Questions

December 4, 2018

1 Arthena Data Science Challenge - Questions

Author: Manksh Gupta

---

1. Which features are most important for your model (for either your individual or pooled model)? Are there any features that surprised you? Given more data, describe other features you feel would be useful in improving the accuracy of your model.

Ans 1)

I am describing the features that were important for the pooled model. The features were very similar in the individual model too. However, the pooled model has more features. The most important features turn out to be:

`'estimate_high': Estimated High by the Auction House 'high_low_diff': Difference betwen the high and low estimate(New created Feature) 'death_auction_diff':Diff between the death of artist and the auction year(New Feature) 'execution_birth_diff': Diff between work execution and artist birth year(New Feature) 'execution_auction_diff: Diff between work execution and auction year`

These were expected time based features that turned out to be important, I made the features expecting these to be important and they so turned out to be. Some Surprises were:

`Dummy Variables for the Auction houses`

I did not, apriori, expect this to be important, however, as seen later in the document, different auctions houses do seem to have significant difference in final pricing of the artwork.

I was also surprised that the `Dummy artist name features` were not found to be the most important, it may be because the three artists I analysed in the model were are pretty popular and the works are selling for similar prices.

It also seems like the model that I am finally using is not very good at predicting artwork of very low value, however, for mid and high values the model is predicting really well. Below a certain threshold, the model just predicts one value for all artwork. This can definiutely be optimized given more time.

More data/Features: I would really have liked to have the image of the artwork itself - One can derive a lot of features about the image(brightness, color use, contrast, if there are faces or not etc.). I beleive that these features can significantly impact the hammer prices of the artwork and thereby increase the accuracy of the model. Th eartwork after all, is what defines why people like/do not like a particular painting and developing some kind of model to identify that would be invaluable!

---

2. Suppose we're tasked with evaluating the price of a work where we're given an incomplete set of features. Say, we're given the artist name, type of work, and size, but we know nothing else. How would you go about estimating the `hammer_price` at an auction tomorrow?

Ans 2)

Assuming we have past data about the same artist and that data has all the features. I would first subset the past data with the name of the artist that i want to predict for. Then, for that artist I would look at his previous work and subset based on the type of work that is the same as the one that we would like to predict the price for, finally, I would group by size and look at the median price for the size in the past data. I would compare this to the data point that I have and reoprt that mean value. Now, If we don't have any past info about the artist, I would just look at all the data we have and subset on type of work and report the median for the size thats closest to the one I have. Finally, If we have no past data at all, any guess is as good at mine, I would probably try scraping some information from online sources to make an 'educated' guess.

---

3. You're starting an art collection. What medium do you purchase and why?

Ans 3)

I am basing this answer on the data for all three artists that I have access to, the answer can be very different if given access to more data. According to the given data 'paintings' seem to be the medium that fetches the highest prices. Paintings are also the 3rd most auctioned type of work(after print and decorative arts). Both Print and Decorative arts fetch a significantly lower price than paintings. Even after paintings are adjusted for outliers(removed paintings over $ 5 Million) Paintings are still highre but pastel and mixed media are close. Bottom line is that this would depend on how much money I am willing to spend on a piece of art. However, any kind of painting is sold far higher than sculptures, at least by these 3 artists. If we include some artist who is more knows for sculptures, the decisions might change.

---

4. Assume we care much more about not overpredicting `hammer_price` than we do about underpredicting the price. Describe how you would go about changing your solution in terms of the model, objective function, etc.

Ans 4)

Since we care more about not overpredicting(its a loss for us), I would change the loss/infromation functions that the algorithms are optimizing to tackle the problem. We are generally optimizing mean squared error(mse), I would choose a new version of 'mse' as follows:

$$L = \frac{1}{n}\sum_i (Y_i - \hat{Y}_i)^2 + \lambda(\hat{Y}_i - Y_i)$$

Now, the second part of the above function is where the interest is. If we predict greater than actual value, the part is positive and thus the loss is increased, however, if the predictions are lower, this value is negative and thus the overall loss will decrease. This will force us ot predict

smaller values. This function is still convex and thus, it is still easy to optimize using out of the box techniques such as Gradient descent.

However, there is one concern with this loss, we can get a degenerate solution where we predict 0 all the time and this may give us a very small loss. This is where the lambda comes in, we must choose 'an appropriate' λ, I susepct something like 0.01, would be an appropriate value, but this can only be decided after more experimentatioon.

---

5. How do auction houses like Christie's and Sotheby's influence prices of works?

Ans 5)

In the data, there are 3 auction houses, Christies, Sootheby's and Phillips. Phillips has the lowest avg hammer price values and also the least number of artwork sold. Christies and Sotheby's look to be very close in terms of pricing and artwork sold. However, I did a t-test to determine if the difference is significant, and it turns out that artwork at Sotheby's is significantly higher than what it is at Christie's. It seems like paintings auctioning at Sotheby's sell at a higher price on average.

---

6. Given more time, but no new features, how much do you think you could improve the accuracy of your models by? Why?

Ans 6)

I thought of trying this but could not finish due to time, esentially, I would bucket the hammer price in a range and reduce the problem to a classification task. Then, after classification, given each class, I would look at the distribution of prices and the features associated. I believe that this would reduce uncertainity. Also, It may be more useful to predict a range rather than predict a value(as given in part 2.2).

I would probably have tried some kind of deep model just to see where it gets me, this would make it really hard to interpret results and understand feature importances. Depending on the goal of the company and what the KPI's are, this may or may not be a good idea. If we only care about the final result and don't care about what the features are contributiing to, this may be a good idea as deep learning models have shown to outperform other techniques. However, I do feel that given the scope and the final questions, the company does care about interpretability and thus, it may not be a good idea.

---

7. Was this fun? Which sections / questions were the most difficult and which were the easiest?

Ans 7)

The whole challenge was definitely fun, I really enjoyed the problem statement and the setting of the pronblem.

I feel the most difficult task was cleaning the data and designing features that have predictive power. The data was extremely messy(as always is in the real world), and there were a lot of missing/-1 values that had to be imputed. Feature engineering and data cleaning took up most of my time in the challenge, the modeling part after this was relatively straightforward. Having

theoretical understanding of the algorithms that I used, it was relatively easy to do the modeling once the right features were in place.

The part about the distribution of predictions was very interesting, I wish I had more time to spend on that and have a good solution for you guys. I did have a preliminary solution however, I did not think that it was good enough to be presented. I was bound with a final and 2 projects during the week and thats why I couldn't make progress on that.

All in all, it was a really interesting machine learning task where I did not feel that I was actually doing a take home 'task' but more like a fun problem that I tried to solve over the weekend.
