


How do machine learning competitions work?

We take a dataset with a target variable – something we're trying to predict

Predicting the sale  
price of a home



SalePrice	SquareFeet	Type	LotAcres	Beds	Baths
\$88k	719	HOME	1.64	1	1
\$164k	2017	APT		3	2
\$72k	697	APT		1	1
\$85k	948	HOME	1.02	2	3
\$271k	3375	APT		3	4
\$482k	3968	APT		4	4
\$88k	790	APT		1	2
\$128k	1341	HOME	0.66	3	3
\$235k	2379	APT		3	3
\$309k	2495	HOME	0.21	3	4
\$163k	1356	APT		1	1
\$375k	3361	HOME	1.64	3	4
\$98k	1060	HOME	0.05	1	1
\$50k	582	HOME	0.61	1	1
\$145k	1640	APT		2	3
\$394k	3546	HOME	0.4	4	4
\$82k	903	APT		2	2
\$105k	1096	HOME	0.04	3	4
\$129k	1280	HOME	0.15	2	2
\$106k	1139	APT		1	1

Split the data into two sets, a training set and a test set

SalePrice	SquareFeet	Type	LotAcres	Beds	Baths	
\$88k	719	HOME	1.64	1	1	Training
\$164k	2017	APT		3	2	
\$72k	697	APT		1	1	
\$85k	948	HOME	1.02	2	3	
\$271k	3375	APT		3	4	
\$482k	3968	APT		4	4	
\$88k	790	APT		1	2	
\$128k	1341	HOME	0.66	3	3	
\$235k	2379	APT		3	3	
\$309k	2495	HOME	0.21	3	4	
\$163k	1356	APT		1	1	
\$375k	3361	HOME	1.64	3	4	
\$98k	1060	HOME	0.05	1	1	
\$50k	582	HOME	0.61	1	1	Test
\$145k	1640	APT		2	3	
\$394k	3546	HOME	0.4	4	4	
\$82k	903	APT		2	2	
\$105k	1096	HOME	0.04	3	4	
\$129k	1280	HOME	0.15	2	2	
\$106k	1139	APT		1	1	

Solution  
“Ground Truth”

Our community gets everything but the solution on the test set

SalePrice	SquareFeet	Type	LotAcres	Beds	Baths	
\$88k	719	HOME	1.64	1	1	Training
\$164k	2017	APT		3	2	
\$72k	697	APT		1	1	
\$85k	948	HOME	1.02	2	3	
\$271k	3375	APT		3	4	
\$482k	3968	APT		4	4	
\$88k	790	APT		1	2	
\$128k	1341	HOME	0.66	3	3	
\$235k	2379	APT		3	3	
\$309k	2495	HOME	0.21	3	4	
\$163k	1356	APT		1	1	
\$375k	3361	HOME	1.64	3	4	
\$98k	1060	HOME	0.05	1	1	
???	582	HOME	0.61	1	1	Test
???	1640	APT		2	3	
???	3546	HOME	0.4	4	4	
???	903	APT		2	2	
???	1096	HOME	0.04	3	4	
???	1280	HOME	0.15	2	2	
???	1139	APT		1	1	

Solution  
“Ground Truth”

Competition participants use the training set to learn the relation between the data and the target



Competition participants apply their models to make predictions on the test set

	SalePrice	SquareFeet	Type	LotAcres	Beds	Baths	
	\$88k	719	HOME	1.64	1	1	Training
	\$164k	2017	APT		3	2	
	\$72k	697	APT		1	1	
	\$85k	948	HOME	1.02	2	3	
	\$271k	3375	APT		3	4	
	\$482k	3968	APT		4	4	
	\$88k	790	APT		1	2	
	\$128k	1341	HOME	0.66	3	3	
	\$235k	2379	APT		3	3	
	\$309k	2495	HOME	0.21	3	4	
	\$163k	1356	APT		1	1	
	\$375k	3361	HOME	1.64	3	4	
Predicted	\$98k	1060	HOME	0.05	1	1	Test
\$41k	???	582	HOME	0.61	1	1	
\$165k	???	1640	APT		2	3	
\$280k	???	3546	HOME	0.4	4	4	
\$76k	???	903	APT		2	2	
\$128k	???	1096	HOME	0.04	3	4	
\$115k	???	1280	HOME	0.15	2	2	
\$94k	???	1139	APT		1	1	

Submission

Kaggle compares the submission to the ground truth

		SalePrice	SquareFeet	Type	LotAcres	Beds	Baths	
		\$88k	719	HOME	1.64	1	1	Training
		\$164k	2017	APT		3	2	
		\$72k	697	APT		1	1	
		\$85k	948	HOME	1.02	2	3	
		\$271k	3375	APT		3	4	
		\$482k	3968	APT		4	4	
		\$88k	790	APT		1	2	
		\$128k	1341	HOME	0.66	3	3	
		\$235k	2379	APT		3	3	
		\$309k	2495	HOME	0.21	3	4	
		\$163k	1356	APT		1	1	
		\$375k	3361	HOME	1.64	3	4	
Delta	Predicted	\$98k	1060	HOME	0.05	1	1	Test
-\$9k	\$41k	\$50k	582	HOME	0.61	1	1	
\$20k	\$165k	\$145k	1640	APT		2	3	
-\$14k	\$380k	\$394k	3546	HOME	0.4	4	4	
-\$6k	\$76k	\$82k	903	APT		2	2	
\$13k	\$128k	\$105k	1096	HOME	0.04	3	4	
-\$14k	\$115k	\$129k	1280	HOME	0.15	2	2	
-\$12k	\$94k	\$106k	1139	APT		1	1	
		Submission						

Kaggle calculates two scores, one for the public leaderboard and one for the private leaderboard

		SalePrice	SquareFeet	Type	LotAcres	Beds	Baths	MeanError	
		\$88k	719	HOME	1.64	1	1	\$14k	Public Leaderboard
		\$164k	2017	APT		3	2		
		\$72k	697	APT		1	1		
		\$85k	948	HOME	1.02	2	3		
		\$271k	3375	APT		3	4		
		\$482k	3968	APT		4	4		
		\$88k	790	APT		1	2		
		\$128k	1341	HOME	0.66	3	3		
		\$235k	2379	APT		3	3		
		\$309k	2495	HOME	0.21	3	4		
		\$163k	1356	APT		1	1		
		\$375k	3361	HOME	1.64	3	4		
		\$98k	1060	HOME	0.05	1	1		
		\$50k	582	HOME	0.61	1	1		
		\$145k	1640	APT		2	3	\$15k	Private Leaderboard
		\$394k	3546	HOME	0.4	4	4		
		\$82k	903	APT		2	2		
		\$105k	1096	HOME	0.04	3	4		
		\$129k	1280	HOME	0.15	2	2		
		\$106k	1139	APT		1	1		
Delta	Predicted								
-\$9k	\$41k								
\$20k	\$165k								
-\$14k	\$380k								
-\$6k	\$76k								
\$13k	\$128k								
-\$14k	\$115k								
-\$12k	\$94k								


Submission

Training

Test



The participant immediately sees their public score on the public leaderboard

381	new	Archit Karandikar	0.50000	1	Thu, 30 Oct 2014 03:14:54
382	new	Martin	0.50000	3	Thu, 30 Oct 2014 09:27:46 (-0.1h)
383	new	<b>Ben Hamner</b>	<b>0.50000</b>	<b>1</b>	<b>Thu, 30 Oct 2014 09:46:58</b>
<b>Your Best Entry</b> Congratulations on making your first submission!					
 Tweet this!					
384	↓68	antonio zippo	0.49927	7	Wed, 24 Sep 2014 13:27:21 (-25.4h)
385	↓68	Konstantina	0.49873	3	Mon, 01 Sep 2014 20:19:46 (-5.9d)
386	↓68	pantheon	0.49809	3	Thu, 04 Sep 2014 00:13:07 (-0.3h)
387	new	Nino Paul Batanay	0.49591	2	Wed, 29 Oct 2014 09:57:38 (-1.5h)
388	↓66	blackout	0.49082	1	Thu, 09 Oct 2014 02:52:45