaggle** competition to get these insights.

I. **Generate** the best “recommended product” for each customer

- Predict what coverage levels under each option the customer is most likely to choose

II. **Evaluate** Allstate’s data analytics practices

SAMPLE DATA

Kaggle submission: take a separate test file with more limited quote viewing history, then predict the product (i.e. 7-digit combination of variables A-G) that each customer will end up purchasing (two columns in submission file: customer_id and product_purchased

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
1	customer_ID	shopping_pt	record_type	day	time	state	location	group_size	homeowner	car_age	car_value	risk_factor	age_oldest	age_younges	married_cou	C_previous	duration_pre	A	B
2	10000000	1	0	0	8:35	IN	10001	2	0	2	g	3	46	42	1	1	2	1	
3	10000000	2	0	0	8:38	IN	10001	2	0	2	g	3	46	42	1	1	2	1	
4	10000000	3	0	0	8:39	IN	10001	2	0	2	g	3	46	42	1	1	2	1	
5	10000000	4	0	0	8:41	IN	10001	2	0	2	g	3	46	42	1	1	2	1	
6	10000000	5	0	0	8:43	IN	10001	2	0	2	g	3	46	42	1	1	2	1	
7	10000000	6	0	0	8:44	IN	10001	2	0	2	g	3	46	42	1	1	2	1	
8	10000000	7	0	0	11:38	IN	10001	2	0	2	g	3	46	42	1	1	2	1	
9	10000000	8	0	0	12:03	IN	10001	2	0	2	g	3	46	42	1	1	2	1	
10	10000000						10001	2	0	2	g	3	46	42	1	1	2	1	
11	10000000						10006	1	0	10	e	4						1	
12	10000000						10006	1	0	10	e	4						1	
13	10000000						10006	1	0	10	e	4						1	
14	10000005	4	0	3	8:58	NY	10006	1	0	10	e	4						0	
15	10000005	5	0	3	8:58	NY	10006	1	0	10	e	4						0	
16	10000005	6	1	3	9:09	NY	10006	1	0	10	e	4	28	28	0	3	13	0	
17	10000007	1	0	4	8:35	PA	10008	1	0	11	c	NA	43	43	0	2	4	0	
18	10000007	2	0	4	8:36	PA	10008	1	0	11	c	NA	43	43	0	2	4	0	

0 = viewed only
1 = viewed and purchased

Unique product view (each customer has at least 3, some up to 15)

Indicates option selected for category A (7 total categories, each with a different number of options)

MENTAL MODEL

Our hypothesis: car status and income level should generate the biggest impact on the level of options chosen.

- Newer car → more expensive options
- Lots of miles driven → more expensive
- Higher income → more expensive
  - If married → higher income + may have children → more expensive
- Older age → more expensive

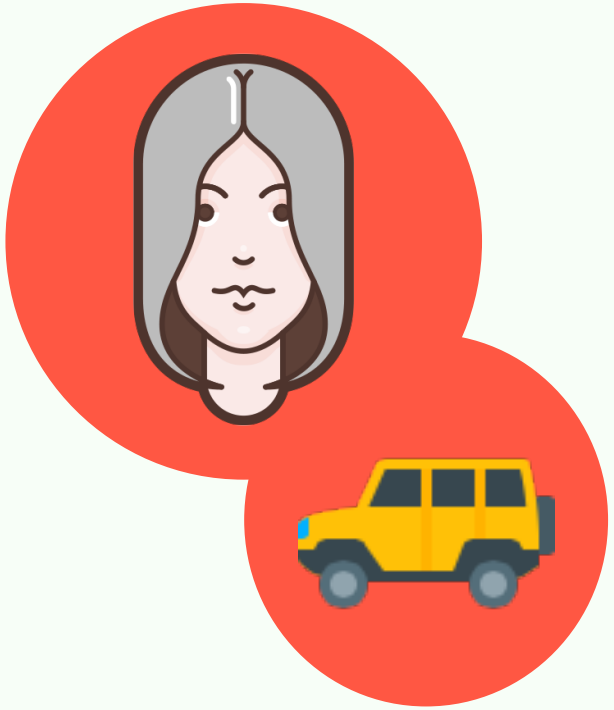
A second, sub-hypothesis is that **preferences for the options might differ among locations.**

PERSONA HYPHOTESIS

Intuition made us draft proto-personas based on our mental models. A few examples:

Bob, the part-timer
On his late 20s, he doesn't want to be held responsible for any liabilities but skimps a bit when protecting his own trusty 2005 Nissan Altima.

Mary, the sporty mom
She is very practical, but seeks maximum protection for her and her kids, and takes all the stops to do so – she has premium protection with every bell & whistle



ANALYSIS METHOD

Create customer segments to understand preferences and use logistic regression to predict likelihood or selecting given options.

For **Customer segmentation** we k means clustering to get 5 segments
We ran **22 Logistic regressions:**

Y variables: options

X variables: demographic data : region, log(local population), log(population density), urbanization level, log(avg income), age of the oldest people etc.

Car information: log(car age+1), car value

Insurance history: previous duration, whether they've chosen option C previously

KAGGLE SUBMISSION SUMMARY

Kaggle submission 1: demographics

Independent variables:

- Zip code demographics (see Appendix A)
- Region (Midwest, Northeast, South, West)
- LN(Car Age + 1)
- Age_oldest
- Duration_previous
- Homeowner (1 or 0)
- Car value (qualitative variable)
- Risk factor (qualitative)
- Married couple (1 or 0)
- C_previous (1, 2, 3, 4)

Range of Adjusted R2 (McFadden): 0.013 (B1) to 0.56 (C4)

Most important variables:

1. LN(Car Age + 1) - 18 significant coefficients
2. Risk factor - 17
3. C previous - 13
4. Region - 13

KAGGLE SUBMISSION SUMMARY

Kaggle submission 1: demographics

Independent variables:

- Zip code demographics (see Appendix A)
- Region (Midwest, Northeast, South, West)
- LN(Car Age + 1)
- Age_oldest
- Duration_previous
- Homeowner (1 or 0)
- Car value (qualitative variable)
- Risk factor (qualitative)
- Married couple (1 or 0)
- C_previous (1, 2, 3, 4)

Range of Adjusted R2 (McFadden): 0.013 (B1) to 0.56 (C4)

Most important variables:

1. LN(Car Age + 1) - 18 significant coefficients
2. Risk factor - 17
3. C previous - 13
4. Region - 13

Kaggle submission 2: quote view history + limited demographics

Independent variables:

- # times each option viewed prior to purchase
- Most recent options viewed prior to purchase
- Region
- LN(Car Age + 1)
- Age_oldest
- Duration_previous
- Homeowner (1 or 0)
- Married couple (1 or 0)
- C_previous (1, 2, 3, 4)

Range of Adjusted R2 (McFadden): 0.56 (G4) to 0.85 (A0)

Most important variables:

1. LN(Car Age + 1) - 9 significant coefficients
2. Behavioral data - almost always significant for the relevant category
3. Duration previous - 7
4. Age oldest - 4

KAGGLE SUBMISSION SUMMARY

Kaggle submission 1: demographics

Independent variables:

- Zip code demographics (see Appendix A)
- Region (Midwest, Northeast, South, West)
- LN(Car Age + 1)
- Age_oldest
- Duration_previous
- Homeowner (1 or 0)
- Car value (qualitative variable)
- Risk factor (qualitative)
- Married couple (1 or 0)
- C_previous (1, 2, 3, 4)

Range of Adjusted R2 (McFadden): 0.013 (B1) to 0.56 (C4)

Most important variables:

1. LN(Car Age + 1) - 18 significant coefficients
2. Risk factor - 17
3. C previous - 13
4. Region - 13

Kaggle submission 2: quote view history + limited demographics

Independent variables:

- # times each option viewed prior to purchase
- Most recent options viewed prior to purchase
- Region
- LN(Car Age + 1)
- Age_oldest
- Duration_previous
- Homeowner (1 or 0)
- Married couple (1 or 0)
- C_previous (1, 2, 3, 4)

Range of Adjusted R2 (McFadden): 0.56 (G4) to 0.85 (A0)

Most important variables:

1. LN(Car Age + 1) - 9 significant coefficients
2. Behavioral data - almost always significant for the relevant category
3. Duration previous - 7
4. Age oldest - 4

Kaggle submission 3: naive - most recent viewed

Set purchase prediction equal to whatever product was most recently viewed

KAGGLE SUBMISSION SUMMARY

Kaggle submission 1: demographics

Independent variables:

- Zip code demographics (see Appendix A)
- Region (Midwest, Northeast, South, West)
- LN(Car Age + 1)
- Age_oldest
- Duration_previous
- Homeowner (1 or 0)
- Car value (qualitative variable)
- Risk factor (qualitative)
- Married couple (1 or 0)
- C_previous (1, 2, 3, 4)

Range of Adjusted R2 (McFadden): 0.013 (B1) to 0.56 (C4)

Most important variables:

1. LN(Car Age + 1) - 18 significant coefficients
2. Risk factor - 17
3. C previous - 13
4. Region - 13

Kaggle submission 2: quote view history + limited demographics

Independent variables:

- # times each option viewed prior to purchase
- Most recent options viewed prior to purchase
- Region
- LN(Car Age + 1)
- Age_oldest
- Duration_previous
- Homeowner (1 or 0)
- Married couple (1 or 0)
- C_previous (1, 2, 3, 4)

Range of Adjusted R2 (McFadden): 0.56 (G4) to 0.85 (A0)

Most important variables:

1. LN(Car Age + 1) - 9 significant coefficients
2. Behavioral data - almost always significant for the relevant category
3. Duration previous - 7
4. Age oldest - 4

Kaggle submission 3: naive - most recent viewed

Set purchase prediction equal to whatever product was most recently viewed

Kaggle submission 4: naive - first viewed (default option)

Set purchase prediction equal to the first "default" option presented

KAGGLE RESULTS

600 submissions (from place 400-1000) were naive predictions, all with the same exact score of 53.269%Current winning score: 53.743% - only 0.5 pps above Regression 2!

Regression 1 (full demographic): 1.587% (place 1,466)

Regression 2 (quote view history + limited demographic): 53.243% (place 1,112)

Regression 3 (last viewed): 53.269% (place 990)

Regression 4 (first viewed): 16.025% (place 1,396)

FINDINGS

Data quality played a huge role in the usefulness of the outcome of the regression.

“Default” option is very important and determines purchase

Focus on algorithms for MAX profits

Poor info = poor hit rate

Previous insurance info helps for better predictions

LIMITATION OF DATA & NEXT STEPS

Data provided is difficult to interpret because of obtuse labeling, and is missing metrics that impact customer decisions.

Recommendation on data collection for further analysis:

1. Include non-customer data to get full picture (not on file)
2. Clearly labeled data for better recs
3. The breakdown the risk factor:

Driver's information

- How long has he/she been driving
- No. of incidents before
- Income level and credit record

Car information:

- Make/Model/Year
- Usage of the car (commercial/fun/commute)
- Mileage
- Ownership Status(rent/owned/leased)

4. Have "user id" under "customer id"
5. Include chat/call information to justify whether the communication from Allstate influence customer's decision
6. Information about discounts/bundle choices offered to customers

Q&A

APENDIX A

Data Structure: 5000 observations, 48 variables.

I. Purchase information

•Customer ID
•Browsed options and purchased options (including time and day)
•Number of browsed options

II. Demographic information

•Marital status
•Group information (size, age of the oldest/youngest person)
•Homeowner
•Location (including states, population density, household income, etc.)

III. Car status information

•Car age
•Car Value

IV. Historical record information

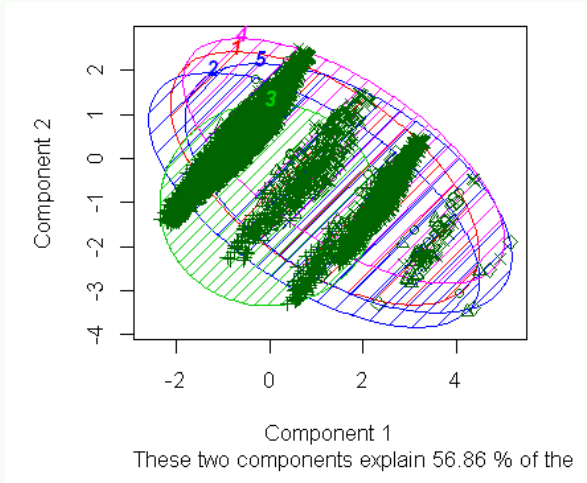
•Previous purchase record (including coverage duration / whether they've chosen a specific option before)

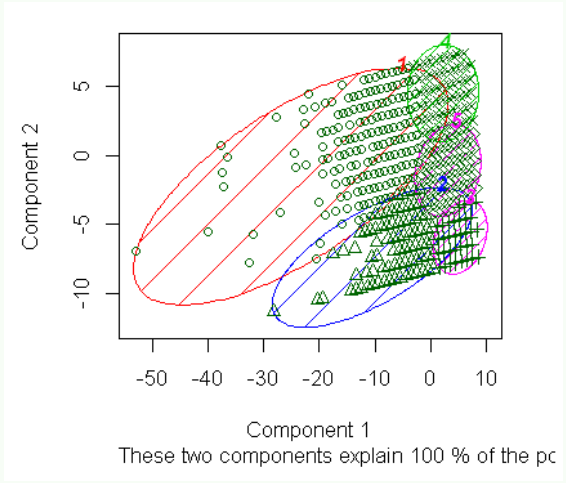
APENDIX B

We tried different approaches for the segmentation (with different number of segments and metrics used), but the results failed to help us generate more insight. There are many overlaps among segments. Here below are examples

	Group.1	car_age	age_oldest	duration_previous	homeowner	married_couple	group_size
1	1	15.37622	57.29435	6.97076	0.6744639	0.2475634	1.309942
2	2	7.268849	42.09028	6.702381	0.6259921	0.2152778	1.268849
3	3	7.416892	26.33378	4.756081	0.2945946	0.09662162	1.103378
4	4	7.365854	71.53297	8.365854	0.831075	0.3550136	1.35682
5	5	4.358556	57.18743	8.065192	0.7613504	0.2596042	1.335274

	Group.1	car_age	duration_previous
1	1	14.57293	3.97188
2	2	12.21113	13.45336
3	3	3.357372	14.30929
4	4	4.667075	2.69344
5	5	4.504673	7.786085





APPENDIX C

Zip-code-level data appended for regression 1:

1.Log population
2.Log population density
3.Rurality index (1 = most urban; 9 = most rural)
4.Urban influence index (1 = most adjacent to urban area; 12 = least adjacent to urban area)
5.Urban commuter score (% of commuters that drive to urban areas of varying size)
6.Log mean household income

All data derived from http://www.ers.usda.gov

#s 1-5 are from 2003 (unable to find more comprehensive recent data)

6 is from 2010

APENDIX D

Do visitors “game” the system to reduce cost, or do they change purchase options for other reasons?

No: on average, customers actually increase quote cost by 0.64% (or about $2.29) from their initial default option

This is consistent across different types of customers - most demographic variables have extremely little impact on the % change in quote cost (see table to the right)