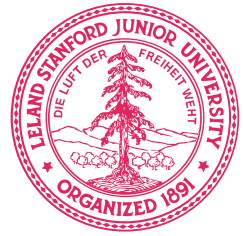
A DATA MINING COURSE USING A KAGGLE COMPETITION

Susan Holmes, Nelson Ray ©

April 24, 2011



How to get started in datamining?

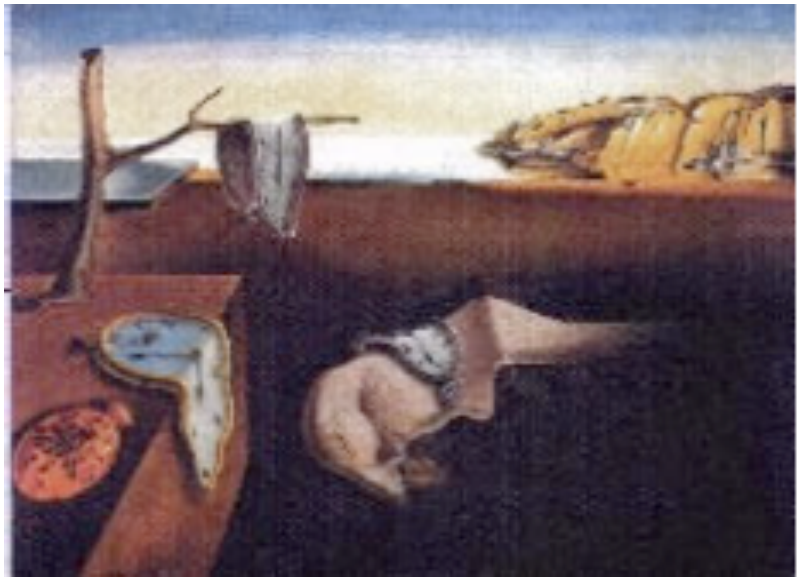
If you have plenty of time:

- Learn statistics
- Learn R
- Learn linear algebra
- Learn functional analysis (kernels,..)
- Learn data base management.

How to get started in datamining?

If you have plenty of time:

- Learn statistics
- Learn R
- Learn linear algebra
- Learn functional analysis (kernels,..)
- Learn data base management.



BEST NEURAL NETWORK AVAILABLE

You have it:

BEST NEURAL NETWORK AVAILABLE

You have it:

BEST NEURAL NETWORK AVAILABLE

You have it:

DATAMINING IN ONE COURSE

We need a hands on approach.

How to get really started ?

HOW TO GET REALLY STARTED ?

AVOID BLACK BOXES

LINEAR ALGEBRA: THE IGON VALUE PROBLEM

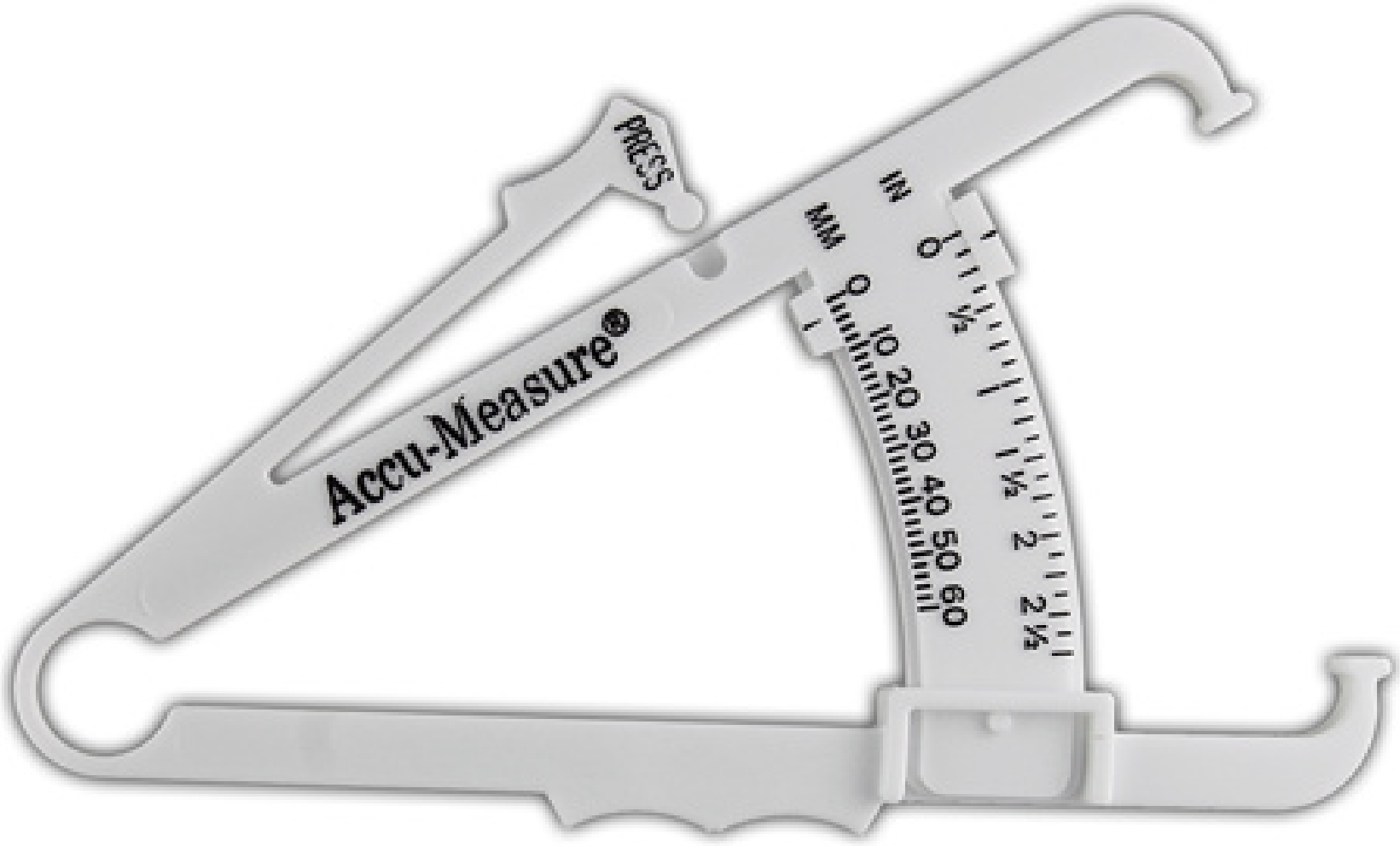
Malcolm Gladwell made an unfortunate error in
`What the dog saw' (in his spelling of eigenvalue).
Igon Value Problem is one of dilettantism.
We have to understand the methods and not treat them as black boxes.

START WITH DATA YOU KNOW WELL

- Because you invented it (simulation studies).
- It is your domain of expertise.
- You have already studied it with other tools.

START IN THE SUPERVISED CONTEXT

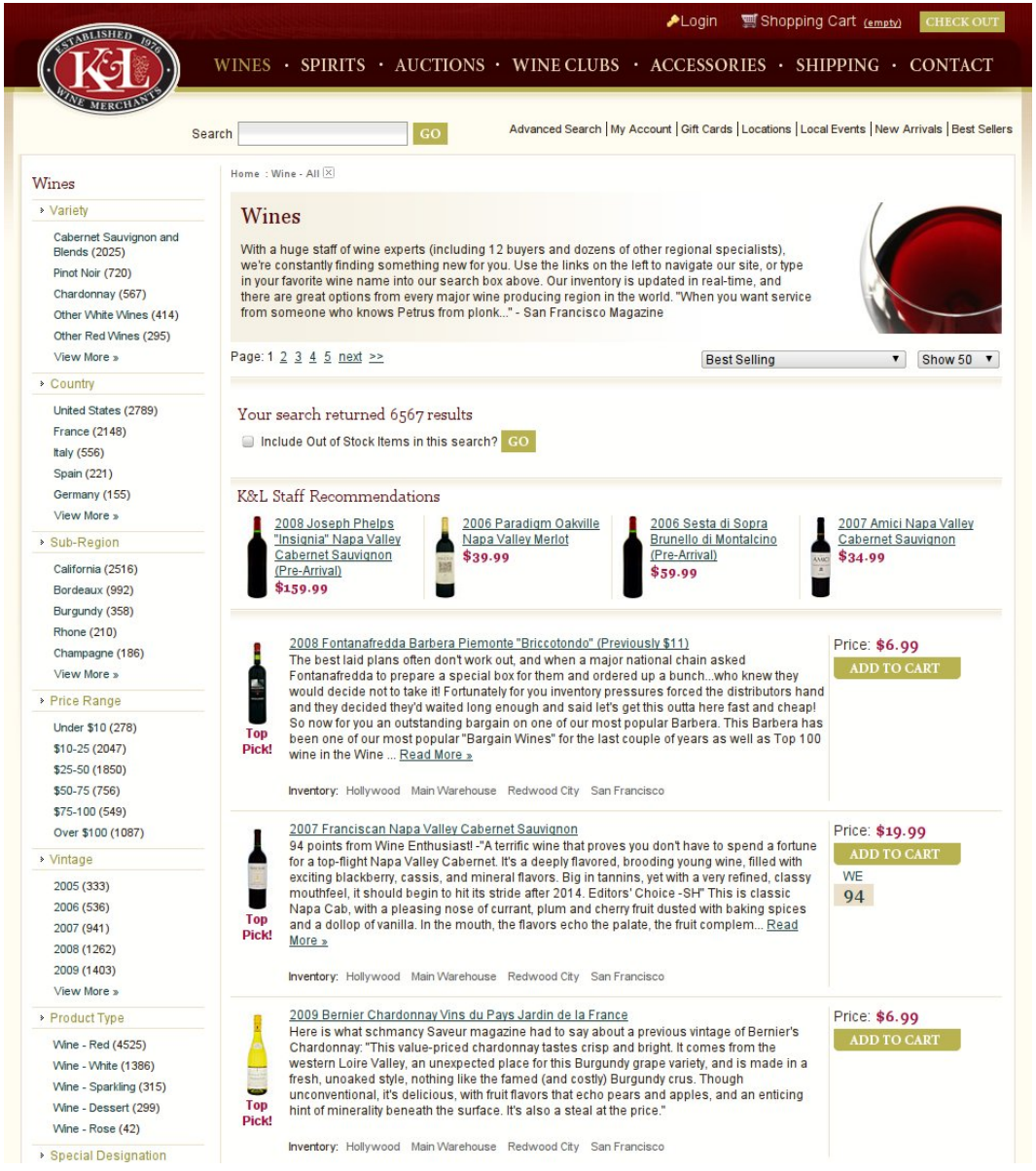
Because there is a gold standard for evaluation.



FINISH OFF BY DOING A REAL COMPETITION

Data Mining competitions with Kaggle.
Started at the middle of the course, students who ranked in top 3 didn't have to take the final.

PICK AN INTERESTING DATASET



WINE HAS MANY GREAT STORIES!

- Robert Parker is arguably the most influential wine critic.
- ``The Robert Parker Effect'': wines ranked 90 points cost significantly more than those ranked 89 points.
- Each spring since 1994 Robert Parker evaluates many wines in Bordeaux.
- In 2003, at the request of his family, he didn't make the trip.
- Economists have estimated that this depressed wine prices by about 15%.

PLENTY OF STRUCTURED INFORMATION

Home : 2008 Ridge Vineyards "Lytton Springs" Dry Creek Zinfandel

ZOOM

2008 Ridge Vineyards "Lytton Springs" Dry Creek Zinfandel

SKU #1061158

93 points from Robert Parker: "Striking, intense black cherry and blackberry fruit with some spice and earth jump from the glass of the 2008 Lytton Springs, a blend of 74% Zinfandel, 21% Petite Sirah, and 5% Carignan. Dark ruby with a nice tannic overlay, the wine was aged 15 months in American oak. Spicy, impressively rich, with good acids and loads of concentration, this is a beauty to drink over the next 5-7 years." (Feb. 2011) 93 points and two stars from the Connoisseurs' Guide to California Wine: "Ripe, but not overly so, and carefully crafted with a certain claret-like polish that is the Ridge signature, this year's Lytton Springs bottling is a deep, very well-focused wine that keys on varietal berries with complexing notes of dusty and dry spice lending a little more range than everyday Zinfandel gets. It is nicely balanced with a fine spine of tannins for grip but maintains its sense of finesse and composure right to the end, and, if not so astringent that it cannot be enjoyed now, it is built to get better for several years and will hold for many more." (Jan. 2011) Ripe nose of raspberry, plum, pepper and chapparel. Blackberry, mineral and vanilla notes dominate the palate. Well integrated tannins typical of this classic vineyard add to the long finish.

- View Additional Information
- Recommend this item to a friend
- Share
- Bestsellers from our entire inventory
- Bestsellers from within Domestic Zinfandel

Price: $27.99

Quantity: 1

ADD TO CART

CG	RP
93	93

Real Time Inventory by location:

The item you have chosen is in stock, and has inventory in our warehouse and one or more stores. Below is the up-to-the-minute **quantity on hand** information for this product within our database (updated each time you refresh this page):

Main Warehouse:	28
Redwood City:	>48
San Francisco:	>48
Hollywood:	14

Product turnaround time varies by location of inventory and your chosen method of shipping/pickup. For a detailed explanation click here .

Product Reviews: Add your own review of this item

You also might be interested in...

- 2009 Buehler Napa Valley Zinfandel **$16.99**
- 2007 Kalinda El Dorado Zinfandel **$12.99**
- 2009 Shenandoah Vineyards "Special Reserve" Amador Zinfandel **$9.99**
- 2006 Rosenblum "Richard Sauret" Paso Robles Zinfandel **$18.99**
- 2006 Rosenblum "Planchon Vineyard" San Francisco Bay Zinfandel (Previously $23) **$12.99**

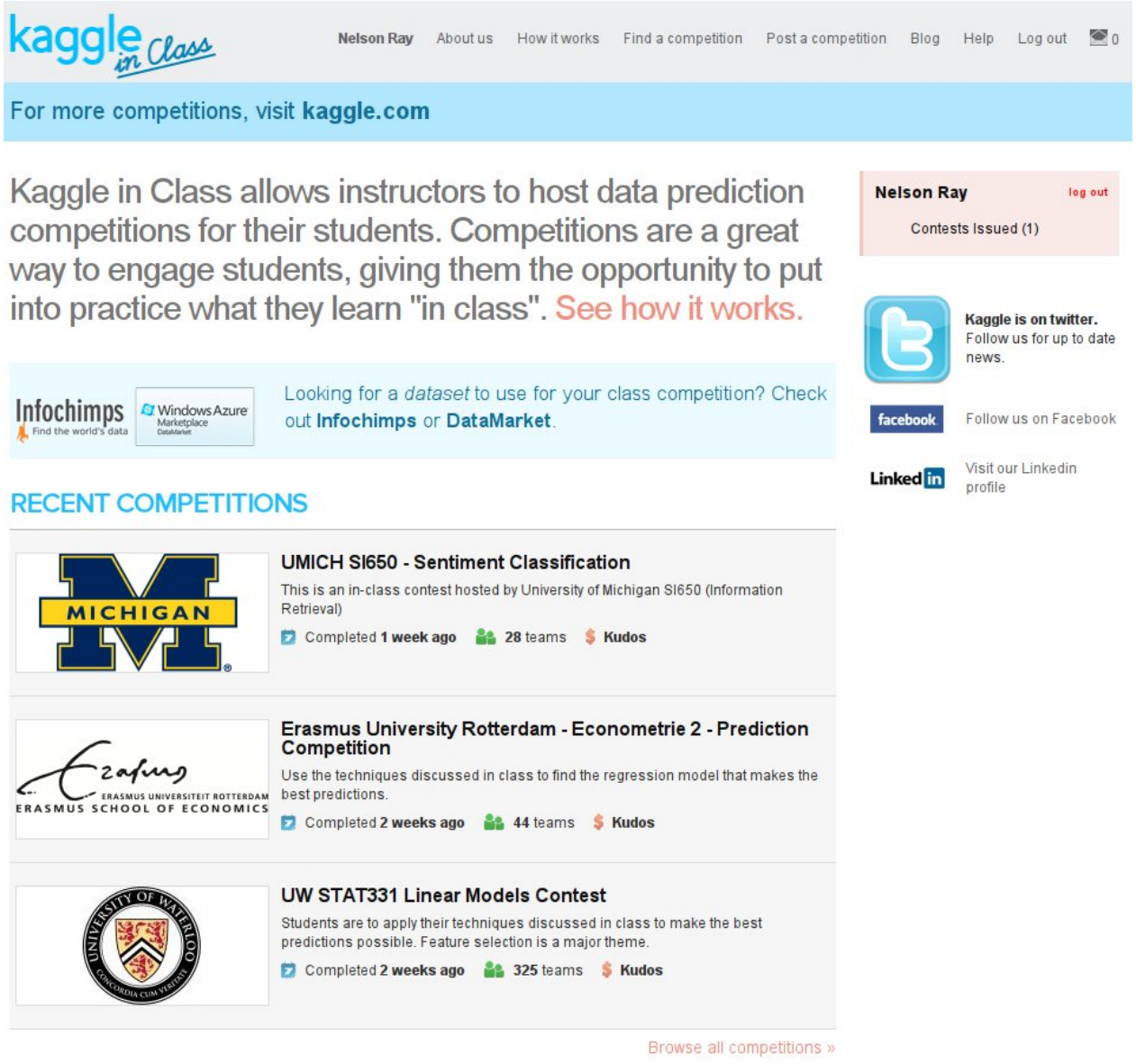
Additional Information:

Varietal:	**Zinfandel** - The bid to name Zinfandel California's "State Varietal" may have failed, but this red wine grape, grown extensively in California since the mid-1800s, is grown in few other places in the world. Sadly, much of what's cultivated today is planted where it's too hot and flat. But when planted to well-drained, hillside vineyards that are warm but not too hot, like those in Sonoma County's Dry Creek Valley and Amador County in the Sierra Foothills, Zinfandel can produce wines with plenty of character. High in natural alcohol and tannin, grown carefully it can be rich and complex, with dark fruit berry fruit and peppery spice. The most known example of Zinfandel outside of California is Italy's Primitivo, which can be similar in style, but is often a bit lighter and less alcoholic than West Coast examples.
Country:	**United States** - When people consider domestic wine, they normally think about the state of California. The fine viticultural Region within California, including the Napa Valley, Sonoma, Santa Cruz Mountains, Mendocino and Santa Barbara, are capable of growing grapes of world-class quality. But there's plenty of fabulous wine coming from other states, too. Oregon, Washington and New York are also causing eyebrows (and glassware) to be raised around the world. Click for a list of bestselling items from the United States.
Sub-Region:	**California** - With the explosive growth that California's wine industry has seen the past several years, it's easy to view winemaking and grape growing in the Golden State as a recent phenomenon. And while it's true that California's viticultural history is brief compared to several European countries, this state's roots date back well over 200 years. Due to the enormous response to California wine within the United States and worldwide, there are thousands of excellent and diverse wines being produced within the state each year. For our entire selection of California wines, please visit this link.
Specific Appellation:	**Sonoma County** - Second in fame only to Napa, this "other" valley offers just about every climate and topography imaginable. From its cool and fog-enshrouded coastal regions on the far west, to the sprawling Alexander Valley on the boarder of Napa and the many little dips and peaks in between, Sonoma has been a vital wine-grape-growing region since the mid 1800s. Important sub-AVAs include Chalk Hill (known for chardonnay and sauvignon blanc), Dry Creek Valley (where zin is king) Knights Valley (largely cabernet land), Russian River Valley and Sonoma Coast (both celebrated pinot and chardonnay zones).

IN ONE TIDY PACKAGE

```
"SKU","name","year","price","WS","RP","ST","WE","CG","GR","WN","BH","W.S","Varietal","Country","SubRegion","Appellati
1060779,"2005 Perrin & Fils ""Les Sinards"" Châteauneuf-du-Pape (Elsewhere $35)",2005,19.99,92,91,90,NA,NA,NA,NA,NA,N
1037599,"2005 Bodegas Castano ""Solanera Vinas Viejas"" Yecla (Previously $14)",2005,7.99,NA,91,90,NA,NA,NA,NA,NA,NA,
1060093,"2003 Bodegas Poesia ""Poesia"" Lujan de Cuyo (Elsewhere $50)",2003,19.99,NA,90,90,93,NA,NA,NA,NA,NA,"Caberne
1038331,"2007 Sacred Hill Savignon Blanc Marlborough (Elsewhere $15)",2007,5.99,NA,NA,NA,NA,NA,NA,NA,NA,NA,"Sauvignon
1045475,"2006 Rosenblum ""Monte Rosso Vineyard"" Sonoma Valley Reserve Zinfandel (Previously $30)",2006,19.99,NA,NA,N
1060504,"2008 Clos Chanteduc Côtes du Rhône",2008,10.99,NA,90,NA,NA,NA,NA,NA,NA,NA,"Rhone Blends","France","Rhone","
1060512,"2008 Slipstream Sauvignon Blanc Pemberton Western Australia (Elsewhere $14)",2008,6.99,NA,NA,89,NA,NA,NA,NA,
1060774,"2008 Talbott ""Logan"" Santa Lucia Highlands Chardonnay",2008,15.99,NA,92,NA,NA,NA,NA,NA,NA,NA,"Chardonnay",
1049412,"2008 Bernier Chardonnay Vins du Pays Jardin de la France",2008,6.99,NA,NA,NA,NA,NA,NA,NA,NA,NA,"Chardonnay",
1060794,"2009 Perrin & Fils ""Reserve"" Côtes du Rhône",2009,7.99,NA,NA,NA,NA,NA,NA,NA,NA,NA,"Rhone Blends","France",
1042222,"2005 Celler de Capcanes ""Mas Donis"" Barrica Montsant  (Previously $11)",2005,8.99,NA,91,NA,NA,NA,NA,NA,NA,
1061214,"2009 Banshee Santa Lucia Highlands Pinot Noir",2009,19.99,NA,NA,NA,NA,NA,NA,NA,NA,NA,"Pinot Noir","United St
1059788,"2008 Beringer ""Private Reserve"" Napa Valley Chardonnay",2008,27.99,92,94,92,NA,NA,NA,NA,NA,NA,"Chardonnay"
1046284,"2004 Fanti Brunello di Montalcino (Pre-Arrival)",2004,29.99,94,NA,NA,92,NA,NA,NA,NA,NA,"Sangiovese","Italy",
1060283,"2008 D'Alessandro Cortona Syrah",2008,13.99,NA,90,91,NA,NA,NA,NA,NA,NA,"Shiraz/Syrah","Italy","Tuscany","Sup
1048803,"Urbanite Cellars ""Redart"" Red Blend",NA,9.99,NA,NA,NA,NA,NA,NA,NA,NA,NA,"Other Red Wines","United States",
1060475,"2007 Bisci ""Fogliano"" Verdicchio di Matelica (Elsewhere $20)",2007,14.99,NA,91,NA,NA,NA,NA,NA,NA,NA,"Other
1060200,"2008 Chateau Ste. Michelle ""Indian Wells"" Columbia Valley Cabernet Sauvingnon",2008,12.99,90,NA,NA,NA,NA,N
1059653,"2008 Patient Cottat ""Le Grand Caillou"" Sauvignon Blanc",2008,10.99,NA,NA,NA,NA,NA,NA,NA,NA,NA,"Sauvignon B
1050286,"2007 Fort Ross Sonoma Coast Pinot Noir",2007,29.99,91,NA,NA,95,NA,NA,NA,NA,NA,"Pinot Noir","United States","
1060467,"2007 Andrew Rich Willamette Valley Pinot Noir (Elsewhere $30)",2007,19.99,NA,NA,90,NA,NA,NA,NA,NA,NA,"Pinot
1061215,"2007 Banshee Napa Valley Cabernet Sauvignon",2007,29.99,NA,NA,NA,NA,NA,NA,NA,NA,NA,"Cabernet Sauvignon and B
1057997,"2009 Weingut Allram Grüner Veltliner (1L)",2009,10.99,NA,NA,NA,NA,NA,NA,NA,NA,NA,"Gruner Veltliner","Austria
1060162,"2006 Bodegas Lan Crianza Rioja",2006,9.99,90,NA,NA,NA,NA,NA,NA,NA,NA,"Tempranillo","Spain","Rioja","","13.5"
```

KAGGLE-IN-CLASS



THE COMPETITION PAGE

Stanford Stats 202 Wine Price Prediction

PRIZE POOL	TEAMS	COMPLETED
Kudos	32	19 weeks

Information Data Submissions Forum Results

- Description
- Evaluation
- Prizes

Login

Username

Login or register now

Forgot your Username/ Password?

2 discussions in this competition's forum

post-competition model sharing — 19 weeks ago

looking for a teammate — 23 weeks ago

Results

01.	S202 Team Eureka
02.	S202 Team Cleopuptra
03.	S202reb
04.	S202 Team Hops
05.	winepycho
06.	wine
07.	bei
08.	S202 Team JZ
09.	outsider
10.	Joan

Use wine vintage, varietal, country, ratings, and other information to predict price. The top 3 teams from Stats 202 will not have to take the final exam.

The price of a wine is influenced by many factors. Wines from Pomerol, a region in south-western France often fetch a pretty penny. First growths such as Château Margaux and Château Lafite Rothschild are often quite expensive, compounded especially for a stellar, old vintage. A Robert Parker rating of 100 is pretty much a license to print money.

The training set comprises 44,151 observations, and the test set 18,923. Submissions will be evaluated via L1 error, or mean absolute error. One submission is allowed per day, and the leaderboard ranking is calculated based on 50% of the test set. The entire test set is used for the final ranking.

The top 3 teams from Stats 202 will not have to take the final exam. If you are in Stats 202, teams are limited to 3 people.

If you are a member of Stats 202, please prepend an identifier to your name. Something like [S202]*name*. This is so we can easily identify the winners in the class. We have provided the names of the wines for the extra information they supply. It would be possible to get 0 test error in this competition simply by looking up the prices corresponding to the names on www.klwines.com. For this reason, we are requiring that the winners also provide their code.

Outside information is allowed, up to a point. For instance, we would be thrilled to see some team extract winery information from the title and then connect that with outside information (the winery is located ____ with climate ____ and usually prices their wines ____) in a useful fashion. However, simply searching for the wine prices is unacceptable. You must provide a non-trival model.

Update: We have gotten rid of wines that were priced over $1,000 (there was a pricing error for these wines in the previous test and training data). If you have downloaded the data before November 10, please make sure to download the new versions. There are now 43,715 training observations and 18,741 test observations.

This competition ended at 1:59am, Tuesday 7 December 2010 UTC.

Visit the forum | Download the data | View the leaderboard | Enter competition

THE RESULTS

#	Team Name	MAE	Entries	Latest Submission
1	S202 Team Eureka *	14.2701	20	6:44pm, Monday 6 December 2010
2	S202 Team Cleopuptra *	16.3072	8	6:41pm, Monday 6 December 2010
3	S202reb *	17.4029	14	10:03pm, Sunday 5 December 2010
4	S202 Team Hops	17.9048	10	11:01pm, Sunday 5 December 2010
5	winepycho	21.8187	4	6:31pm, Monday 6 December 2010
6	wine	21.821	2	11:41am, Monday 6 December 2010
7	bei	21.9851	2	12:45pm, Monday 6 December 2010
8	S202 Team JZ	22.134	6	12:50am, Monday 6 December 2010
9	outsider	22.3803	4	7:59am, Monday 6 December 2010
10	Joan	22.3803	2	12:42am, Monday 6 December 2010
11	MQSBF	24.3192	4	6:35pm, Monday 6 December 2010
12	S202 Fortified Wine	24.7443	10	5:46pm, Monday 6 December 2010
13	S202 Team Franzia	25.7653	9	4:06pm, Monday 22 November 2010
14	winelover	26.5527	8	3:23pm, Saturday 4 December 2010
15	s202 Outlier	26.7379	6	8:16pm, Sunday 5 December 2010
16	S202 skmenon	27.8962	19	5:44pm, Monday 6 December 2010
17	S202 Qin Zeng	40.0686	11	7:47pm, Friday 26 November 2010
18	miner99	40.3637	3	9:47pm, Tuesday 23 November 2010
19	S202 Team Pososhok	42.7053	6	8:07pm, Tuesday 16 November 2010
20	S202 WorseThanNaive	42.7172	2	5:07am, Sunday 21 November 2010

THE WINNERS REVEAL THEIR STRATEGY

How we did it: Jie and Neeral on winning the first Kaggle-in-Class competition at Stanford

13 December 2010
by Jie Yang
(@jacksheep)

How I Did It

3 Comments

Neeral (*@beladia*) and I (*@jacksheep*) are glad to have participated in the first Kaggle-in-Class competition for Stats-202 at Stanford and we have learnt a lot! With one full month of hard work, excitement and learning coming to an end and coming out as the winning team, it certainly feels like icing on the cake. The fact that both of us were looking for nothing else than winning the competition, contributed a lot to the motivation and zeal with which we kept going each and every day. Each of us may have spent about 100 hours on this competition, but it was totally worth it.

Even though the professor and textbook already mentioned some points which ended up to be very useful, the experiences we gained from this practice were much more influential to us. We strongly recommend every data mining student to participate in such competitions, make your hands dirty, and eventually every minute you invest on it will reward back.

ANALYZE THE DATA FIRST

The most important lesson we learnt was: we should always analyze the data first.

As newbies in data mining, we had thought a cool model could give us the best result, so the main effort was to find such advanced models and keep tuning it. We used the features from the raw data, did some feature transformation to deal with missing values, and tried some supervised learning model, such as lm, randomForest and glmboost in R. It seemed working, because Random Forest could improve 18% comparing to linear regression.

POST-MORTEM

- At least 20 teams from Stats 202 participated.
- Teams were limited to three people.
- Roughly 160 students in the class.
- Contest ran for about a month with a one submission per day limit.
- The top 3 teams averaged more than one submission every other day.
- Received many thank yous from students for the competition.

IDEAS FOR IMPROVEMENT

- Started the competition in the final month and had some integration with the class (displayed the leaderboard at the beginning of class and had some homework assignments involving the data)
- Could start even earlier and provide sample code using increasingly sophisticated techniques (i.e. starting with linear models and moving to basic text mining for features in boosted trees).
- Code would be provided at regular intervals to allow new participants to easily catch up and push the leaders even further.
- Have ``progress prizes'' so the top teams don't entirely neglect studying for the final.