Finding the Optimal Age of Wine

Andrew Prudhomme
A05419855
University of California, San Diego
9500 Gilman Dr.
La Jolla, California 92093
aprudhomme@eng.ucsd.edu

ABSTRACT

For this project, I attempted to characterize how the perceived quality of wine relates to its age. To do this, I performed analysis on the Cellar Tracker data set of wine reviews. Through exploration of the data set I found that wine does not age consistently. There was a noticeable shift in the trend of review scores as the wines' age increased. I performed various forms of regression analysis in an attempt to characterize this shift. Most importantly, I tried to determine the point at which this shift takes place, as it could be considered the optimal time to drink the wine (or the point of diminishing returns on quality). I came up with a model to this end and was able to obtain reasonable estimates for this value.

1. INTRODUCTION

It is often said that a fine wine will just get better with age, but is this really the case? And if so, in what manner does it improve? There are many potential improvement models that are possible. If it improves strictly linearly or exponentially, it would benefit a person to wait as long as possible before drinking a bottle of wine. Also, as a result, you would expect all wine to eventually converge to the highest rating as time passed.

There is another question that must be considered. Does the improvement of wine overtime follow a consistent trend? Is there a point in the life of a bottle of wine where it begins to see diminishing returns on its improvement? It might even be the case that quality might begin to decline as time continues to pass. If this is true, it would not be of benefit to save a bottle of wine indefinitely.

In the preceding case, one where quality would eventually decline, it is of great benefit to characterize where there is a shift in the aging trend. This point would represent a best time to drink the bottle of wine. Even in the event that the later aging trend is still an increasing one (though just at a slower rate), this would represent the point of diminishing returns. As the cellar space for a large wine owner can be a precious resource, this point could be used to determine what bottles should be consumed sooner to make room for younger bottles that would benefit more.

Seeing as this optimal age would be interesting and useful to know, it is what I decided to attempt to find. Through the remainder of this document, I walk you through the process I followed to develop a model for this goal. In Section 2 I describe the data set I have chosen to use. Section 3 describes that task I am attempting to solve. Section 4 contains information of previous literature relating to the data set and creation of a model. Sections 5 and 6 describe the data set features I used and the steps I took to evolve my model. Finally, Sections 7 and 8 summarize my results and conclusions.

2. THE DATA SET

The data set I chose to explore is one from the Cellar Tracker. Cellar Tracker is a website that caters to wine enthusiasts. It allows a user to catalog and manage their collection of wine. Additionally, it contains a large collection of reviews of wine that have been submitted by users. It is these user reviews that make up the data set.

The data set as a whole consists of 2,025,995 reviews made by 44,268 different users over 485,179 unique wines. Each entry contains expected meta data, such as a wine and user name/ID with a time stamp. There is some additional meta data on the wine, mainly the variant and year. Each review has the option of giving the wine a point score, with a value up to 100. Finally, there is a text body to the review that was written by the user. The text provided is relatively small (compared to some of the beer reviews) with a median word count of 29. Also, unlike the beer reviews, there are no subcategory scores.

I found this data interesting because of the presence of both a wine year and a post time stamp. What I quickly discovered was that not all of the reviews contained complete information. It was possible for a reviewer to select 'N/A' for either the wine year or the point score. As I felt this information was necessary, I created a filtered data set comprised only of reviews with valid information. The size of this reduced data set was 1,521,471 entries.

As I understood it might be useful to not think of all wines as equivalent, I wanted to know how much variety there was

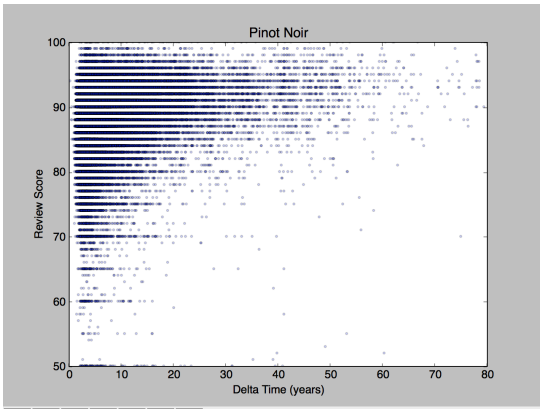

Figure 1: A graph of scores for a single variant.

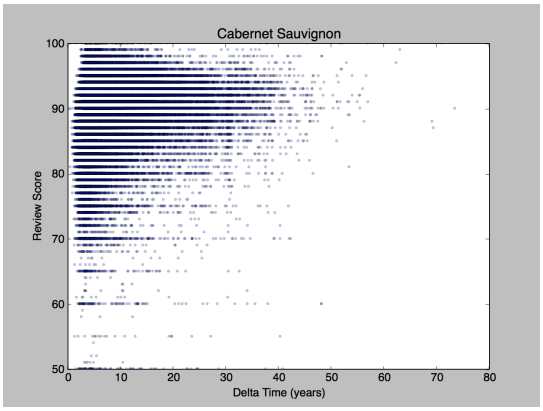

Figure 2: A graph of scores for a single variant.

in terms of styles. There turned out to be 752 different variants present in the data. Many of them were very obscure styles. There were only 79 that had more than 1000 reviews. These were the variants I chose to focus on. I knew having variants with so few reviews might cause some problems fitting a model, but I chose this threshold to see how robust my models were. The largest of the variants, 'Pinot Noir', had 202,714 reviews. This was plenty of data to try to identify a trend.

The following section describes the continuation of my data exploration as I formulated a task.

3. THE PREDICTIVE TASK

Having identified the primary wine variants in the data set, I next needed to determine if there was an interesting trend present. I started by grouping the data into its variants. I then decided to visually inspect the data for interesting formations. Figures 1 to 3 each show the score of a single wine variant graphed against the age of the wine. I used partially transparent points to better emphasize the concentration of data.

The figures had some amount of variation, but the general trends seemed the same. There was a period where there was a high spread within the scores, but gradually the concentration of lower scores moved upward. The top scores

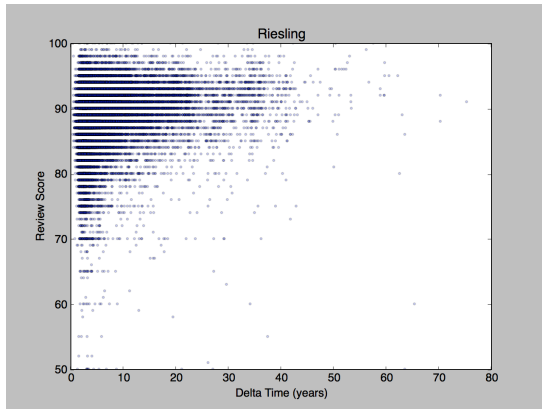

Figure 3: A graph of scores for a single variant.

remained high in this period, then started to sag a bit down around an average value that was lower than the top.

The graphs of the data, by variant, confirmed my suspicion that the affects of aging on wine does not follow a single consistent trend. It looked like there was an initial period of ramp up followed by a convergence to an average value.

My main interest was creating a model to find where this trend transition occurred. I also needed to be able to determine that my model was valid. To that end, I took the grouped data set and broke each group into a training and test set. I put half of the reviews in one set and half in the other. Noticing that the reviews were ordered by user, I put the odd entries in the training set and the even in the test.

To determine if my model was good, I used the mean squared error (MSE). I first created a few baseline measurements. From the training set, I determined the mean and created a linear regression solution using the wine age in the feature vector. The MSE of these models on the test data served as my baseline for comparison with my model. My model must be constructed in such a way that the transition point I am looking for is readily measurable.

4. RELATED WORK

It is important to also discuss the origin of the data set. This data was collected for use in [3]. The topic of the work was to create a model that took into account user experience level to model reviews. They leveraged the fact that there was a large set of time series of reviews by the same users. This allowed them to create an expertise model on a per user basis.

There were other similar data sets that were also used in [3]. There were two sets of beer reviews, one from BeerAdvocate and one from RateBeer. Additionally, there were two sets of Amazon reviews (fine foods and movies). This data was used in the same way as the Cellar Tracker data, to use a user expertise model to improve predictions.

Additionally, the beer data sets were also used in [2]. They leveraged the fact that the review scores were split into multiple parts. By analyzing the review text, they were able to

associate certain text with certain scores. This allowed them to use text analysis to predict missing scores.

There were naturally many other review based works (the Netflix challenge [1] came to mind), but my chosen task did not have similar goals. I needed to be able to identify the transition point between two trends. To do this, I needed to model these trends either independently or jointly. The best way to do this was through regression analysis.

The simplest form of regression analysis is linear regression [6]. In this type of analysis, you model the relation between an input feature vector and an output by finding a set of linear parameters Θ. The model is in the form: $y = \Theta X$, where X is the input feature and y is the output value. The value of Θ can be determined by solving a set of linear equations using linear algebra. The resulting values are linear parameters that minimize the MSE. This type of analysis was a useful tool in the deriving of my model.

Another type of regression I found might be useful was non-linear regression [5]. This type of analysis also solves for weights to fit a function to data. However, this input function does not need to be linear. This approach can be much more powerful as it can be used to model phenomenon that does not have a linear structure. The catch is that you must understand the data well enough to guess what class of function is a good fit. Solving for this type of model does not have a closed solution. The algorithms focus on an iterative approach that eventually converges to an optimal value. One such algorithm I ran across for this work was the Levenberg-Marquardt algorithm[4]. This algorithm uses a form of gradient decent to find a solution that minimizes the MSE for the function.

5. FEATURE IDENTIFICATION

I briefly mentioned some of the features I used in the earlier sections, but I shall go over them in more depth here. There are a few features that were important to this analysis.

The variant of the wine must be used to separate the different types. As mentioned above, I divided the data set into smaller sets grouped by variant. My motivation was that different wines will have different aging characteristics. It was certainly not expected that you would want to save a sauvignon blanc as long as you would a pinot noir. It would not be possible to extract any meaningful trend if all the data was grouped together. As seen in the figures above, the individual variants do allude to the trends I am looking for.

The review scores were the next necessary feature. The score represented the measure of quality. By assessing how the scores changed over time, I could find the optimum wine age I was looking for.

The final important feature was naturally the age of the wine. Each data item had a year associated with it. There was also a time stamp for the review. The time stamp was in a Linux format. I first converted this to a usable date and time. I treated the wine year as if it was bottled on the first day of the year. I used the difference between the time stamp and the wine year as the wine age. To add a little more precision, I also used the month. It did not seem reasonable to use any more precision, given how coarsely the bottling date was estimated. I found that the wine age varied from 0-290 years with an average of 5.986 years.

As a small note, the data set contained a few time stamp errors. There were a couple of time stamps that had negative values. These values were present in the raw data set. It was most likely the result of an error during collection. There were also a couple of items that had times the preceded the wine year. These were thrown out.

These three classes of features together were what was needed to derive a model for my task.

6. MODEL DERIVATION

In this section, I describe the different models I tried to accomplish my task. All of these models were able to run in reasonable time on the full data set. Not all these models were successful. This will only describe the reasoning and setup of the models. See the Results section for the model evaluations.

6.1 Binned Linear Regression

My first attempt at a model was an application of linear regression. As stated above, I used a simple linear regression model as part of my baseline. That version used a feature vector in the form of [1.0,age]. For this model I used a longer feature vector where each entry represented an age bin. For each entry in the vector, the value would be zero unless the age of the wine for that data item fell into that entry's bin.

The first 25 feature entries were each one year bins. This was followed by 5 five year bins and a final bin for anything greater than 50 years. This was indeed a lot of entries per feature and there was definitely a risk of over fitting (as will be shown later). A solution might have been to create fewer, larger bins. However, even if this would have yielded some results it would have been at too coarse a granularity to be useful for my task.

The idea was to examine the resulting theta vector from the solution and look for either the maximum value of theta or where there was a relative maximum. This critical point might be a potential solution to my task.

6.2 Two Part Linear Regression

This next model was inspired by wanting to find the location where two data trends diverge. This model was also powered by linear regression, but it worked differently than either of the other implementations.

The idea was to form a model that used one linear regression solution for the first part of the data and a second linear regression solution for the rest. The assumption was that there was some way to divide the data set (on the time axis) that splits it into two separate models. The feature vector was the same as for the baseline model, [1.0,age]. The main trick was finding out the optimal point to split the data.

I started by marching through time (the age of the wine). At each interval I would split the data set into two, with

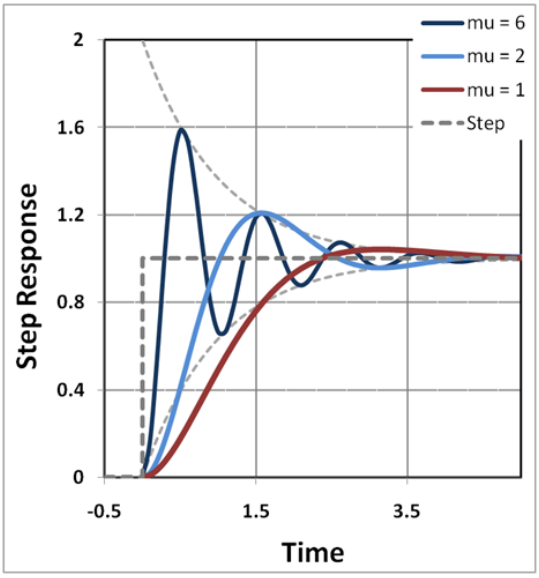

Figure 4: A set of dampened oscillation functions, provided by wikipedia.org.

everything below the current age in the lower set and everything above the current age in the upper set. I would then solve for the theta weight vector for both the lower and upper sets. With these vectors I would calculate the MSE for the training set.

For my first attempt at this model, I would keep track of the calculated MSEs as I stepped through time and would stop when I found a minima, using this as my solution. My initial assumption was that there would only be one. I quickly discovered that this was not the case. My next attempt at the model stepped through and located the global minima.

For my final attempt at this model, I recorded all the local minima as I stepped through time. In order to figure out which of these solutions was the best, I used the one with the least MSE for the testing set.

The motivation for this model was that it created a clear division between the two trends in the data. The solution to my task (finding an optimal time to drink the wine variant) would be the wine age used to divide the data set for the given solution.

6.3 Curve Fitting

Additionally, I wanted to see if I could get a better fit by trying to use a non-linear regression solution. A solution of this type tries to fit a specified non-linear function. For my solver I used the curve fit function from the scipy package. This function allows you to provide a fitting function and an initial set of parameter and uses the Levenberg-Marquardt algorithm to find an optimal solution. I decided to try a number of different functions to fit.

My initial thought was to use a function from physics. The solution form I was looking for reminded me of a dampened oscillation function. Figure 4 shows a set of oscillation functions with various dampening parameters. The though was that this function type would allow for the presence of a peak and then converge towards a lower value. It was hoped that the location of this peak would be the optimal wine age I was looking for.

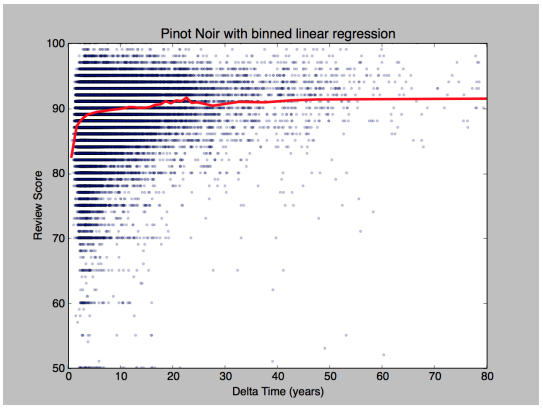

Figure 5: A binned linear regression solution.

The function was of the form:

$$\exp(-\gamma * x) * A * \cos(\phi * x - \alpha) + \beta$$

The other concept I was interested in trying was similar to my two part linear regression model. I used the curve fit function to fit piecewise functions. I first tried a combination of a quadratic function and a linear function. The function was designed so that it would be continuous. Both functions shared the same y intercept and were shifted (by the solver) an equal amount. The idea was to have a function with a quadratic ramp up that at some point became linear. The point where the functions connected would be considered the point of the trend shift and, thus, the optimal wine age.

The other function I tried was also piecewise. This one, however, started as a linear function and was connect to a logarithmic function. The thought was that the connection between these two functions would represent the start of the trend shift, which would be the point I was looking for.

7. EVALUATION AND RESULTS

The above models were trained and tested. The following subsections describe the model evaluations.

7.1 Binned Linear Regression

This was the first model I attempted. As stated above, it used a feature vector of length 31 to solve against the wine scores. The results for three wine variants are shown in Figures 5, 6 and 7. The resulting solutions do not present any consistent trends. As an artifact of having such small bins, there was a relatively high degree of variance between adjacent bins. I attempted to find both the highest theta component and the first relative maxima, but neither proved to be a solution point to my task.

Additionally, Figure 8 shows the MSE of the binned linear solutions for all 79 wine variants compared with the baseline

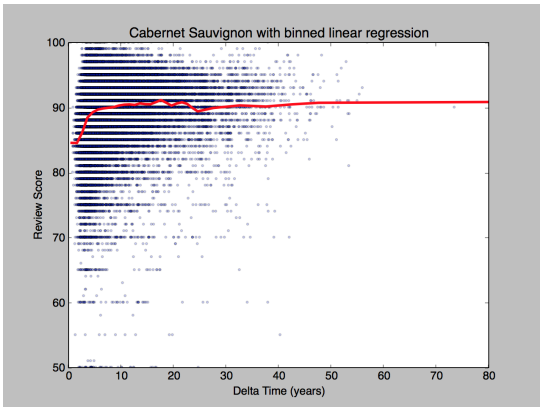

Figure 6: A binned linear regression solution.

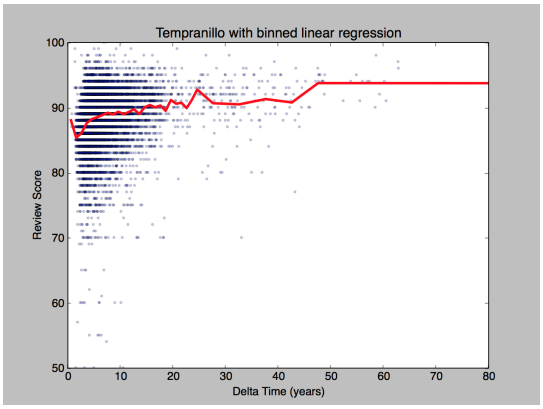

Figure 7: A binned linear regression solution.

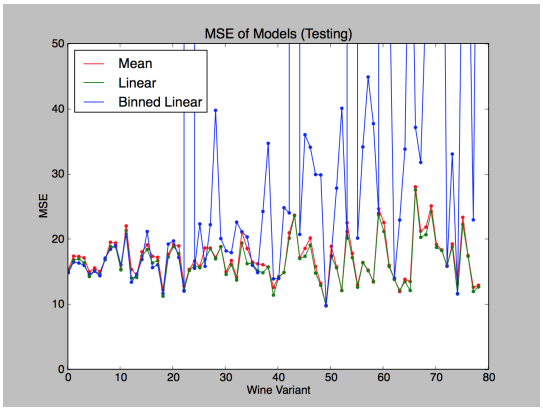

Figure 8: MSE comparison for binned linear model.

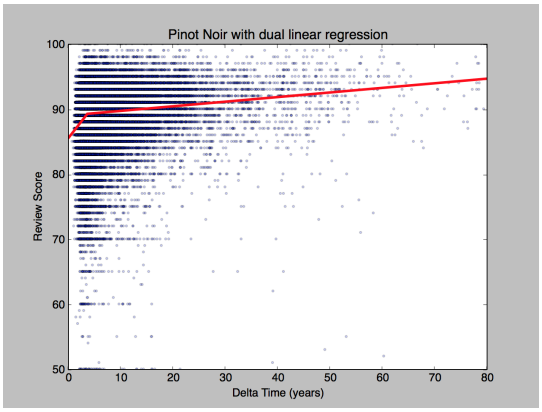

Figure 9: A two part linear regression solution.

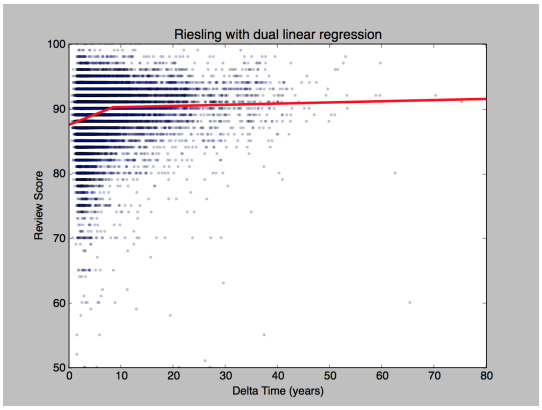

Figure 10: A two part linear regression solution.

mean and linear models on the testing data set. The variants are arranged from most reviews (left) to least reviews (right). The binned linear regression solution is clearly very over fitted to the training data as the size of the data sets decrease. This model did not prove useful in the accomplishment of my task.

7.2 Two Part Linear Regression

As the previous model did not meet my needs, I moved on to try a two piece linear regression model. Figures 9 and 10 show solutions for two wine variants. The figures clearly show the division of two linear trends within the data.

As stated above, I tried three different methods to determine the optimal division of the data sets. I used the first minima, the global minima, and the best of all minima on the test set. It was now necessary to determine if there models were a good fit. Figure 11 shows a graph of the wine variants' MSE on the testing set for the three methods and the two baseline methods (mean and single linear models). The models do indeed beat the baseline MSEs in almost all cases. To clarify the difference, Figures 12 and 13 graph just the difference between the three two part models and the mean and linear baselines.

It is important to note that the three models produced a similar MSE for the earlier variants and started to vary more

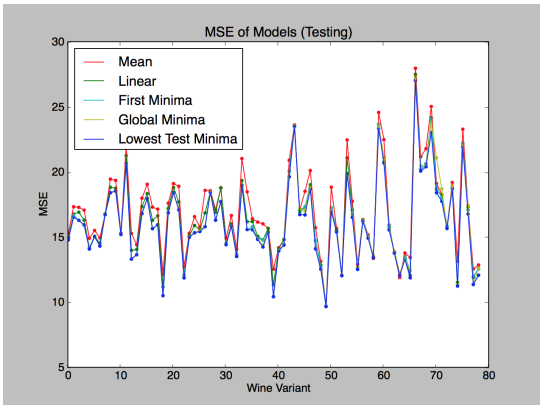

Figure 11: Graph of MSE of two part linear models with baselines.

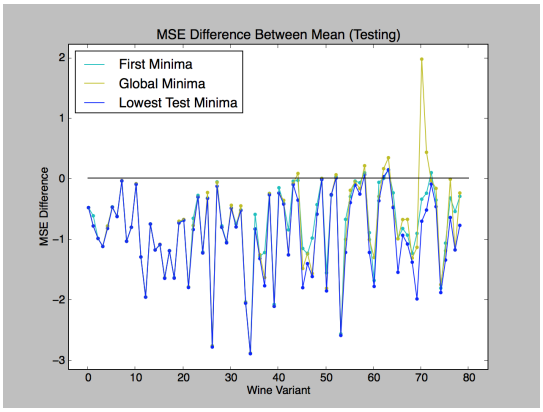

Figure 12: Graph comparing two part model with mean baseline.

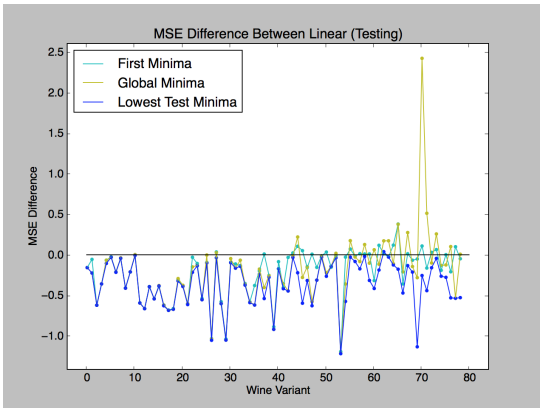

Figure 13: Graph comparing two part model with linear baseline.

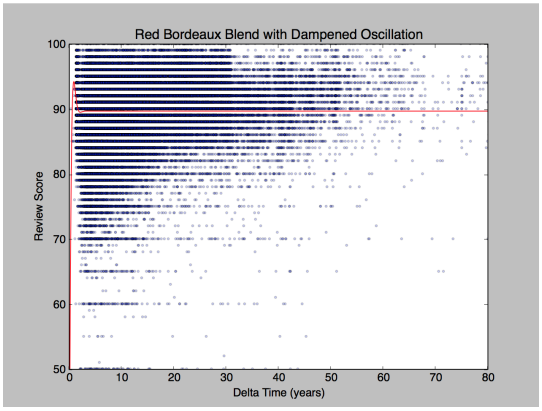

Figure 14: Graph with fitted dampened oscillation.

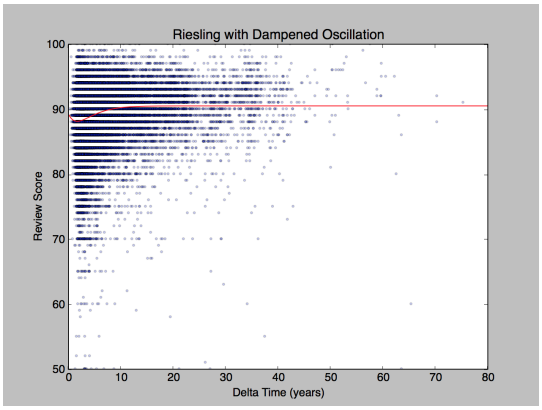

Figure 15: Graph with fitted dampened oscillation.

for the later ones. Since the variants were arranged from most data items to least data items, this showed that a cutoff threshold of 1000 (500 test, 500 train) was probably too low for these models and resulted in a lower quality fit. Even so, the best of the three methods still produced a fit that beats the baselines for almost all variants.

This model appeared to be a viable solution. It was shown to fit the testing data better than the baselines, reducing the probability of over fitting. What was also good about this model was that the point of division between the two linear pieces was very clear. This division age was a good candidate for what I was attempting to find.

7.3 Curve Fitting

The two part linear regression model appeared to give me a useful result, but I wanted to try a non-linear model as well. I did this as a check to make sure my linear assumption was not too ridged. To this end, I tried the three models: the dampened oscillator, the quadratic-linear piecewise function, and the linear-log piecewise function.

I first attempted to use a dampened oscillation function. The problem I ran into was that the function had too many degrees of freedom. This resulted in odd fittings that, though locally optimal, did not reflect the trend I was looking for. Even after fixing some of the parameters, the best results I

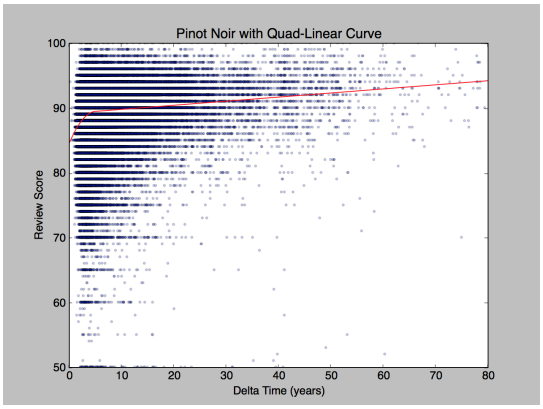

Figure 16: Graph with fitted quadratic-linear function.

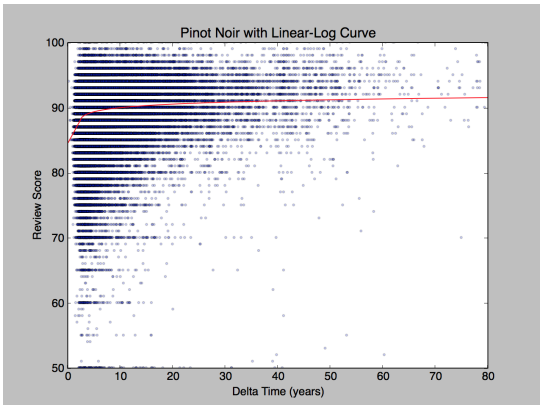

Figure 17: Graph with fitted quadratic-linear function.

got are reflected in Figures 14 and 15. My goal was to use the function maximum to find the optimal wine age. However, these graphs do not fit in a way that allows that. This was not the function I was looking for.

I next tried to fit two different piecewise functions. Figure 16 shows a graph with a quadratic-linear function fitted to it. Figure 17 shows a graph with a fitted linear-log function. Both functions seem to fit the data trend fairly well.

To determine if the models were reasonable, I compared them to the baseline values on the testing set. Figure 18 shows the two models compared to the linear regression baseline over the 79 wine variants. With the exception of two variants, both these models beat the baseline linear MSE. The two outliers are again likely caused by over fitting on these smaller data sets. The graph shows that though both models are valid, the linear-log fitting is better than the quadratic-linear fitting. I used the linear-log model as my potential solution from the curve fittings.

7.4 The Best Solution

I now had two potential models to use for my task. I needed to determine which was best suited. I first checked to see if there was a dramatic MSE difference between the two mod-

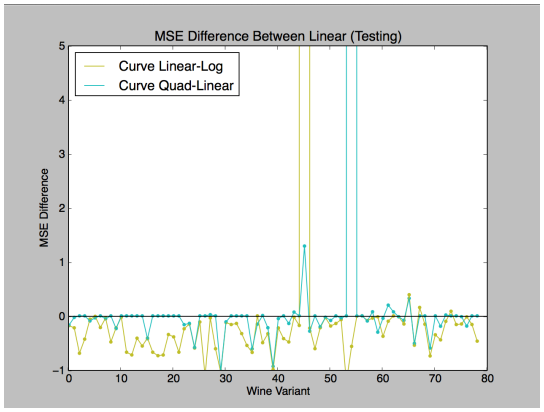

Figure 18: Comparison of curve fit MSE to linear baseline.

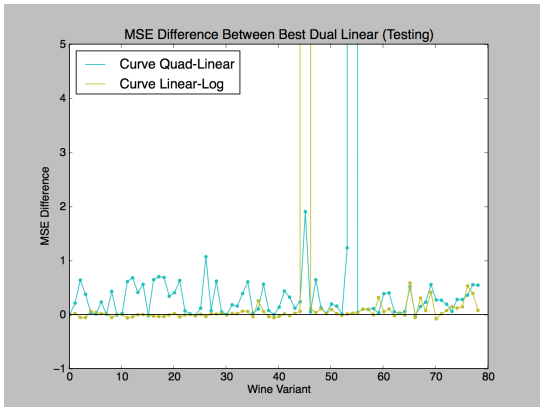

Figure 19: Comparison of linear-log MSE to best two part linear regression.

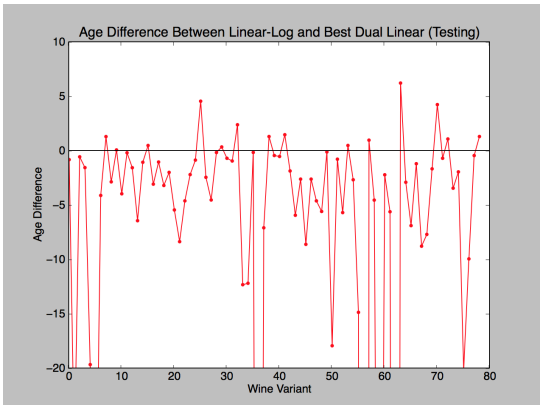

Figure 20: Comparison of linear-log optimal age to best two part linear regression.

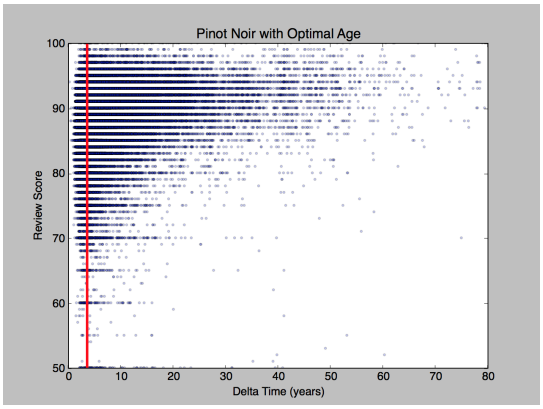

Figure 21: Graph with Optimal Age.

els on the test set. Figure 19 shows the difference between the curve fit MSEs and the best two part linear regression fitting. The linear-log model is shown to have a reasonably similar MSE to the two part linear model.

It was then a matter of determining which of these models would allow me to accomplish my task. Using the point of discontinuity as the optimal wine age, I decided to compare the models' solutions. Figure 20 shows the difference between the optimal curve fit value and the optimal two part linear value. There is a very large difference between most of the values.

To explain this difference, I went back to the graph of the linear-log curve fit solution (Figure 17). I realized that even after the function transition, the log function still increases at a reasonable rate before leveling off. This largely increasing portion of the log function was causing an underestimation of the optimal age value. The age at the start of the log trend was being used, but the logical solution would really be the point where the log sufficiently slowed its increase. The point I used as my solution was therefore not the correct one. Also, finding an actual solution from this model was not a strait forward task. Though the model may be a good fit to the data, it did not help me accomplish my task.

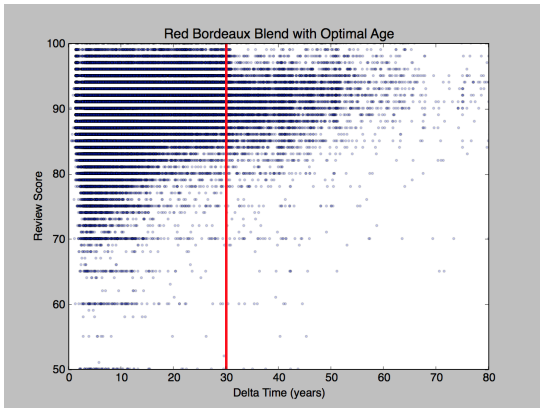

Figure 22: Graph with Optimal Age.

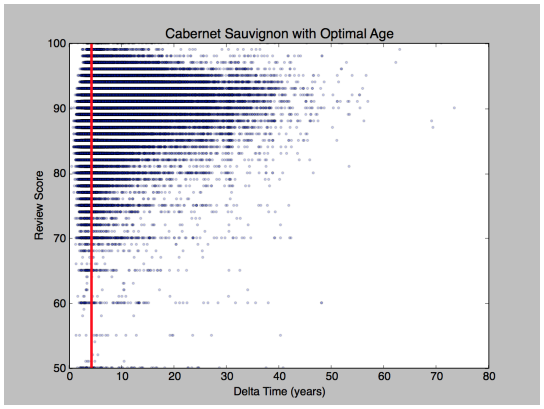

Figure 23: Graph with Optimal Age.

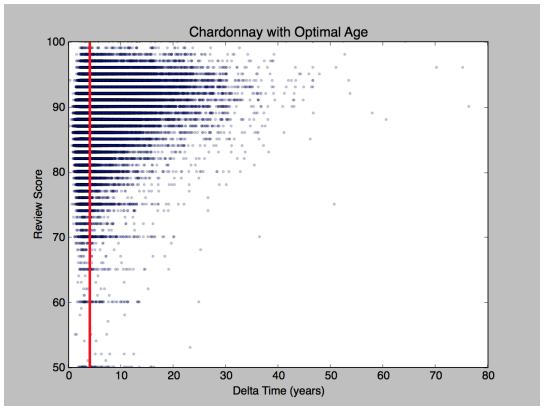

Figure 24: Graph with Optimal Age.

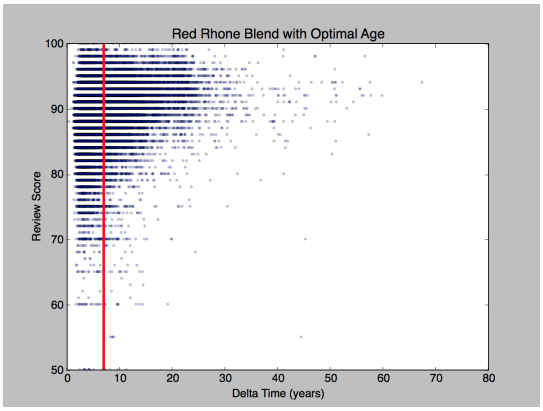

Figure 25: Graph with Optimal Age.

The two part linear regression model, however, fit the data and had a very clear solution point. For those reasons, I chose this to be the model for my solution. To provide visual validation to my solution, Figures 21 through 25 are the top five wine variants with their optimal age indicated. For all of these graphs the optimal age was found to be the point right as the upper end of the ratings began to decline. This matched my expectations of a solution.

Additionally, I have included a table at the end of the paper which contains the optimal age for the top 20 wine variants (by review abundance).

7.5 Implications

Trend	Proportion (79)	Proportion (40)
+/+ 2>1	0.088608	0.000000
+/+ 2<1	0.658228	0.750000
+/-	0.177215	0.225000
-/+	0.050633	0.025000
-/-	0.025316	0.000000

: Theta trend for two part linear model.

It is also worth discussing some implications of the model. The solution consists of two theta vectors, each containing a y intercept and a slope. By looking at the slope values, we can infer some information on the trends in the data. If both are increasing, and the second is larger than the first, the wine will get better faster after it passes the optimal age. If the second is smaller, there will be diminishing returns after the optimal age. If the first slope is positive and the second is negative, quality will begin to decline after the optimal age. If the opposite, quality will start improving. In the event both slopes are negative, the quality will perpetually decline.

The above table shows the proportion of the variants that fall into each of these categories. There is one column for the top 79 variants and another for the top 40. Since the lower variants have less data items, it gives an idea of how tolerant the model is to a smaller amount of training data. The trends seem more reasonable when just looking at the 40 larger data sets. With only one exception, all of those data sets either show an increase that begins to have diminishing returns or an increase that leads into a decline. This matched my intuition from the initial data exploration. It also showed that the odder trend combinations likely came as a result of fitting to too little data.

8. CONCLUSION

This project has been a complete lesson in data mining. I found a data set and, through exploration, identified a task to examine. Using the features of the data, I derived and tested models to try to accomplish my goal.

The cellar tracker data set provided an interesting challenge. Through the features present, it gave me the opportunity to use peoples' tastes (through reviews) to attempt to characterize the effect of age on the quality of wine. I created several models in an attempt to find one that represented the data trends I observed. Through testing and validation, I was able to eliminate the models that were unfit or did not suit themselves to my task. In the end I derived a model that, given a sufficient amount of data, produced a plausible solution to the optimal age to drink a wine variant.

Variant	Optimal Age
Pinot Noir	3.50
Red Bordeaux Blend	30.00
Cabernet Sauvignon	4.25
Chardonnay	4.00
Red Rhone Blend	7.00
Syrah	17.00
Riesling	8.25
Zinfandel	3.5
Red Blend	6.00
Shiraz	8.00
Sauvignon Blanc	7.5
Merlot	3.75
Sangiovese	8.25
Nebbiolo	10.75
Tempranillo	4.50
Malbec	3.25
millon-Sauvignon Blanc Blend	7.00
Sangiovese Blend	5.50
Champagne Blend	11.00
Grenache Blend, Grenache	6.25

: Optimal age for top 20 variants.

9. REFERENCES

[1] J. Bennett and S. Lanning. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35, 2007.

[2] J. McAuley, J. Leskovec, and D. Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1020–1025. IEEE, 2012.

[3] J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908. International World Wide Web Conferences Steering Committee, 2013.

[4] J. J. Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978.

[5] H. J. Motulsky and L. A. Ransnas. Fitting curves to data using nonlinear regression: a practical and nonmathematical review. *The FASEB journal*, 1(5):365–374, 1987.

[6] G. A. Seber and A. J. Lee. *Linear regression analysis*, volume 936. John Wiley & Sons, 2012.