Machine Learning project

Team members: Jack, Harry & Abhishek

Homesite problem: Predicting Quote conversion

- Homesite sells Home-insurance to home buyers in United States
- Insurance quotes are offered to customers based on several factors

What Homesite knows

- Customer's geographical, personal, Financial Information & HomeOwnership details

- Quote for every customer

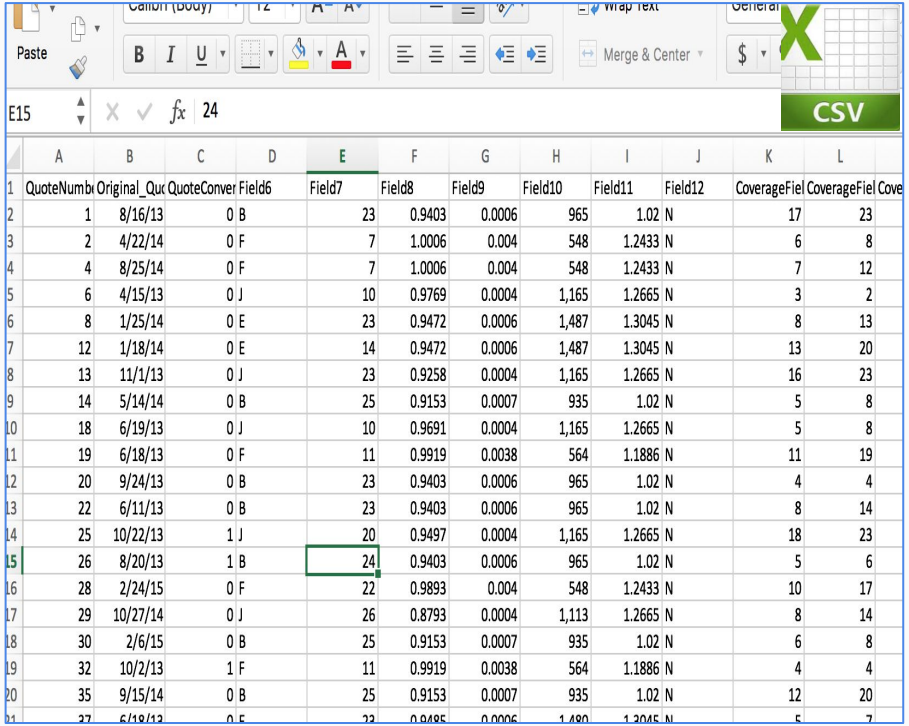
What Homesite Doesn't know:

Customer's **likelihood** of buying that Insurance contract

Enter

kaggle

Data shared: Training & Test



E15 | 24

	A	B	C	D	E	F	G	H	I	J	K	L	
1	QuoteNumb	Original_Qu	QuoteConver	Field6	Field7	Field8	Field9	Field10	Field11	Field12	CoverageFiel	CoverageFiel	Cove
2	1	8/16/13	0	B	23	0.9403	0.0006	965	1.02	N	17	23	
3	2	4/22/14	0	F	7	1.0006	0.004	548	1.2433	N	6	8	
4	4	8/25/14	0	F	7	1.0006	0.004	548	1.2433	N	7	12	
5	6	4/15/13	0	J	10	0.9769	0.0004	1,165	1.2665	N	3	2	
6	8	1/25/14	0	E	23	0.9472	0.0006	1,487	1.3045	N	8	13	
7	12	1/18/14	0	E	14	0.9472	0.0006	1,487	1.3045	N	13	20	
8	13	11/1/13	0	J	23	0.9258	0.0004	1,165	1.2665	N	16	23	
9	14	5/14/14	0	B	25	0.9153	0.0007	935	1.02	N	5	8	
10	18	6/19/13	0	J	10	0.9691	0.0004	1,165	1.2665	N	5	8	
11	19	6/18/13	0	F	11	0.9919	0.0038	564	1.1886	N	11	19	
12	20	9/24/13	0	B	23	0.9403	0.0006	965	1.02	N	4	4	
13	22	6/11/13	0	B	23	0.9403	0.0006	965	1.02	N	8	14	
14	25	10/22/13	1	J	20	0.9497	0.0004	1,165	1.2665	N	18	23	
15	26	8/20/13	1	B	24	0.9403	0.0006	965	1.02	N	5	6	
16	28	2/24/15	0	F	22	0.9893	0.004	548	1.2433	N	10	17	
17	29	10/27/14	0	J	26	0.8793	0.0004	1,113	1.2665	N	8	14	
18	30	2/6/15	0	B	25	0.9153	0.0007	935	1.02	N	6	8	
19	32	10/2/13	1	F	11	0.9919	0.0038	564	1.1886	N	4	4	
20	35	9/15/14	0	B	25	0.9153	0.0007	935	1.02	N	12	20	
21	37	6/18/13	0	F	23	0.9485	0.0006	1,480	1.3045	N	5	7	

Task:

Binary classification

Training set:

261k rows, 298 predictors, 1 Binary response

Test set:

200k rows, 298 columns

Predictors:

Customer Activity, Geography, Personal, property & Coverage

Response:

Customer Conversion

What's good about Homesite data:

- 296 variables don't have NA's or bad data entry points
- Not many Levels in Nominal variables
- Plenty of binary variables
- Plenty of ordinal variables
- No unbalanced variables
- No missing values
- No Textual columns

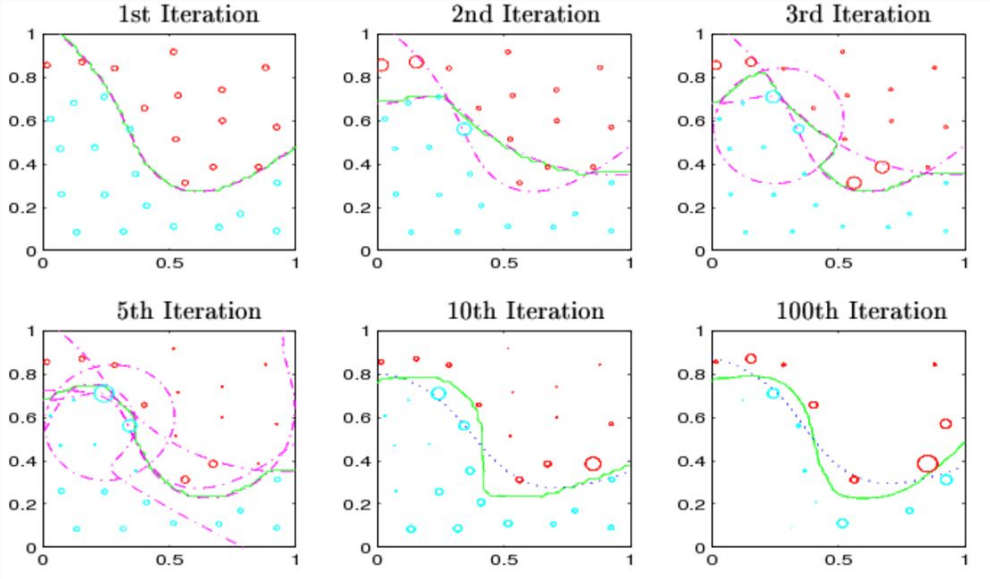
Data cleaning steps

1. Removing Constants
2. Removing Identifier rows
3. Synthesizing Date column
4. Treating NA variables
5. Treating bad levels (-1)
6. Treating false categorical
7. Categorical to dummy

Gradient Boosting (Iterative corrections)

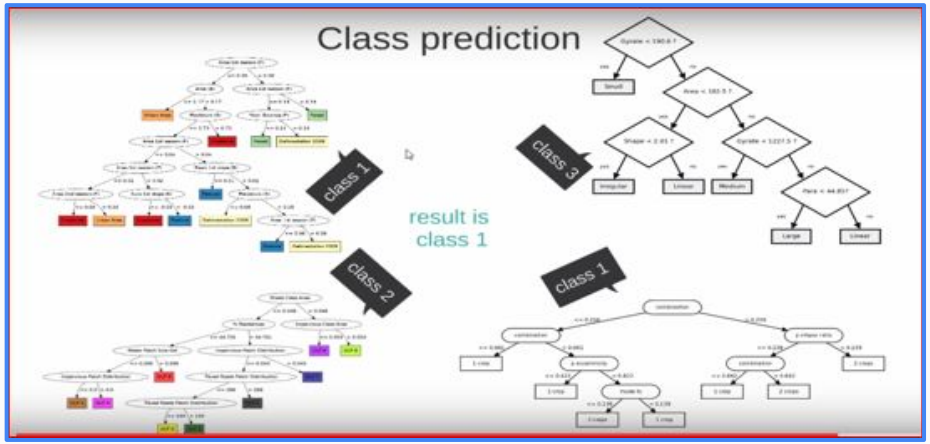
- Learning from past mistakes
- Could get nearly 0 training error
- Weighted scoring of multiple trees

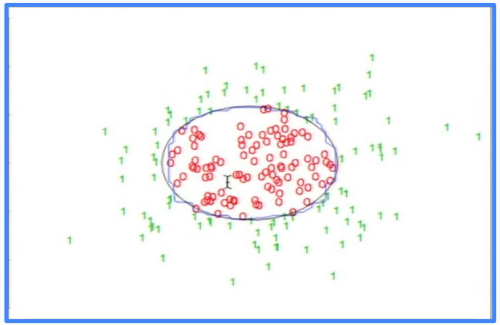
- Hard to tune, as there are too many parameters to adjust
- Often overfit and hard to decide the stopping point



Random Forests (Majority wins)

- Handles missing data
- Handles redundancy easily
- Reduces variations in results
- Produces Out of Bag error rate
- Produces De-correlated trees
- Random subspace & split





Bias sometimes Increases as Trees are shallower

Gradient Boosting + Random Forest

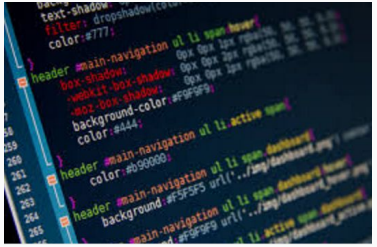
- Handles missing data
- Handles redundancy easily
- Reduces variations in results
- Produces Out of Bag error rate
- Produces De-correlated trees
- Random subspace & split
- Does not overfit
- Little bias, due to correction
- Easy to tune

Quite slow and Computationally expensive. Optimizing these constraints could be an excellent area for research

Calculating AUC

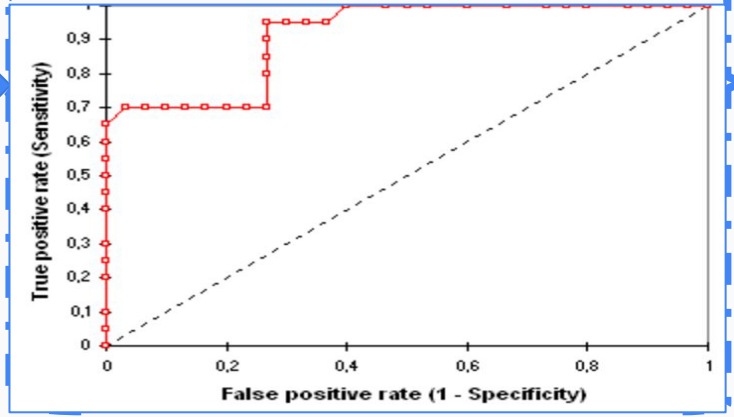
ID	True class	Predicted probability
1	1	.8612
2	0	.2134
3	0	.1791
4	0	.1134
5	1	.7898
6	0	.0612

kaggle



AUC

- Randomly decide a threshold
- Calculate True Positive Rate (y) & False Positive Rate (x)
- Based on (x,y) plot the point
- Repeat steps for each value of threshold [0,1]
- We now have a curve and we call it **ROC**
- **Area under this curve becomes AUC**



War for the highest AUC

Homesite HOME INSURANCE

$20,000 • 627 teams

Homesite Quote Conversion

Merger and 1st Submission Deadline

Mon 9 Nov 2015 — Mon 8 Feb 2016 (2 months to go)

Dashboard

Public Leaderboard - Homesite Quote Conversion

Our Score
AUC = .9645

This leaderboard is calculated on approximately 30% of the test data.
The final results will be based on the other 70%, so the final standings may be different.

#	Δ6d	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	clustifier *	0.96920	73	Tue, 01 Dec 2015 07:21:47 (-2.2d)
2	—	Dmitry Larko *	0.96903	18	Wed, 02 Dec 2015 15:02:03
3	—	Victor *	0.96895	21	Tue, 01 Dec 2015 12:52:38
4	↑4	Ivanhoe	0.96860	69	Fri, 04 Dec 2015 03:56:01 (-0.1h)
5	↑51	MrOoijer	0.96856	81	Fri, 04 Dec 2015 00:20:01 (-12h)
6	↑40	Daniel FG	0.96855	6	Sun, 29 Nov 2015 22:38:52
7	↑2	PierreNowak	0.96852	45	Wed, 02 Dec 2015 11:27:37

What we have already employed

- Categorical to Continuous conversion
- Continuous to Ordinal conversion
- Variable bucketing
- SVM / Logistic Regression
- Random Forest/ Trees
- Lasso / Ridge / Elastic Net
- Gradient Boosting
- Multicollinearity elimination
- Outlier treatment
- K-Fold Cross validation

What we look forward to use

- Imputation for NA's
- Model tuning
- Variable transformation
- Most importantly, **Your Suggestions**

THANK YOU