A Look at Cataloguing Artworks

CS 412: Final Project

Anya Ellis

12/10/2021

1 Introduction

I would like to start off by reflecting on my experience which I had in this class, "CS 412: Introduction to Machine Learning". I learned how to apply machine learning algorithms from scratch in python, as well as some algorithms theories that I did not know at the start of the coursework. I was particularly unfamiliar with support vector machines and some of the theory behind neural networks. I don't feel confident in either of these areas, but that's to be expected as brushing up old stuff (like k-means or k-nearest neighbors) is very different from obtaining mastery at new stuff. My familiarity with the "numpy" library has also increased. I wish to spend more time learning about neural networks and maybe nanocubes, but this, I believe, is just for personal satisfaction. Now, we move on to my opinions on machine learning in general.

Cathy O'Neal wrote a pivotal book titled "Weapons of Math Destruction". It detailed her experience and knowledge related to machine learning and data science and the sometimes willful or unintentional consequences of biased computer science. This book, itself, is only a drop in the expanding literature surrounding computer science ethics and fairness. It is within this field that most of my opinions on machine learning are viewed from, as this field is also my area of research. So, to expound upon my views of machine learning, I believe the field is not only making new theoretical and technical advances, but also expanding on new and old ethical frameworks to go with it. From this standpoint, we can move on to the topic of my research for the class.

2 Problem & Purpose

Representation of art is known to be very problematic in terms exhibiting certain groups over others to the disregard and degradation of the artwork while in inappropriate storage facilities.

In this study, we look to see if there is any relation between if the artwork is catalogued and the other attributes which it might have. Cataloguing artworks is extremely important as it not only showcases the importance which the museum might place on the artwork, but it also essential for risk management, research and exhibition development.

We look at the Museum of Modern Art's open database to make some connections within their facilities. The Museum of Modern Art (commonly known as MoMA) is located in New York city and is one of the biggest museums and galleries within the United States. MoMA has over 200,000 artworks, and with this size, it make sense that a portion of its collection is not catalogued. But, we have to ask ourselves which portion? And is there a trend?

3 Data

The MoMA has over 90,000 artworks in its database with 29 attributes. We list them the original at Table 5 and the cleaned attributes at Table 6. Twelve being a numeric type (includes date types) and seventeen being a string type. Out of the attributes, ... were dropped for either being repetitive, not essential or difficult to clean. For example, "Artist" and "Title" are basically the same as "ConstituentID" and "ObjectID" (i.e. their respective unique int ids). The URL related attributes were seen as not essential for the purpose of this project. "ArtistBio" and "Nationality" were seen as difficult to correctly clean, for the limited time of the project. Moving on, we will explain the attributes that were looked at, in terms of their comparable groups.

ArtistID: This attribute is a tuple int type within the data; each int corresponds with an unique artist. We replaced missing values with -1 and pulled the first listed int to create the attribute "FirstArtistListed". This was done to simplify the cleaning process, but still include some artist unique ids. This is important on the assumption that a work by Pablo Picasso might be more highly favored than a work by Hilma af Klint or Liu Xiaodong.

Numbers Of Details: To track the artist ids left out, we create three new attributes: "NumArtists", "NumFemales", and "NumMales". The first attribute created takes the number of artist based on

the tuples in ”ArtistID”, and the following attributes take the info from ”Gender” by counting the occurrences of the words ”Females” and ”Males”. We see these attributes as comparable as they count the number of artists involved in the particular artwork in different novel ways.

Dimensionality: These attributes describe the artwork in terms of size or duration. They are all numeric types, and for the purpose of this paper, we change all missing values to 0.

Descriptive Details: All these attributes were string types that describe the work in terms of where its from, what its made out, where it is, or what it is. These are: ”CreditLine”, ”Medium”, ”Department”, and ”Classification”, respectively. All of these attributes were converted to categorical numerics, except for medium which was transformed to bits for phrases of importance based on top 60 mediums. For example, ”Gelatin silver print” and ”Albumen silver print” are both prints, and both use silver print process (a photography process). However, they use different soluble solution for the print process (i.e. gelatin or albumen). Therefore, we can make four medium categories from this problem: ”print”, ”silver print”, ”gelatin”, ”albumen”.

Time-Related Details: There are three attributes under what we call the time-related details: ”DateAcquired”, ”TimeStarted” and ”TimeFinished.” The ”DateAcquired” refers to the acquisition date for the museum, and the others refer to when the artist started or finished that particular work. The ”DateAcquired” is transformed into a period object as its in datetime format, going back before 1600s. This is then converted to an int type, while ”TimeStarted” and ”TimeFinished” are directly converted to an int type.

4 Method

We started our research through exploring the relation between the classes on three attributes: ”AcquiredDate”, ”Department”, and ”CreditLine”. Through this, we found that ”AcquiredDate” had little difference between top days of acquisition and the classes. The departments on the other hand had 2 departments which had no works catalogued. The credit line showcased a need for a more complex analysis, as certain credit lines had different names with different time periods. The Judith Rothschild Foundation has made multiple donations to the museum under different names. We did notice a trend though, that certain creditlines were more likely to be not catalogued or catalogued. However, since creditlines has over 7,000 different varieties, we believe this is to be expected, but we can’t say if this can be learned from reliably in a supervised manner. We showcase some of these as graphs in Section

8.2, titled ”Exploratory Data Analysis.”

At this point, we moved on to perform two different supervised learning classifiers, eventually under two different sets of features. Are first set of features were:

> ”CreditLine”, ”Classification”, ”Department”, ”DateAcquired”, ”Circumference (cm)”, ”Depth (cm)”, ”Diameter (cm)”, ”Height (cm)”, ”Length (cm)”, ”Weight (kg)”, ”Width (cm)”, ”Seat Height (cm)”, ”Duration (sec.)”, ”NumMales”, ”NumFemales’, ”NumArtists”, ”FirstArtistListed.”

And, To be honest, we set about splitting the features, because we forgot to included them all the first time around, but it did evaluate to some interesting analysis. For this reason, we showcase them.

Our chosen classifiers were on the idea that we would use all the features, and so the performance between the first feature set wasn't really expected as we didn't really think that we would be doing a feature sub-selection. In any case, the chosen classifiers were: Naive Bayes and Logistic Regression.

Naive Bayes is a generative approach to classifying, and works best on detection through phrases. Since we split the medium attribute into top phrases detected, we thought that Naive Bayes would perform somewhat well. We showcase the equation for Naive Bayes below, where C_k stands for class k, and x is the input.

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} \tag{1}$$

Logistic Regression is a discriminative-based classifying approach which also relies on the idea that the attributes are independent from each other in some way like Naive Bayes. We showcase the Logistic Regression formula as well, where w stands for weights.

$$P(C_k|x) = \frac{e^{w_k \cdot x}}{\sum_j e^{w_j \cdot x}} \tag{2}$$

We went with these two somewhat opposing models as we had many attributes other than medium types that could have had the most indication that an artwork is catalogued.

We used the ”scikit-learn” library to create our classifiers, and performed a three fold cross-validation. The cross validation took the form of splitting the data randomly into three group, and using one group for testing and the other two were joined together for training. This was repeated

until each group could be a testing group. In addition, we tracked the time for fitting the training data. We then repeated this for the full feature set listed at Table 6.

5 Observations & Results

To start off, we make general comparisons between the two models. Overall, the Naive Bayes had a faster fit time as shown in Table 1, below. On the hand, the Logistic Regression classifier performed overwhelming better in terms of accuracy.

fold/ms	Partial Features (LR)	Partial Features (NB)	Full Features (LR)	Full Features (NB)
Fold1	915	33.5	2270	411
Fold2	1300	33.3	2060	347
Fold3	950	33	2130	329

Table 1: The fit times between the different folds

However, the added feature sets did not help either classifier gain a better accuracy. The average accuracy between classifiers was 64% and 39%, respectively, and this did not change with the expanded feature set 2. As said earlier our initial assumption was Naive Bayes would perform adequately because we have over 50 attributes (i.e. those based on medium) that tracked phrases.

fold/accuracy	Partial Features (LR)	Partial Features (NB)	Full Features (LR)	Full Features (NB)
Fold1	64%	39%	64%	39%
Fold2	64%	39%	64%	39%
Fold3	64%	39%	64%	39%

Table 2: The prediction accuracy between the different folds

Though, we cannot say at this time why the accuracy did not change between the feature sets, we did unveil some interesting findings. When looking at the weights that the Logistic Regression model placed between the different feature sets, we see interesting parallels and divergences on the effect that a particular weight might have had. We define *effect of a weight* to be the absolute value of the weight.

5.1 Feature Set 1: Logistic

The first striking observation that we look at are from our initial feature set. The Logistic Regression classifier had declared no negative weights. This means everything leans towards the positive class (i.e. being catalogued). Our next observation is that out of the dimensionality attributes, height &

width have the most effect. On the assumption that most artworks will have a height/width, this makes sense. We also saw that the number of male artists had more of an effect than the number of females artist. This could just be related to the overwhelming number of male artists in the collection.

Then we see that the classification of the artwork (i.e. painting, sculpture, etc) had 26 times more effect than the Department (i.e. “Painting & Sculpture”, “Film”, etc). And this leads us to the question, since classification is a sub-field of department in a sense, can we see this same trend if medium, a sub-field of classification, is added to the feature set?

5.2 Feature Set 2: Logistic

This leads us to our evaluations of the full feature set. The Logistic Regression classifier now has multiple negative weights, some of them being: ”Department”, ”Classification”, ”NumMales”, and ”TimeFinished”. This is a stark contrast to the weight evaluation in our original feature set. And, we continue to see stark contrasts between the two feature sets.

The ”Department” attribute now has 1.6 times more weight than ”Classification.” This can be seen in Table 3. A lot of the ”Classification” attribute’s weight was spread among the medium types where eight specific medium types achieved comparable weights to the ”Classification” attribute. They are:

> portfolio”, ”black”, ”lithograph”, ”drypoint”, ”pencil”, ”illustrated book”, ”paper”, and ”color”

We define comparable in this case as them having absolute weights between 0.0001 and 0.0004. Though these weights have small values, we can expect that with the abundance of attributes. The weights between the first feature set and the second are of different magnitude but can be comparable in terms of ratio between elements in their comparable groups.

For this reason, we concluded that the number of female artists has more of an effect than the number of males artist when these additional attributes are added. As said earlier, the number of males now has a negative weight and goes in the opposite direction of its comparable counterpoint. Another two attributes that showcases this correlation through opposites is ”TimeStarted” and ”TimeFinished”. They literally only have a difference of less than 0.00003 between each others magnitudes. This is an interesting finding as most years which an artist started a work is also the year which they finished it. Another interesting fact is that the bias stayed the same between both feature sets.

Attributes/Weights	Partial Features	Full Features
CreditLine	0.0000792228492692299	0.000088948261338642
Classification	0.000354249048608818	-0.000392030963468615
Department	0.0000134262800344238	-0.000664225236220298
Height (cm)	0.00774078024853215	0.00968727642749201
Length (cm)	0.000335741923731944	0.00220809501978425
Width (cm)	0.00778873570200354	0.00661838465340963
NumMales	0.0000642569442557297	-0.000100976227707734
NumFemales	0.0000440280718731516	0.000313042311912789
NumArtists	0.0000929067930057628	0.0000677136714064633
FirstArtistListed	0.00000193503852215108	0.00000215290666241975

Table 3: A selection of weights by feature set

5.3 Naive Bayes

As said previously, the Naive Bayes classifier did abhorrently worst than its counter classifier, but what did we learn from it? During the process, we learned the variance & mean of all the attributes. And this gave a better idea of the spread of the attributes beyond our initial exploration of the data. We noticed through this that the duration of a film had stark contrast between the classes. A catalogued film had an average duration of around 63 seconds and a not catalogued film had around 28 seconds. Another interesting insight was that medium types with the most weight also had starkest averages between group. Some of these insights can be seen from Table 4, below.

	”Class 0: Variance”	”Class 0: Mean”	”Class 1: Variance”	”Class 1: Mean”
”DateAcquired”	37549060564.1296	-52278.6001104646	17330885144.6098	-18164.4687442975
”Height (cm)”	629.878315913411	24.3761401591057	3260.91696229464	37.8703745889191
”Length (cm)”	30.438174695765	0.077101334115042	1007.58146986752	0.712889210524184
”Width (cm)”	1197.57641023223	24.4395240891918	5783.78440458517	38.1952892915496
”Duration (sec.)”	511128.487660909	27.915567018475	465257391.294589	136.509536673519
”portfolio”	24.8603211366383	0.026482579861738	24.9095875731256	0.081727467638129
”screenprint”	24.8563554024483	0.022313703167881	24.8736010766537	0.040719474406435
”chromogenic”	24.8531764736892	0.018997523338106	24.8536026669663	0.019440611439567
”black”	24.9343862979825	0.112505069970958	24.8458850610186	0.01147683017981
”lithograph”	24.9540005125847	0.138698558962618	24.9835581614895	0.182224500679839
”drypoint”	24.8616357825451	0.027872686852813	24.8802468773933	0.048012095817111
”TimeStarted”	1156.24044593182	1950.20366827434	1212.37938394503	1958.74276504768
”TimeFinished”	1139.12065538303	1950.69424398922	1205.68373623819	1958.94527223111
”NumMales”	25.217082268371	0.880567760556212	25.3420604309958	0.86174504778515
”NumFemales”	24.9348604026486	0.099834829537178	25.015559445744	0.18911238364772
”NumArtists”	25.4977976442561	1.09224256581752	25.4762037334498	1.12468902828081

Table 4: A selection of variances and means per class

6 Conclusion

Despite not having a classifier that gives a reliable accuracy, we have revealed interesting facts about the data. We see as more features were added into consideration, the Logistic Regression model changed the weights from all positive to also showcase negative weights. This leads us to believe that the addition of more features are pivotal to a better model. We are currently missing the ”ArtistBio” and ”Nationality” from the original dataset. In addition, creating dummy variables for price of the artwork as well as doing a better job at string analysis so ”CreditLines” attribute understands similiarities between different strings.

In terms of re-evaluating our Naive Bayes classifiers, we have three phrase-based attributes in the data: ”CreditLines”, ”ArtistBio” and ”Medium”. There is a real case, in trying to doing a specialized string analysis on these different attributes.

In addition, we saw that ”Departments”, ”Classification”, and ”Medium” had a type of hierachial relationship with each other. This showcases that the data is not as independent as we originally assumed, and so in the future, we plan to perform a Random Forest classifier to account for this.

7 Remarks

My final remarks about this Machine Learning class is that some of the topics felt fast. The lectues go over a lot of material in one sitting, and I just don’t have enough time to digest what is going on. I know since the class attendance was low, it discourage asking us, the students, questions, but I think it would have helped more for my understanding. I also think forcing us to write in mathematical notation more often for the question parts of the homework would have helped me in remembering and understanding the math formulas. (Everytime I saw a partial derivative symbol after not seeing it for awhile, I could not remember what that was.) In terms of class material, there was a lot of interesting sections that were covered. I did wish that we could have gone over Singular Value Decomposition, and the Netflix prize. (I have been researching that by myself, and it would be helpful if we could have gone over the idea of matrix completion.)

8 References

8.1 Tables

This subsection showcases some of the longer tables referenced throughout the paper or those who aren't directly related to understanding the paper.

Original Attributes
Title
Artist
ConstituentID
ArtistBio
Nationality
BeginDate
EndDate
Gender
Date
Medium
Dimensions
CreditLine
AccessionNumber
Classification
Department
DateAcquired
Cataloged
ObjectID
URL
ThumbnailURL
Circumference (cm)
Depth (cm)
Diameter (cm)
Height (cm)
Length (cm)
Weight (kg)
Width (cm)
Seat Height (cm)
Duration (sec.)

Table 5: List of attributes before cleaning, or removal.

”Attributes After Cleaning”	
”CreditLine”	”albumen”
”Classification”	”gouache”
”Department”	”gelatin”
”DateAcquired”	”poster”
”Circumference (cm)”	”illustrated book”
”Depth (cm)”	”colored pencil”
”Diameter (cm)”	”video”
”Height (cm)”	”tracing paper”
”Length (cm)”	”paper”
”Weight (kg)”	”silver print”
”Width (cm)”	”sound”
”Seat Height (cm)”	”oil”
”Duration (sec.)”	”cardboard”
”aquatint”	”inkjet print”
”watercolor”	”woodcut”
”pigmented”	”matte”
”canvas”	”charcoal”
”linoleum cut”	”paint”
”silkscreen”	”photolithograph”
”photogravure”	”color”
”offset”	”dye transfer print”
”letterpress”	”glass negative”
”silent”	”ballpoint pen”
”platinum print”	”color print”
”portfolio”	”white”
”screenprint”	”engraving”
”chromogenic”	”etching”
”black”	”collotype”
”lithograph”	”TimeStarted”
”drypoint”	”TimeFinished”
”wood”	”NumMales”
”board”	”NumFemales”
”pencil”	”NumArtists”
”bronze”	”FirstArtistListed”
”ink”	

Table 6: We see the new attributes that were extrapolated and the entirity of the full attribute set.

8.2 Exploratory Data Analysis

This section showcases graphs that were made from some of the initial exploratory analysis on differences between the two classes on three attibutes: acquisition dates, departments, and credit lines.

Figure 1:

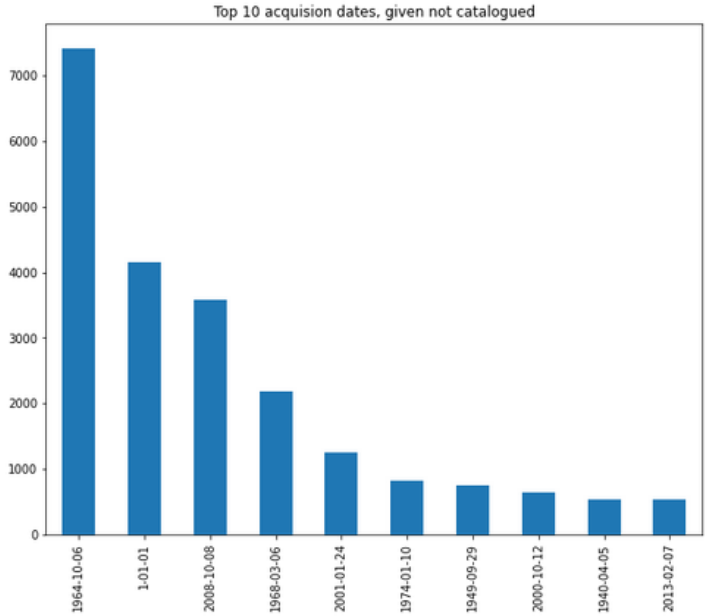
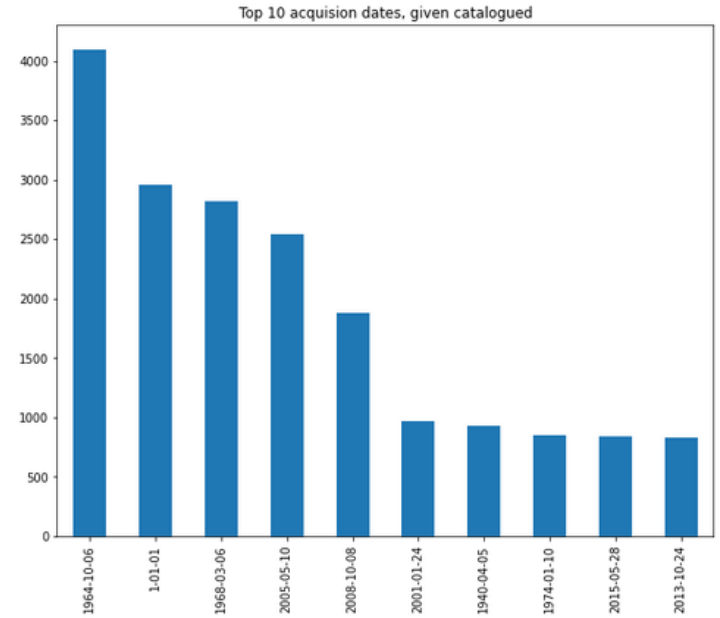


Figure 2:



As seen from the two graphs above, most of the top dates are shared between the two classes, just the order is different. Its interesting to see that thousands of works from over 50 years ago still have not been catalogued, yet.

Figure 3:

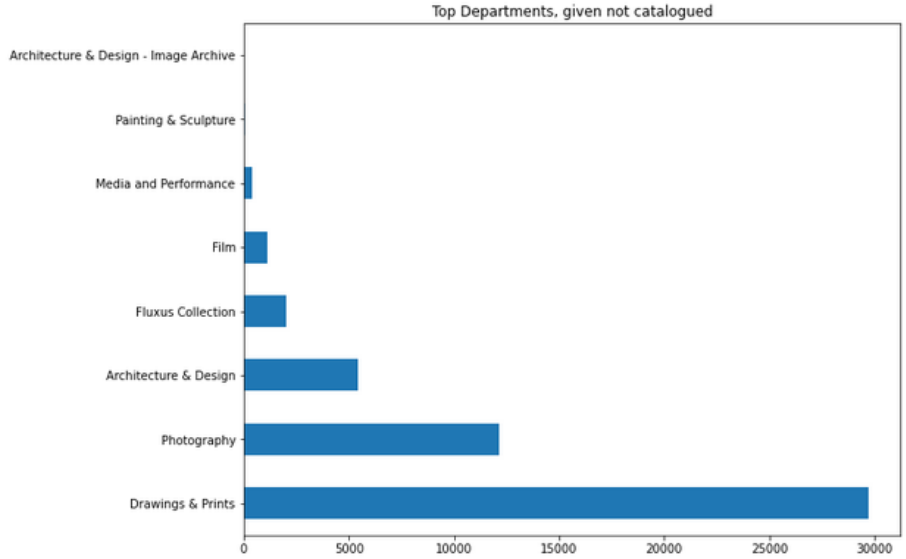
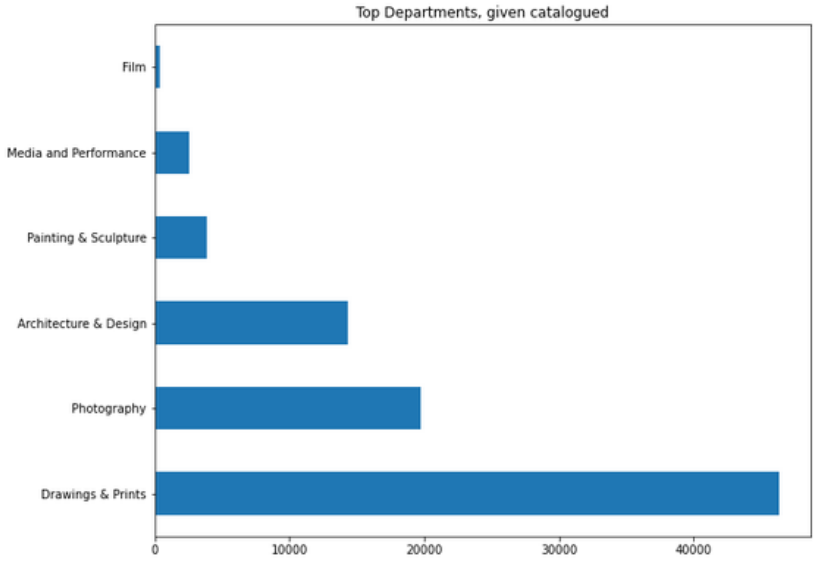


Figure 4:



We see the same idea as with Figure 1 and Figure 2, that is, most elements are shared between the two classes. However, we see that two departments have not been catalogued at all. Through some research, we found out that the Fluxus Collection was mostly active between the early 1960s to the late 1970s. The reason behind this class difference might be due to a Museum policy such as don't catalogue any artwork in a department till its guaranteed that you can catalogue a certain percentage, or some other policy.