

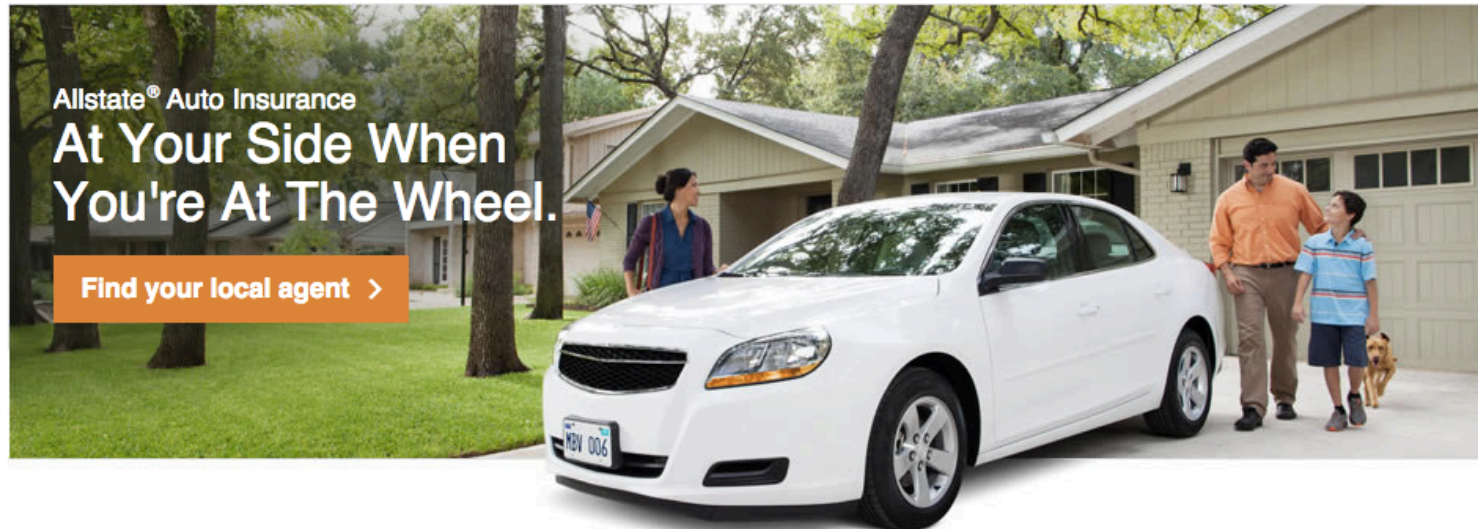
Kaggle – AllState Prediction Challenge


Priyanka Talwar

December 17, 2014

GA DAT10

All State Insurance

[GET A QUOTE](#)[Insurance & More](#)[Claims](#)[Support](#)[Tools & Resources](#)[My Account Login](#)

Auto 

Coverage

Policy Features

Auto Discounts

Teen Drivers

[Home](#) > [Auto Insurance](#)

[+ Share](#)

542

Text Size

[A](#) [A](#) [A](#)

Enjoy Quality Car Insurance Coverage.

Allstate car insurance gives you quality protection at a great price. Every policy comes with the support of a knowledgeable and friendly Allstate agent. With personal attention and service, you'll have the information you need to choose the car insurance coverage that makes

Introduction & Competition Goal

- Goal : Predict purchased policy options for a customer , given customer transaction history
- Details : 7 insurance options; 2 - 4 possible values
- A customer will receive number of quotes with different coverage options; a customer may purchase a product that was not viewed

Example of coverage options

Option name	Possible values
A	0, 1, 2
B	0, 1
C	1, 2, 3, 4
D	1, 2, 3
E	0, 1
F	0, 1, 2, 3
G	1, 2, 3, 4

Example of coverage options

Customer_ID	shopping_pt	record_type	A	B	C	D	E	F	G
10000000	1	0	1	0	2	2	1	2	2
10000000	2	0	1	0	2	2	1	2	1
10000000	3	0	1	0	2	2	1	2	1
10000000	4	0	1	0	2	2	1	2	1
10000000	5	0	1	0	2	2	1	2	1
10000000	6	0	1	0	2	2	1	2	1
10000000	7	0	1	0	2	2	1	2	1
10000000	8	0	1	0	2	2	1	2	1
10000000	9	1	1	0	2	2	1	2	1
10000005	1	0	1	1	3	3	1	0	2
10000005	2	0	1	1	3	3	1	0	2
10000005	3	0	1	1	3	3	1	0	2
10000005	4	0	0	0	3	2	0	0	2
10000005	5	0	0	0	3	2	0	0	2
10000005	6	1	0	0	3	2	0	0	2
10000007	1	0	0	0	2	3	0	0	3
10000007	2	0	0	0	2	3	0	0	1
10000007	3	0	0	0	1	1	0	0	1
10000007	4	0	0	0	1	1	0	0	1
10000007	5	0	0	0	1	1	0	0	1
10000007	6	0	0	0	2	2	0	0	1
10000007	7	0	0	0	2	2	0	0	1
10000007	8	1	0	0	1	2	0	0	1

DataSet

Training set:

- 97,009 customers that included quote history along with purchase

Test Set:

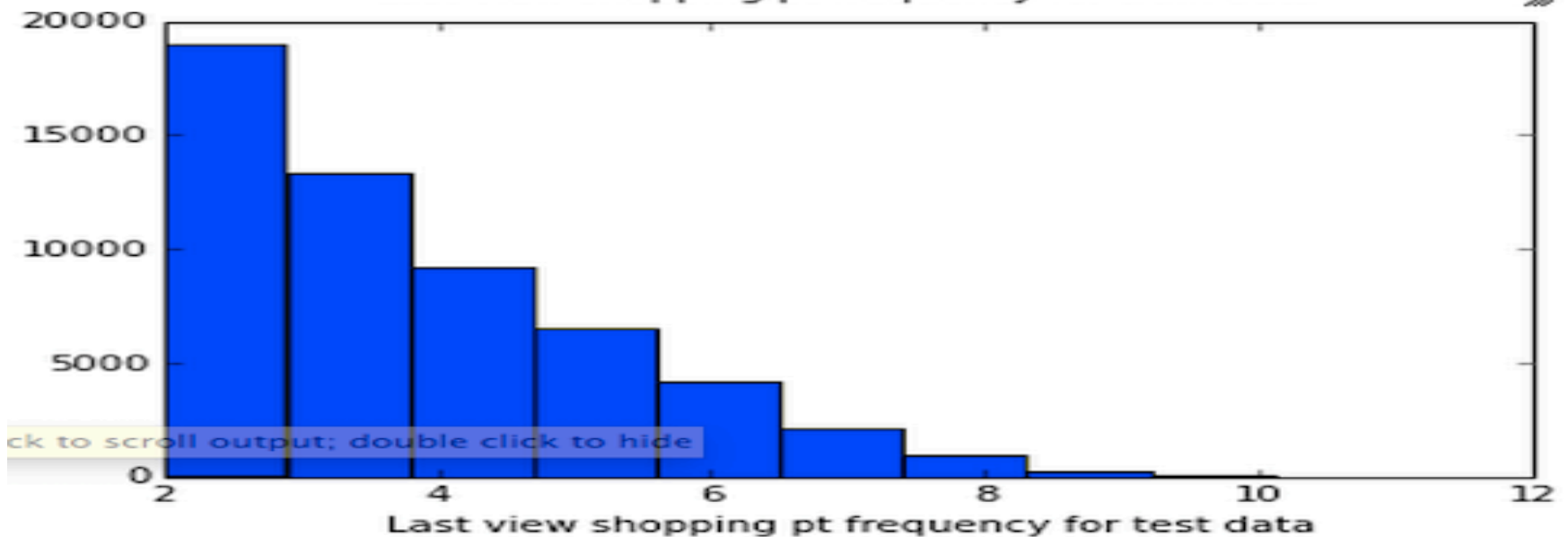
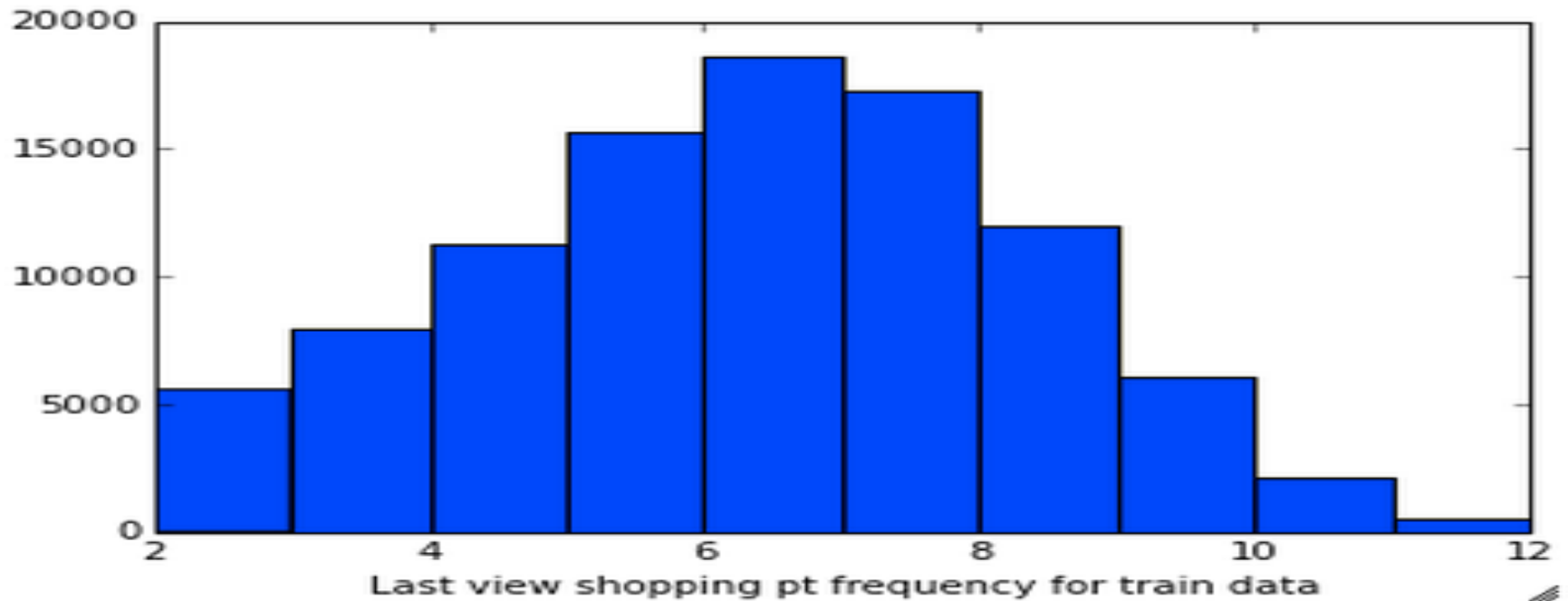
- 55,716 customers with the truncated quote history

Variable Name	Description	Type
customer_ID	Customer's identifier	identifier
shopping_pt	Plans presented to a customer - Unique identifier	categorical
record_type	Record type can either be 0 = shopping point or 1 = purchase point	categorical
Day	Day of the week (0 - 6, 0 = Monday)	categorical
Time	Time of day (HH:MM)	
State	State where shopping point occurred	categorical
Location	Location ID where shopping point occurred	categorical
group_size	Number of people covered under the given policy. Ranges between 0 - 4	continuous
Homeowner	Whether the customer owns a home or not (0 = no, 1 = yes)	categorical
car_age	How old is the customer's car	continuous
car_value	What was the value of the customer's car when new	categorical
risk_factor	How risky is the customer. Ranges between 1 - 4	categorical
age_oldest	Age of the oldest person in customer's group	continuous
age_youngest	Age of the youngest person in customer's group	continuous
married_couple	Is anyone in the customer group is married. (0 = no, 1 = yes)	categorical
C_previous	What the customer formerly had or currently has for product option C (0 = nothing, 1, 2, 3,4)	categorical
duration_previous	For how long (in years) the customer was covered by their previous issuer	continuous
A,B,C,D,E,F,G	Different coverage options	response
cost	Cost associated with the quoted coverage options	continuous

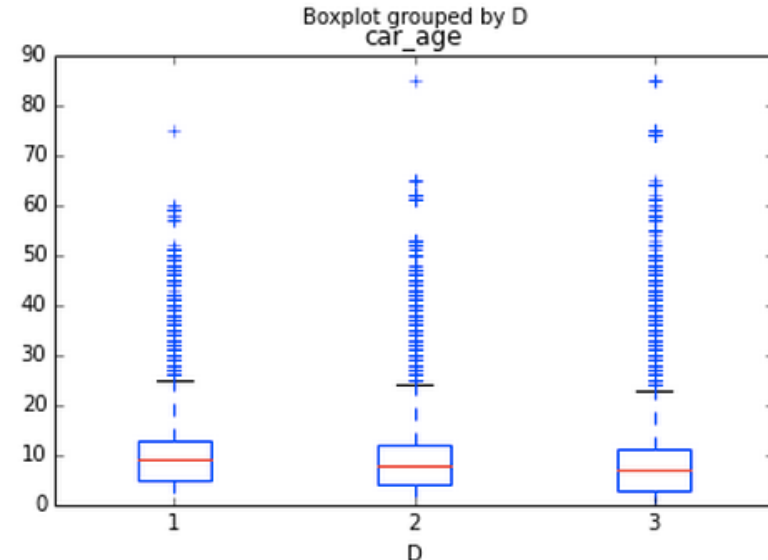
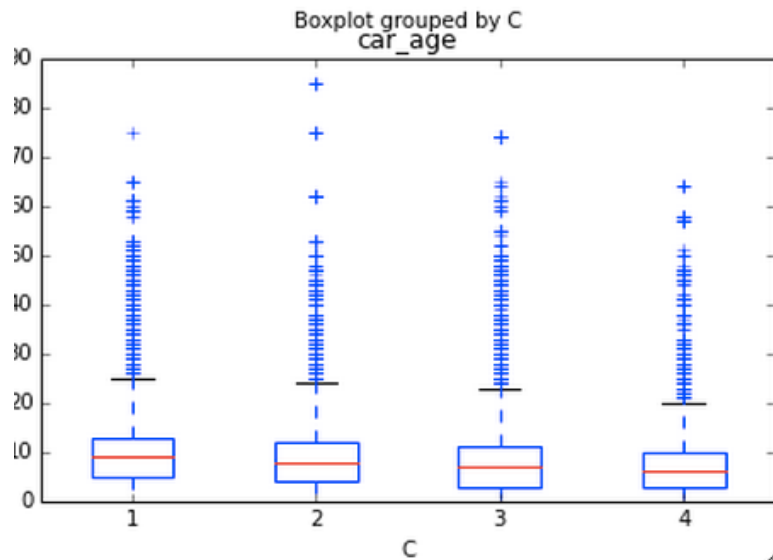
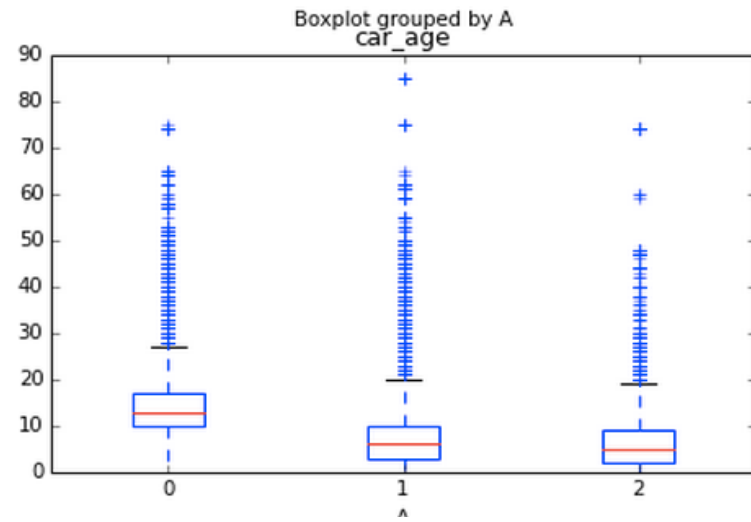
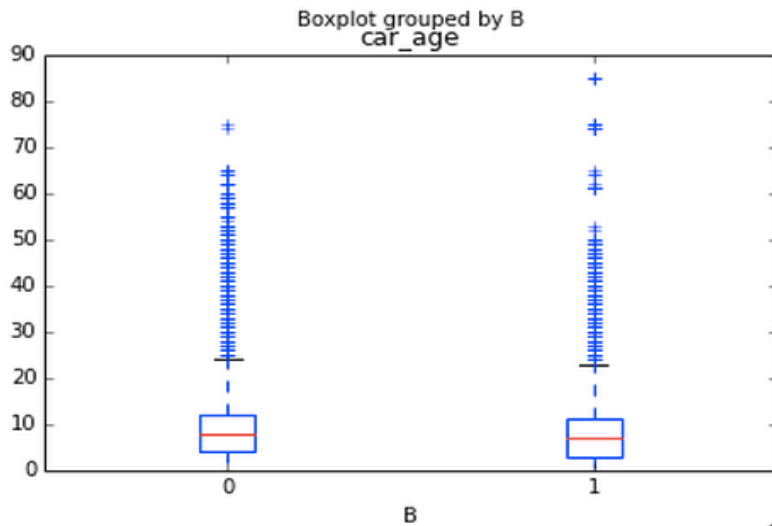
Why is this interesting / challenging?

- Dataset included feature set in both rows and columns for each customer; the cols represented feature set of a customer and each customer had multiple rows
- For a prediction to be correct, you need to get all 'seven' options right
- Data was not thoroughly explained

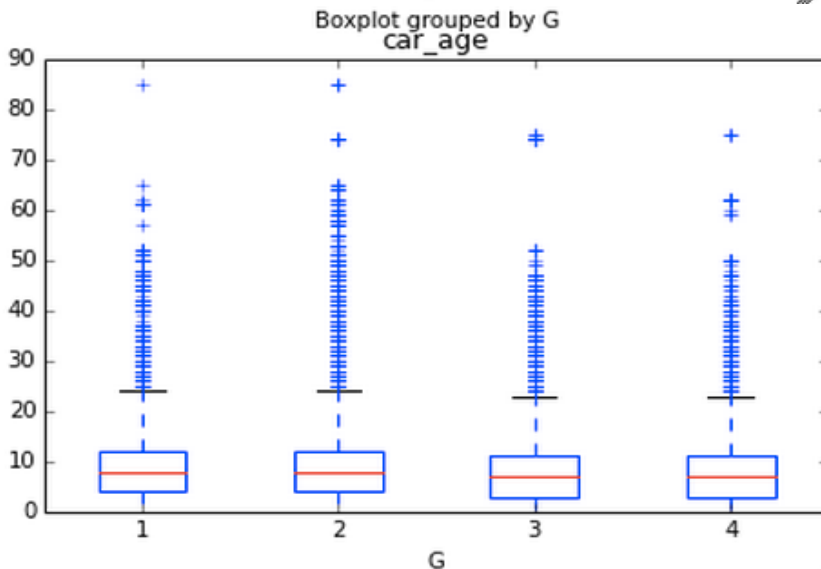
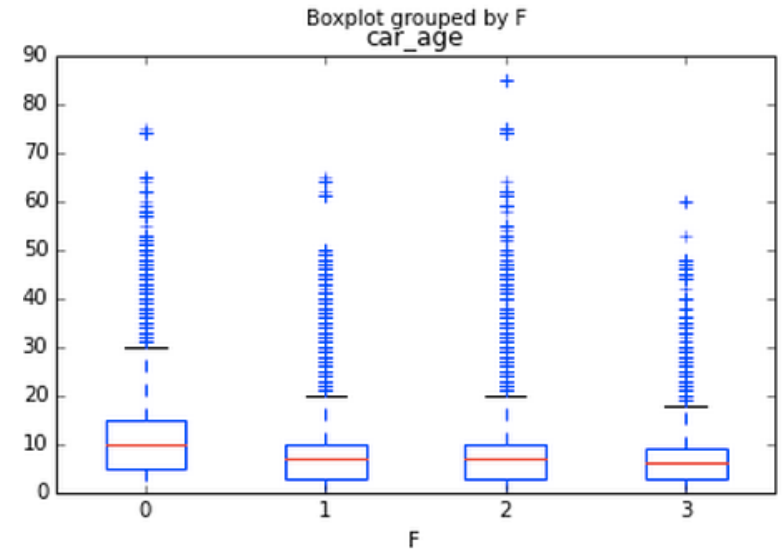
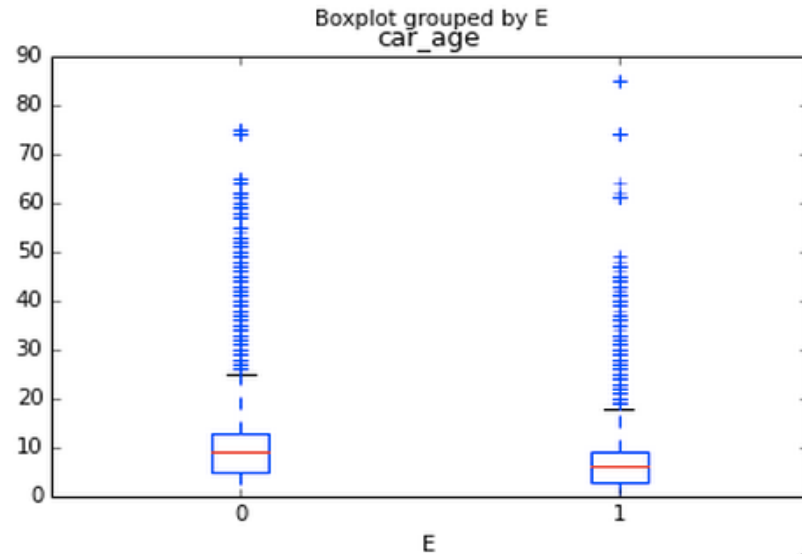
Significantly Truncated Test



Exploratory Analysis: Individual Options Vs features



Individual Options Vs features : No results!!



Feature Engineering and What would I as a customer do?

- Started with a basic approach of predicting based on last view
- Random Forest Model used


	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimension reduction</i>	<i>clustering</i>

Submission Entry

	customer_ID	plan
1	10000001	2113132
3	10000002	2023122
6	10000003	1021022
15	10000004	2011123
17	10000006	0011001
23	10000008	0013023
27	10000009	0112002
29	10000010	1012012
31	10000011	1033022
35	10000012	2023033
37	10000015	1133102
39	10000017	1133111
42	10000018	1133111
46	10000020	0123022
49	10000021	0113012
54	10000022	1143113
56	10000024	0011002
59	10000029	0033002
64	10000030	0011001
66	10000035	1113123
68	10000044	2133113
71	10000045	1133112
73	10000046	1033021
75	10000050	1022002
79	10000051	1121122
82	10000053	1143022

Results

- **Cross Validation Score** : .68045233433819
- Past Deadline so no \$\$, but here were the Kaggle results; Accuracy of .52979

1186	↓186	cts309	0.52992	10	Fri, 09 May 2014 02:47:59 (-0.8h)
1187	↑16	Panacea	0.52982	1	Thu, 06 Mar 2014 13:08:32
-		priyanka talwar	0.52979	-	Wed, 17 Dec 2014 18:06:27 Post-Deadline
Post-Deadline Entry If you would have submitted this entry during the competition, you would have been around here on the leaderboard.					
1188	↑13	Forio 	0.52977	10	Tue, 29 Apr 2014 16:01:49 (-19.2h)

Next Steps

- More feature engineering: Use the frequency percent of times an option appears in a plan
- Stability : How likely a customer switches from one option to other
- Suggestion : Use weighted avg on plans
- More Models : SVM or Naive Bayes