Artsy Propensity Analysis

Predicting Customer Quality

Summary

One of Artsy's largest revenue opportunities is a SaaS offering to art galleries around the world. For much of Artsy's gallery business's lifetime we have been able to rely on inbound applications and cherry picking from the global art community. In 2018 the company looks to greatly expand its investment in the gallery product and sales team. As such, the business needs to learn more about it's addressable market as well as be able to better funnel qualified leads directly to the sales team. This analysis tries to peel back some of the layers around the addressable market and formulate a classifier that can qualify prospects to a satisfactory degree.

Current State of Data

Overview

Artsy has been collecting data on art galleries for a few years now. Our sources range from web scrapes to bought lists from aggregators such as Factual. As such we have varying degrees of data coverage, with the most being user habits from our current subscribers, to the least being just a name and a website. There are additional problems with the data since it has come from so many different sources over so many years, some of them being accuracy and staleness.

Our goal over the next quarter is to acquire as much information about these prospects as possible in order to better explore and classify.

We estimate that there are ~90,000 art galleries in the world and somewhere between ~20,000 - ~26,000 are eligible for current product offerings. We arrived at this conclusion by doing a sampling test of ~3,500 galleries over a variety of different internal databases. The context of each isn't important, but the graph below gives an understanding of the nuances behind our current state:

Category	Total quantity	Estimated % Qualified	Estimated Qualified quantity
SF Active Artsy Subscribers	2,155	100%	2,155
SF Qualified Gallery Accounts	5,428	95%	5,157
SF Qualified Fair Partners	2,631	95%	2,499
SF Fair Partners	440	90%	396
SF Gallery Accounts	3,771	80%	3,017
SF Qualified Gallery Leads	1059	70%	741
SF Gallery Leads	17,520	40%	7,008
Total SF	**30,849**	**68%**	**20,995**
Bearden (art organization database) "Arts Organizations"	50,000	16%	8,000
Total Overall	**80,849**	**36%**	**28,995**

Stack

Our database stack in this context is comprised of Salesforce, a native app Bearden (mostly Ruby), and a host of core native applications. We plan to use this loose funnel when thinking about what data lives where:

Bearden > Salesforce > Core Applications

Bearden holds all of our data on organizations regardless of coverage or confidence in how relevant the prospect is to our business. In Salesforce we put organizations we think are galleries, and in our core applications we only keep galleries that have used Artsy's product at some point (as well as all our B2U users). We use AWS Redshift as our main analytics database and do most of the ETL ourselves, minus a few tools that use a plug and play Segment.io product.

The analysis was done almost entirely in an IPython Notebook in the Artsy Analytics repo Minotaur. This allowed for easy querying access to Artsy's database as well as a way to share and get feedback from other members of the Artsy team, primarily Anil and Will.

Analysis

Even though we know there is a lot of data collection ahead of us, we tried an initial exploratory analysis which contained some unsupervised learning. We included organizations in our database that are galleries, have an artist roster, a gallery tier, and a location. The reason why we only looked at galleries that have already been tiered is that ultimately we wanted to run some supervised learning models on this dataset. A supervised learning model is where you give an algorithm the raw data and the result you are trying to predict. It then trains itself on the data, hopefully in the process becoming a good predictor of quality. Our plan is to revisit the EDA and unsupervised learning part of this project after we have attained more data, broadening the dataset to include organizations that are not yet classified.

Data Wrangling

In order to start the analysis we had to pull in data from many different sources. Organization information came from a combination of Salesforce, Bearden and Gravity. An existing data pipeline infrastructure with Redshift built by the analytics team made easy work of this. Additional features such as artist roster and Google keyword search volume were pulled in from spreadsheets. The artists rosters were attained by the Marketing Team in the first half of 2017 for another project. While there are more features to be had, especially for galleries who have interacted with the platform, we tried to only use features we knew we could attain for the entire universe of galleries. All of the data wrangling can be found HERE.

After we decided on what features and pulled everything in we had to combine, sift through, and start chugging away. Using the Python Pandas library we easily created a DataFrame with a sample below:

gallery_tier	Clicks	Impressions	inquiry_requests_count	bid_count	on_artsy	search_volume	artist_slug	min_distance_to_art_city	has_roster	disqualified_reason
50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	133.392199	NaN	NaN
50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3499.671448	NaN	NaN
50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NaN	True	NaN
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3674.192945	NaN	Too Low Quality
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NaN	NaN	Too Low Quality

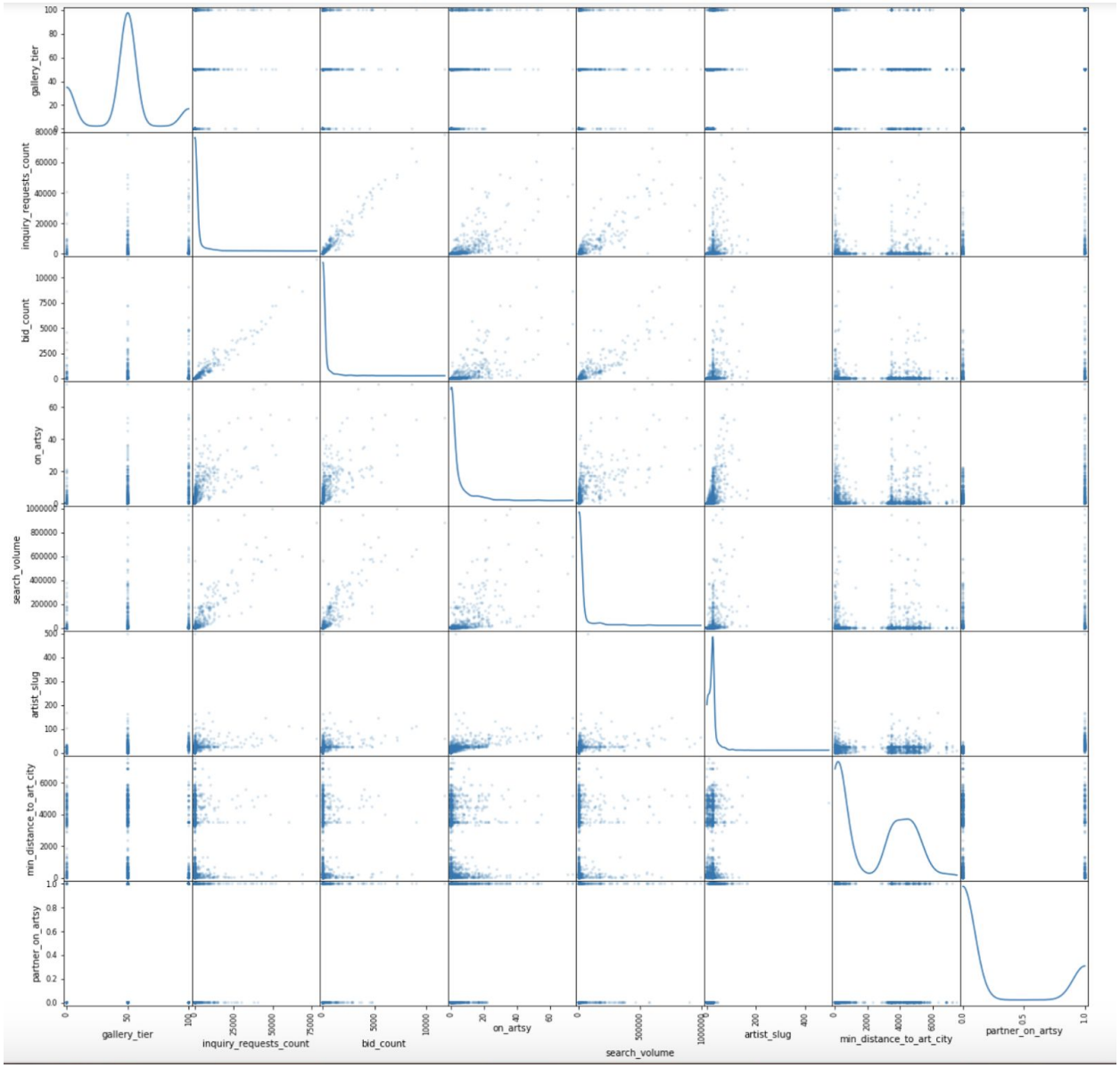
Most of the numerical columns or sums are based on the gallery's artist roster. For instance, the "Search Volume" column is a sum of the google search volume for each artist's name that belonged to that gallery's artist roster. "Minimum Distance to Art Center" is a generated feature, in which we took the distance between the galleries location and one of 10 "Art Centers" across the globe such as London, New York, or Buenos Aires. Finally we consolidated gallery tier from 5 classes to 3: very qualified, qualified, not qualified. An additional feature not pictured above is a text analysis of the companies website that will be used in the supervised learning part of the analysis.

Exploratory Data Analysis

The goal of exploring the data is to see if we can find any correlations, as well as gain a better understanding of each class. We are able to eliminate some highly correlated and lowly covered features. There is also a large class imbalance, ie we have a much greater proportion of qualified than not qualified galleries. Below we can see what our artist roster coverage is, which is a driving factor of much of the features included in this report. There are ~3300 qualified and ~900 not qualified galleries that have an artist roster. You'll also notice there are few very qualified galleries, which is consistent with our qualitative understanding of the gallery universe.

		domain
qualified	**has_roster**	
not_qualified	**False**	233
	True	895
qualified	**False**	345
	True	3012
very_qualified	**False**	22
	True	315

Next we look at a correlation matrix:



The squares that show clusters of points moving up and to the right represent two features that are highly correlated with each other. You will notice that gallery tier is split into 3 lines as there are 3 categories: qualified, very qualified and not qualified. Below the correlations are represented numerically.

	gallery_tier	inquiry_requests_count	bid_count	on_artsy	search_volume	artist_slug	min_distance_to_art_city	partner_on_artsy
gallery_tier	1	0.15	0.14	0.31	0.13	0.23	-0.08	0.37
inquiry_requests_count	0.15	1	0.98	0.76	0.88	0.32	-0.0065	0.23
bid_count	0.14	0.98	1	0.72	0.86	0.31	-0.0043	0.21
on_artsy	0.31	0.76	0.72	1	0.68	0.48	-0.027	0.42
search_volume	0.13	0.88	0.86	0.68	1	0.31	0.026	0.2
artist_slug	0.23	0.32	0.31	0.48	0.31	1	0.059	0.44
min_distance_to_art_city	-0.08	-0.0065	-0.0043	-0.027	0.026	0.059	1	-0.038
partner_on_artsy	0.37	0.23	0.21	0.42	0.2	0.44	-0.038	1

We can see that much of the artist roster data that comes from Artsy is highly correlated. After digging into this more we also see that coverage is very low for a lot of these features, with only ~1500 galleries having coverage across all of them. Also the range across some of these features is huge, which introduces a lot of noise to the analysis. This leads us to believe that we need to get much better coverage of these data features as well start our analysis with a more limited feature set.

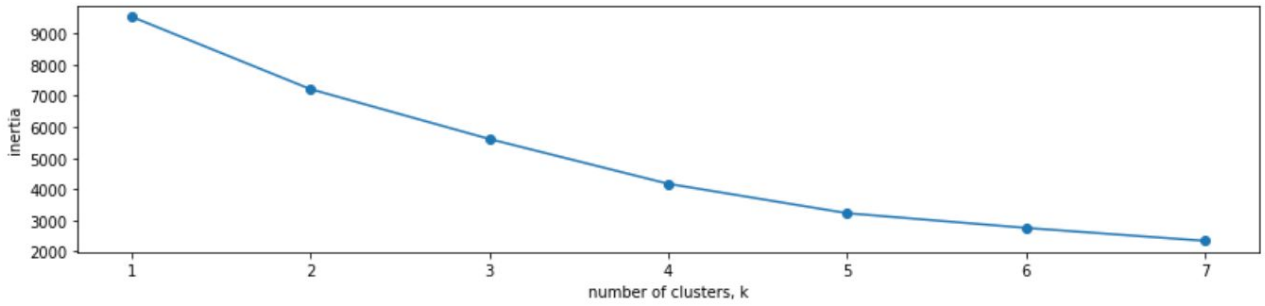
Unsupervised Learning

The features our EDA led us to include in the Unsupervised and Supervised Learning models are: 'on_artsy', 'search_volume', 'artist_slug', 'min_distance_to_art_city', 'has_roster', and 'partner_on_artsy'. We eliminated most of the Artsy artist information at this point in the analysis because the coverage was too poor. For the unsupervised part of the analysis we used a few forms of clustering to gain some insights into our TAM as it relates to qualification. Unfortunately with the data we have the results are inconclusive. Ultimately this first analysis was trying to predict qualification more than understand our TAM. There is definitely more to be done here in terms of gaining gallery “profiles”.

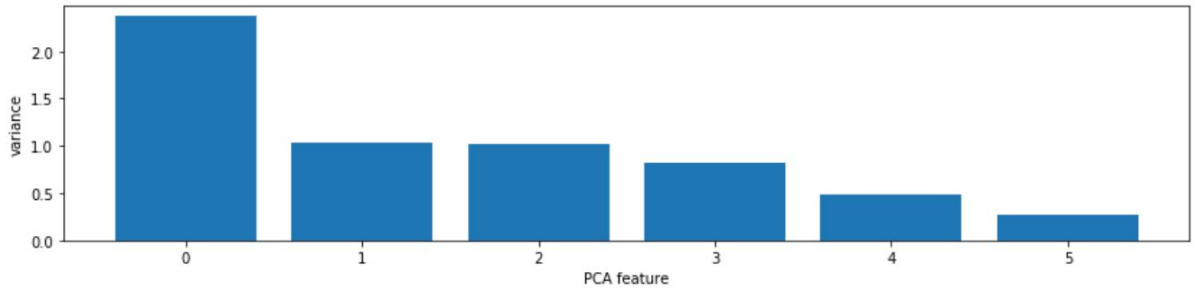
Below are the crosstab analysis, some feature reduction using PCA, and hierarchical clustering using only galleries that have full coverage of the features mentioned above.

qualified labels	not_qualified	qualified	very_qualified
0	72	104	5
1	271	719	48
2	13	207	89
3	7	31	21

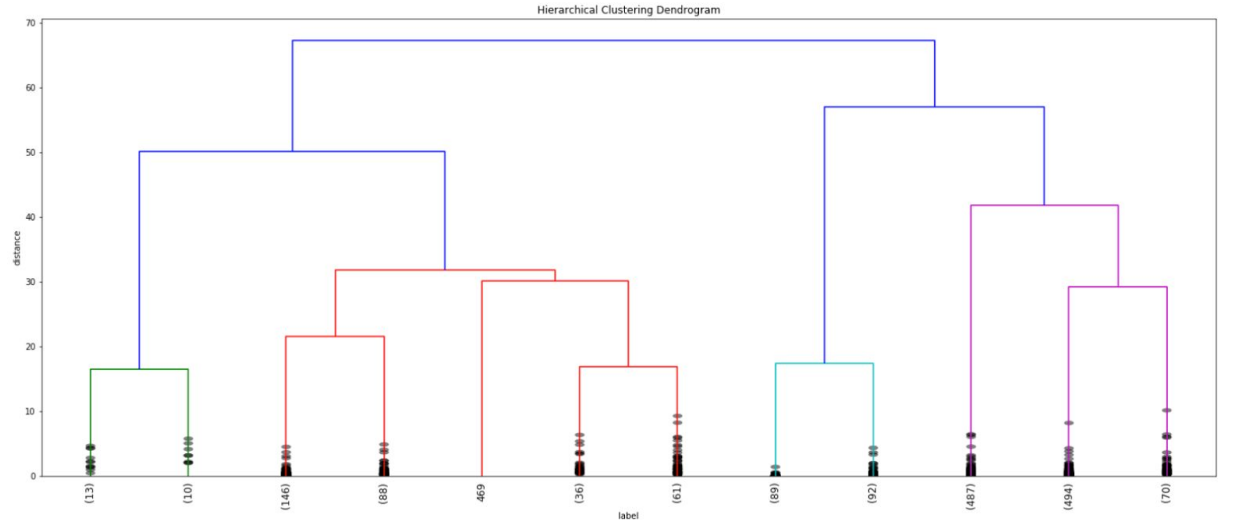
As you can see above the clusters that the KMeans model created did not group along the qualification classes.



This graph plots the distance between the sum of squares of each clusters centerpoint. Ideally we would want this to be more of an elbow graph which would mean the data forms distinct clusters.



There seems to be one group of features that has high importance, 3 have some importance and the rest are not very important.



Heirarchical clustering shows that there may be four distinct clusters in the data. This is somewhat consistent with our KMeans model.

Supervised Learning

For the supervised learning part of the analysis we focused on three models: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. We have also limited our dataset down to 3200 rows of information that contain an artist roster and location. For the analysts reading this we used GridSearchCV to determine the best hyperparameters each time a model was run.

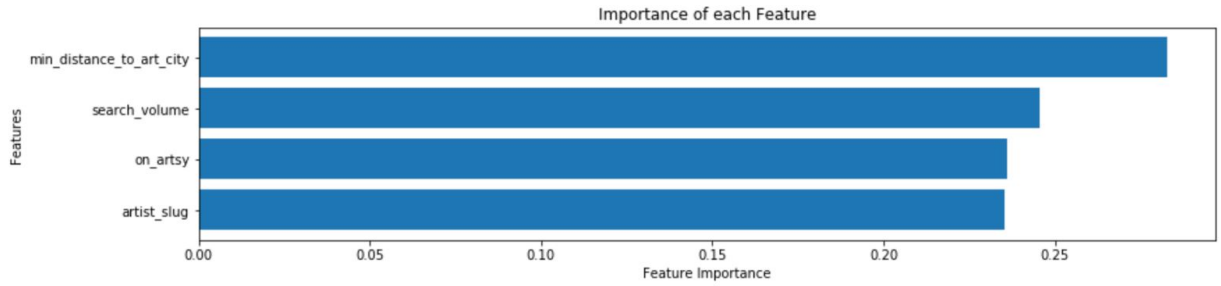
Within this dataset there are only 262 “not qualified” galleries which leads to a major class imbalance problem. As already mentioned we are going to get more data as one solution, but I also tried undersampling. After many iterations across the 3 models the best score occurred when limiting the dataset to only “very qualified” and “not qualified” using a Random Forest Classifier.

Classification Report	Precision	Recall	f1-score	support
not_qualified	0.59	0.71	0.65	76
very_qualified	0.9	0.84	0.87	227

Above is the classification report for the Random Forest model trained only on “not qualified” and “very qualified”. I won’t dive into too much detail but note that the f1-score is the harmonic mean of precision and recall, and support is the number of true responses that lie in that class. Even though some of the numbers here are high, ultimately it is still not a great model, particularly when it comes to predicting not qualified galleries.

Confusion Matrix	not_qualified	very_qualified
not_qualified	54 True Positives	22 False Positives
very_qualified	37 False Negatives	190 True Negatives

The confusion matrix does highlight some more positive results with 54 true positives for the predicted not_qualified class out of 66 samples.



There is a healthy feature importance across all features. While the model may not be good enough to use to surface prospects to sales, it at least tells us we are on the right track with determining what features to include in the analysis.

In conclusion, a supervised learning model is really only as good as its weakest link which in this case is 65%. So one could say that this model can predict accuracy a minimum of 65% of the time. Yet, since we only used not_qualified and very_qualified classes this is not very representative of our real world problem.

Conclusion

Through unsupervised learning and EDA we were able to determine that our 3 class approach is on the right track, and we removed many extraneous features. With supervised learning we were able to verify that the features we did choose were predictive, and create a model that was significantly better than chance. However, while these are interesting results that lead us to believe we are on the right track, the ultimate conclusion is that we don't have enough examples or enough features to come away with anything conclusive. As such we must go out and get more data.

Important Links

- Unsupervised Learning and EDA
- Supervised Learning
- Presentation