1. Data Set

The complexity of wine has developed many sophisticated palates, some of which are self-proclaimed, which can distinguish even the most nuanced tastes, such as notes of "forest fruits", "earth" and even "old leather boot". Prominent wine critics and sommeliers have emerged over the years, and have guided the general attitudes and appeal towards different wines, thus also the prices. The most influential has been Robert Parker, who instated the 100 point wine rating system, and made wine reviews more digestible by the general public. The benefits and detriments of such a simplified wine rating system can be debated, but nonetheless it has enabled consumers to make decisions on which wines to purchase simpler, and thus has helped propel wine consumption around the world.

The data set for this assignment is the CellarTracker data, which is the largest collection of wine and ratings provided by its users. As shown in Table 1, the original data set contains over 2 million reviews, and from that, a training set and test set was produced at random.

Table 1. Data Set Statistics

	Counts		
	Original Dataset	Training	Test
# Reviews	2025995	305389	76572
# Unique Wine Variants	830	624	454
# Unique Wines by ID	485179	156087	56296
# Unique Wines by Name	Not counted	155370	56192
# Users	44268	21902	12593

In order to understand the data, and later form a model, some initial analysis must be performed. Table 2 and Table 3 show the 10 most and least reviewed wine variants, respectively. It is no surprise Pinot Noir tops the list of most reviewed, since it has become known over the years as being easy to pair. And, also unsurprising, are those in the least reviewed list, comprising mostly of rare varietals, including Phoenix, a German grape, and Piculit-Neri, an Italian grape.

Table 2. Most Reviewed Wine Variants

	Wine Style
1	Pinot Noir
2	Red Bordeaux Blend
3	Cabernet Sauvignon
4	Chardonnay
5	Red Rhone Blend
6	Syrah, Riesling
7	Red Blend
8	Zinfandel
9	Shiraz
10	Syrah

Table 3. Least Reviewed Wine Variants

	Wine Style
1	Cabernet Cubin
2	Cagnulari
3	Cagnina
4	Karaoglan
5	Pear-Apple Blend
6	Karalahna
7	Phoenix
8	Cabernet Moravia
9	Cabernet Gernischt
10	Piculit-Neri

The 100 point system put forth by Robert Parker and duplicated by many other critics is: 96-100 for extraordinary, 90-95 for outstanding, 80-89 for above average to very good, 70-79 for average, 60-69 for below average, and 50-59 for unacceptable. It is thus no surprise that the 10 most reviewed wines, shown in Table 4, generally have very good scores, with average points of around 90 points. There is mental barrier around 90 points, enforced by wine merchants: Consumers seek higher point wines, and merchants typically highlight wines that score close or above 90 points. They may feature them in ads, or identify them and their rating by a special label, easily seen by consumers browsing the wine aisle. What is surprising is that the 10 least reviewed wines, shown in Table 5, also have generally good scores. Intuition would suggest that their good scores would encourage others to purchase them, but perhaps the reviewer or reviewers who provided the scores are not highly regarded in the CellarTracker community.

Note that in Table 4, there are two wines that appear with the same name. The data set was verified, and though they have the same name, they have different wine ids (260581 and 556767).

Table 4. Most Reviewed Wines

	Rating			Wine Name	Approx. Price (USD)
	Ave.	High	Low		
1	90.35	95.00	77.00	2007 Seghesio Family Vineyards Zinfandel Sonoma County	40
2	92.48	98.00	86.00	NV Krug Champagne Grande Cuvee Brut	200-300
3	90.47	97.00	84.00	NV Bollinger Champagne Special Cuvee Brut	30
4	87.80	95.00	78.00	NV Veuve Clicquot Ponsardin Champagne Brut	25
5	87.39	90.00	77.00	2006 Monte Antico Toscana IGT	30
6	92.86	96.00	87.00	2000 Domaine du Pegau Chateauneuf-du-Pape Cuvee Reservee	100
7	89.04	92.00	65.00	2008 Kim Crawford Sauvignon Blanc Marlborough	20
8	89.65	93.00	50.00	2006 Kim Crawford Sauvignon Blanc Marlborough	20
9	88.17	93.00	60.00	2006 Mollydooker Shiraz The Boxer	20
10	92.23	97.00	84.00	2003 Chateau Pontet-Canet	100

Table 5. Least Reviewed Wines

	Rating			Wine Name	Approx. Price (USD)
	Ave.	High	Low		
1	89.00	89.00	89.00	2006 Anglim Pinot Noir Fiddlestix Vineyard	50
2	84.00	84.00	84.00	2006 Beringer Vineyards Zinfandel California Collection	25
3	89.00	89.00	89.00	1997 El Grifo Lanzarote Canari	30
4	88.00	88.00	88.00	2007 Kunin Zinfandel WestSide	25
5	85.00	85.00	85.00	2008 Johann Topf Gruner Veltliner Wechselberg	175
6	86.00	86.00	86.00	2008 Villfane & Guzman Malbec Parados	NA
7	86.00	86.00	86.00	2005 Villa Sant Andrea Chianti Classico Castello di Fabbrica	NA
8	84.00	84.00	84.00	1999 Domaine Jaboulet-Vercherre Vosne-Romanee	NA
9	84.00	84.00	84.00	2006 Mario Marengo Barbera d'Alba Pugnane	24
10	89.00	89.00	89.00	2005 Bodega La Colegiada Vino de la Tierra de Castilla y Leon Pago De Florentino	30

The prices per bottle were manually collected from WineSearcher, and included in these tables, to gain some insight on the prices of wine and how they relate to ratings. The objective quality of the wine aside, the general economics of wine indicate that the price of wine follows the rating; for example if a wine receives a very high Robert Parker score, its price increases. It is also suspected that the rating of wine follows the price. Either or both or none (ie. the wine is objectively excellent) of these may be in effect in the highest rated wines, Table 6. These wines have excellent scores, and have very high prices.

The same mechanism is harder to suspect in the lowest rated wines, Table 7. Many of them were not found in Wine Searcher at the exact vintage, or at all, and so are listed as "NA" for price. Note, however, wine prices are not static, and thus change over time, and are exacerbated by wine speculation. The wine prices shown are the current prices of the wines. To truly see the relation between price and ratings, the price at the time of each rating should be examined and not simply the current price.

Table 6. Highest Rated Wines

	Rating			Wine Name	Approx. Price (USD)
	Ave.	High	Low		
1	97.04	100.00	87.00	1990 Chateau Margaux	800-1000
2	100.00	100.00	100.00	1995 Domaine de la Romanee-Conti Montrachet	4000-6000
3	99.00	100.00	98.00	2009 Chateau Margaux	700-900
4	98.33	100.00	96.00	1990 Krug Champagne Clos du Mesnil	1500-2000
5	99.50	100.00	99.00	NV Daniel Bouju Tres Vieux Brut de Fut	300
6	98.00	100.00	95.00	1994 Harlan Estate	1000-1600
7	99.00	100.00	98.00	2008 Fairchild Estate Cabernet Sauvignon Sigaro	200
8	99.00	100.00	98.00	2001 Albert Mann Gewurztraminer Furstentum Selection de Grains Nobles	NA
9	99.20	100.00	97.00	1959 Chateau Lafite Rothschild	3000-7000
10	96.20	100.00	93.00	2005 Cayuse Syrah Bionic Frog	400

Table 7. Lowest Rated Wines

	Rating			Wine Name	Approx. Price (USD)
	Ave.	High	Low		
1	50.00	50.00	50.00	2007 Brest Pere et Fils Saint Pourcain Vieilles Vignes Les Crechoux	NA
2	50.00	50.00	50.00	1999 Alois Kracher Chardonnay Welschriesling Days of Wine and Roses	NA
3	50.00	50.00	50.00	1889 Chateau La Mission Haut-Brion	1938: 1000
4	50.00	50.00	50.00	2003 Louis Eschenauer Vin de Pays d'Oc	2008: 6
5	50.00	50.00	50.00	2005 Kiss Chassey	2011: 15
6	50.00	50.00	50.00	2006 Salmon Harbor Merlot	NA
7	50.00	50.00	50.00	2006 Vignoble de Sainte-Petronille Voile de la Mariee	2013: 13
8	50.00	50.00	50.00	NV Oovvda Winery Raspberry	20
9	50.00	50.00	50.00	1999 Domaine Saint-Vincent Vin de Pays d'Oc	NA
10	50.00	50.00	50.00	1999 Loudoun Valley Vineyards Merlot	NA

Another interesting phenomenon with the data is that 25% of reviewers provided a review by text but did not provide a rating. Examination of review text samples reveal that there are cases, obviously, where the reviewer liked the wine, and where the reviewer didn't like the wine. Of the reviewers who liked the wine, perhaps some neglected to provide a rating, or simply refused for personal reasons (ie. not confident in their rating, disagrees with the rating system). Of the reviewers who didn't like the wine, some abstained to allow the wine to age and develop. Figure 1 shows two examples of this. Aside from understanding that there is a "right time" with wine, these reviewers are exhibiting caution with providing a rating, and perhaps even reluctance at assigning low scores to wines.

{"wine/name": "2007 Coho Headwaters", "wine/wineId": "669310", "wine/variant": "Red Bordeaux Blend", "wine/year": "2007", "" "review/time": "1271203200", "review/userId": "111964", "review/userName": "cesmd", "review/text": "Medium bodied, fruity and spicy with clear cab notes of blackberry and cherries, but with a modest nose that gave little hint of the fruitiness, and brief, one-dimensional finish. The tannins were somewhat stiff and distracting on first impression, but improved with an hour of air time and the fruit began to show itself and become downright charming. Still, I've banished my second bottle to the cellar for a few years to think about things, though at these prices, one could put a lot away."}

{"wine/name": "2002 Diebolt-Vallois Champagne Brut Blanc de Blancs", "wine/wineId": "278684", "wine/variant": "Chardonnay", "wine/year": "2002", "" "review/time": "1288137600", "review/userId": "415", "review/userName": "Siggy", "review/text": "Brief notes- Large-scaled, intense, and youthful. Ripe, complex apple-infused fruit, lively acidity, and pungent supporting minerality that build to a long finish. Still needs a little time to settle down and harmonize -- should be even better with a couple more years of bottle age."}

Figure 1. Sample Reviews With No Rating

Reviewers are certainly aware of the impact a score can have, especially a low score (below the mid-80s). From this, it can be inferred that there will be more "good" ratings than "bad" ratings, objectivity aside. The histogram shown in Figure 2, does show this trend, but of course does not prove it.

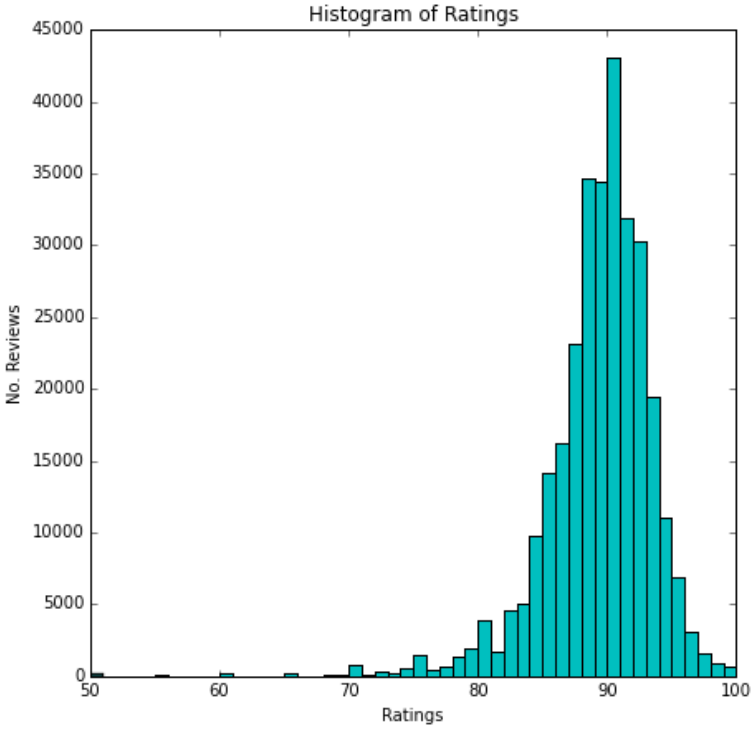

Figure 2. Histogram of Ratings

Most scores are 80 points and above, and there is a long tail towards 50, which is the lowest point achievable. A rating of 90 to 91 points is the highest occurring rating. This reinforces the idea of a mental barrier around 90 points: If a reviewer is reviewing a very good wine, there is a fine line between 89 and 90 points, and so they will likely opt to give the higher score.

As mentioned in the beginning, wine critics have emerged over the years as the dominant voices in the wine industry. In a similar fashion, wine experts may emerge within wine rating communities, providing a general guidance for the rest. If this in fact occurs, then the opinions or ratings on wine will tend to homogenize over time. To find indications of this, the ratings of select wines are plotted over time, shown in Figure 3.

Note that the wines shown are from the top most reviewed list, so as to have enough reviews to see any trends. Also note that the time is in UNIX time, where 1.080777600e9 is equivalent to 1 April 2004, the approximate launch date of the CellarTracker website.

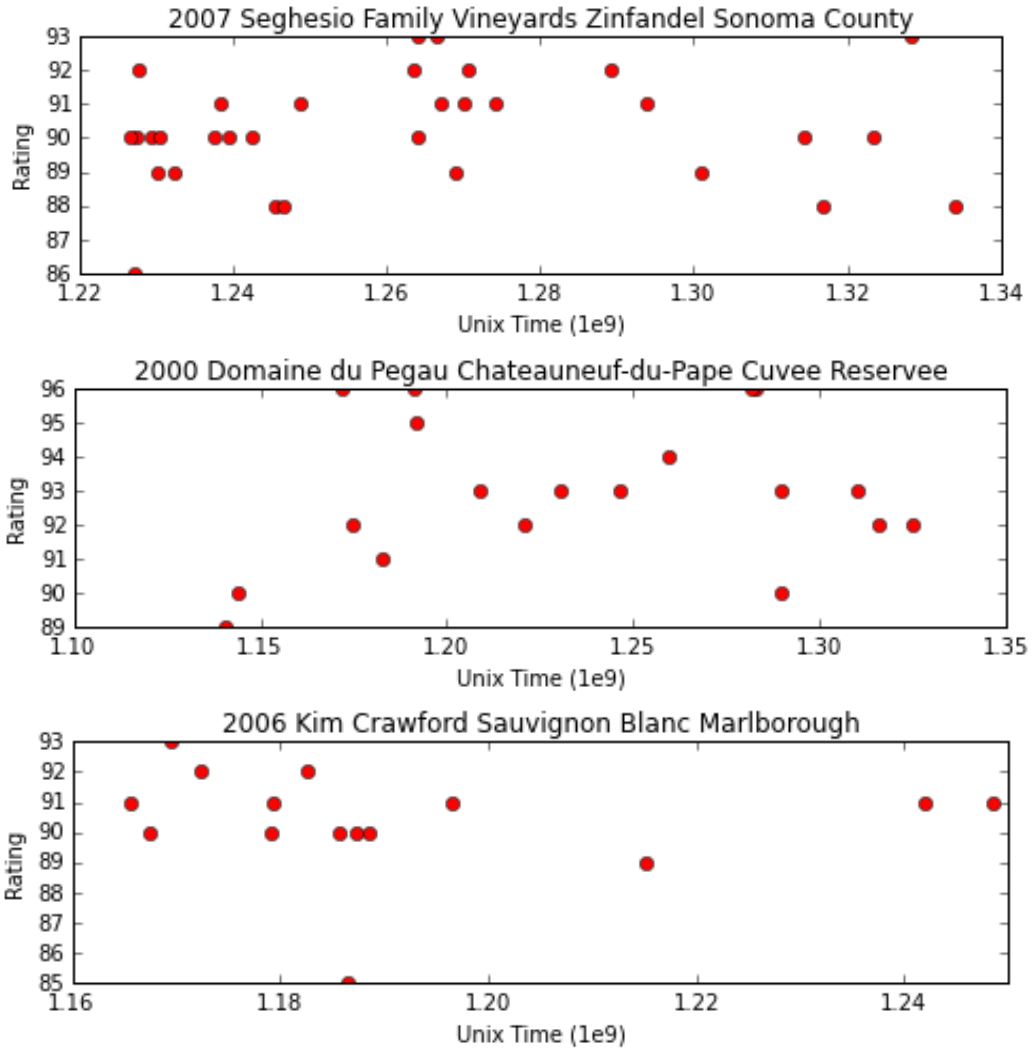

Figure 3. Ratings Over Time for Select Wines

2. Predictive Task

The plots of ratings over time, Figure 3, appear to have less variance at the later epoch than the first epoch, thus the hypothesis of ratings homogenizing over time is fair. The predictive task is then, given the reviewer (user u), the wine (item i) and the time (t), predict the rating that will be given. This differs from the standard model, in that its parameters will be dependent on time:

$$rating(u, i, t) = \alpha(t) + \beta_u(t) + \beta_i(t) \quad \text{Eq 1}$$

Recall that the standard model is:

$$rating(u, i) = \alpha + \beta_u + \beta_i \quad \text{Eq 2}$$

To evaluate the performance, the mean squared error (MSE) on the test set can be calculated against the standard and the time-dependent model:

$$\text{MSE} = \frac{1}{N_{test}} \sum_{j \in test} (\hat{r_j} - r_j)2 \quad \text{Eq 3}$$

In this equation, $\hat{r_j}$ is the predicted rating and r_j the actual rating, or ground truth, for a given test datum j. If the MSE of the time-dependent model is less than that of the baseline, the model is then successful and confirms that to model the user's and the community's behavior more accurately, time must be taken into account.

3. Literature

The CellarTracker data used here is also used in the research paper "From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews" by Julian McAuley and Jure Leskovec. In their research, McAuley and Leskovec examine the temporal dynamics in product ratings, and offer three time-dependent mechanisms that cause a user's attitude of a product to change: the age of the product, the age of the user, and the general attitude of the community.

They conduct several experiments against different models, which differ in whether the parameters are learned for the individual or the community, and with or without the concept of a progression of time and expertise. Their first model, with parameters that evolve for the entire community as a function of time, is most similar in concept to the time-dependent model put forth here.

McAuley and Leskovec also develop a model for the progression of a user's experience and expertise regarding a product. Rather than the model parameters evolving over time, they evolve specifically for each user, based on experience level, which is itself evolving over time. This is a very novel idea, and as shown in their results, provides a more accurate prediction of a user's ratings for a particular product. With the CellarTracker data, their achieved MSE for the first model with the community uniform rate is 0.051, and for their more complex model with the user-learned rate, is 0.045.

4. Features

As mentioned in Section 1, it is hypothesized that time effectively influences a user's review of an item, since any existing ratings for the item and influence of experts work to homogenize ratings. Thus, time has been chosen as a factor in this model.

The model attempts to learn the general offset, $\alpha(t)$, user bias, $\beta_u(t)$, and the item bias, $\beta_i(t)$, as functions of time t. To build parameter $\beta_u(t)$, the items that user u has reviewed and ratings r given, up to time t, are used as features. And similarly, to build $\beta_i(t)$, the users who have reviewed item i and all its ratings r, up to time t, are used as features.

Prior to training, the data was transformed into JSON, and then 'cleaned' of unwanted or unusable data. The cleaning steps performed were:

- Replace the HTML codes of foreign letters to their closest English equivalent, if it exists
- Replace the remaining HTML codes (ie. of symbols) with white spaces
- Remove single and double quotes, and forward and backslashes, and colons from value strings
- Remove data with a UNIX time less than (ie. earlier than) 1080777600, which is 1 April 2014, the approximate launch date of CellarTracker
- Remove data that do not have values for rating ('review/points'); all data had the other features, user id ('review/userId'), item id ('wine/wineId') and time ('review/time')

5. Model

The model equation, as provided in Eq 1, is a linear combination of the parameters α(t), β_u(t) and β_i(t). Their update equations are show below, as Eq 4 to 6. $I_u(t)$ is a vector representation of the items reviewed by user u at time t. Similarly, $U_i(t)$ is a vector representation of the users which have reviewed item i at time t. The regularization parameter, λ, can be tuned using grid search but with the limitation of time, it is simply set to 1.

$$\alpha(\mathrm{t}) = \frac{\sum_{u,i,t \in train} \left(r_{u,i,t} - (\beta_u(t) + \beta_i(t))\right)}{N_{train}} \quad \text{Eq 4}$$

$$\beta_u(\mathrm{t}) = \frac{\sum_{i \in I_u(t)} \left(r_{u,i,t} - (\alpha(\mathrm{t}) + \beta_i(t))\right)}{\lambda + |I_u(t)|} \quad \text{Eq 5}$$

$$\beta_i(\mathrm{t}) = \frac{\sum_{u \in U_i(t)} \left(r_{u,i,t} - (\alpha(\mathrm{t}) + \beta_u(t))\right)}{\lambda + |U_i(t)|} \quad \text{Eq 6}$$

$$argmin_{\alpha,\beta} = \sum_{u,i,t \in train} \left(\alpha(\mathrm{t}) + \beta_u(\mathrm{t}) + \beta_i(\mathrm{t}) - r_{u,i,t}\right)^2 + \lambda[\sum_{u,t \in train} (\beta_u(\mathrm{t})^2)) + \sum_{i,t \in train} (\beta_i(\mathrm{t})^2)] \quad \text{Eq 7}$$

The parameters α and β for time t are updated iteratively until convergence of the optimization equation, Eq 7. Convergence is achieved when the values of the optimization equation at iteration i and i + 1 differ by less than 0.005. At this point, the parameters are stable up to at least 4 significant digits.

For training, the entire span of time from the first review to the last review in the training set is divided into intervals of time, or epochs. The parameters are then optimized for each epoch, and thus each epoch T will have an optimal set of parameters. During prediction, a review at time t' that falls in epoch interval T' will be evaluated with the parameters for epoch T'.

The optimal division of time will best capture the mood of the community including any biases, as it evolves. Ideally, many epoch sets would be attempted, however due to time and computer processing limitations, only 4 were attempted and with at most 8 intervals. The epoch set that produced the lower MSE on a validation set is shown in Table 8.

Table 8. Epochs

Epoch T	UNIX Time	GMT
1	1143849600	01 Apr 2006 00:00:00
2	1207008000	01 Apr 2008 00:00:00
3	1238544000	01 Apr 2009 00:00:00
4	1270080000	01 Apr 2010 00:00:00
5	1301616000	01 Apr 2011 00:00:00
6	1333238400	01 Apr 2012 00:00:00
7	1364774400	01 Apr 2013 00:00:00

6. Results & Conclusions

The resulting MSEs of the test set are shown in Table 9, and samples of the predicted and target ratings for the varietal Pinot Noir are shown in Figure 4.

Table 9. Results

	Mean squared error
Model, time-dependent	13.550551
Model, standard	12.735544

The time-dependent model does well compared to the standard model, but not better. This could mean that the hypothesis is incorrect, under-developed, or implemented incorrectly. Given that the epoch set used wasn't optimized, it is possible it is a combination of the latter two.

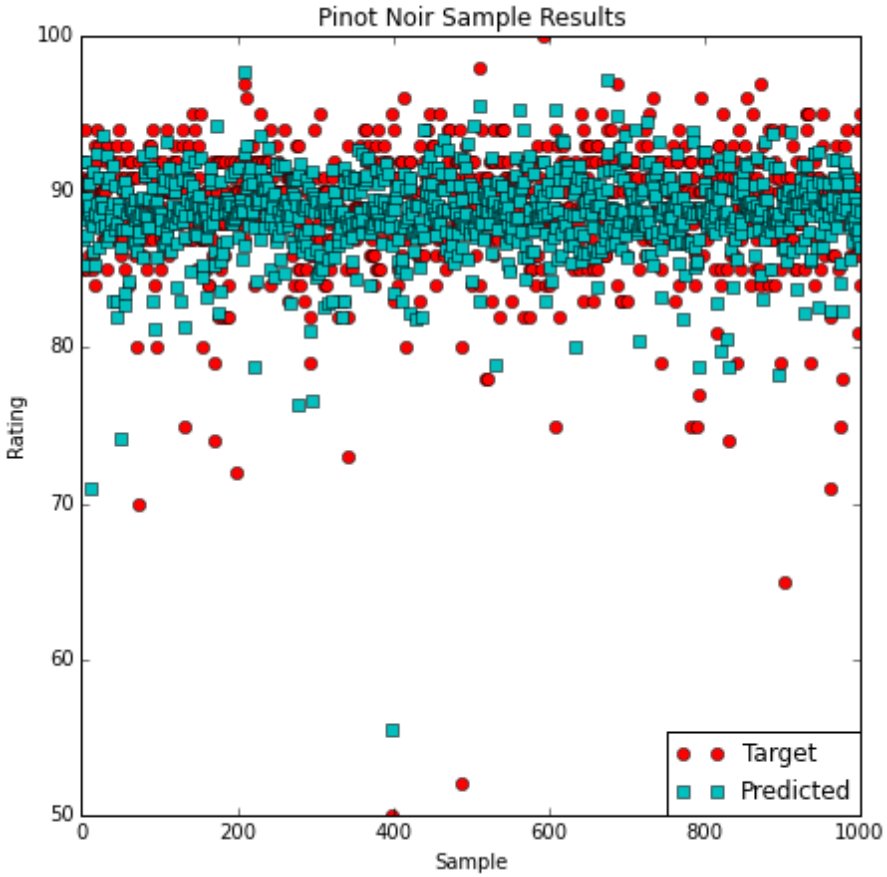

Figure 4. Pinot Noir Sample Results

A possible improvement to both the standard and time-dependent models is the use of latent factors. Currently, both models treat users and items independently, and do not currently model a user's affinity towards certain item features. For example, if a user has rated Pinot Noir wines highly in the past, then they are likely to rate a new Pinot Noir highly in the future. Parameters that may be useful include a users' action given particular varietals or common descriptors, such as 'jammy'. Another possible improvement is to develop a more sophisticated model, like that posed by McAuley and Leskovec, which models a user's expertise and can more accurately predict their experience with a product.

Further work can also be done to include the price of wine as a feature, since there are indications that price does influence ratings, perhaps by setting the expectations of a given wine. This was actually attempted, but scraping the data proved to be quite time consuming, and so was not completed.