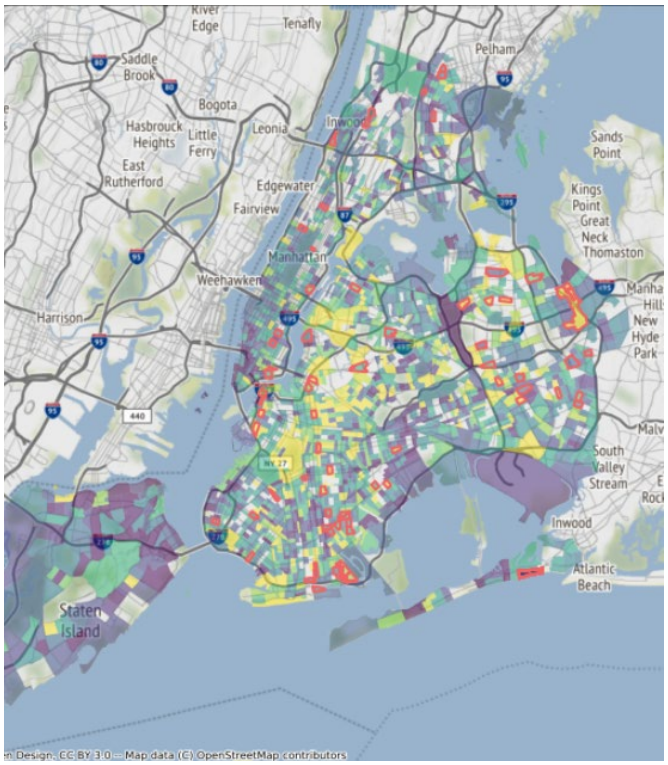


Team 30

Citadel Data Open Submission

ABSTRACT

In this report, we first introduce a new methodology to quantify the gentrification process on each tracts, rather than simply using the binary classification on tracts. We will present our model on predicting gentrification process of all tracts in NYC city based on noises complaints, racial decomposition, ratio of bachelor's degree holder, median home value and median income. We build a linear regression to explain the relationship between gentrification process and various feature explicitly. And we build a hybrid LSTM-DNN network to predict on gentrification process in the next year. We have significant result on gentrification process prediction for next year with 0.636 correlation and 0.581 correlation for home value prediction in next year

Team Members:

Ramunas Genys | Deivis Banys | Richard Xia | Chris Liu

Executive Summary:

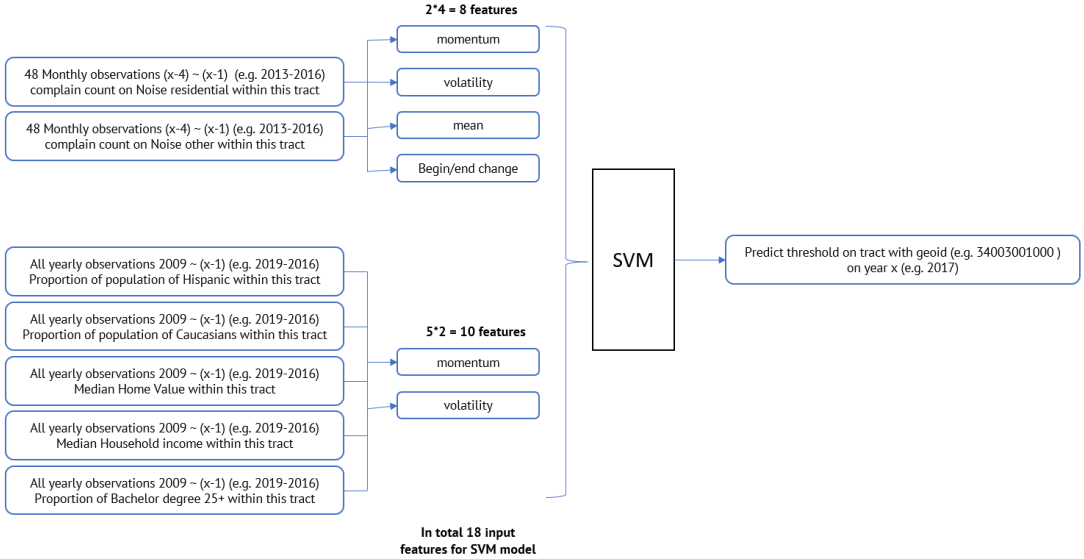
In this project, we are going to forecast the gentrification process and the residential home value movement of one interested tract for a specific year based on:

- 5 sets of all available yearly historical observation including:
    - Proportion of population aged 25+ and had Bachelor's degrees
    - Proportion of population of non-Hispanic Caucasians
    - Proportion of population of Hispanic
    - Median household income
    - Median home value
- 2 set of 48 months historical observation:
    - Monthly number of complaints on Noise residential
    - Monthly number of complaints on Noise others

For example, if I am predicting on gentrification process of geoid 34003001000 for year 2017, I am using yearly observation of these time series from 2009 to 2016 and using two monthly time series on 311 calls from 2013-01-01 to 2016-12-31.

And we have two models using:

1. A SVM regression on 4 features "momentum, volatility, mean and begin/end difference "of 2 sets of monthly timeseries and 2 features of "momentum, volatility" on 5 sets of yearly timeseries.



2. A hybrid LSTM-DNN network by building LSTM layer on 2 monthly time series and a dense layer on 2 features of "momentum, volatility" on 5 sets of yearly timeseries.

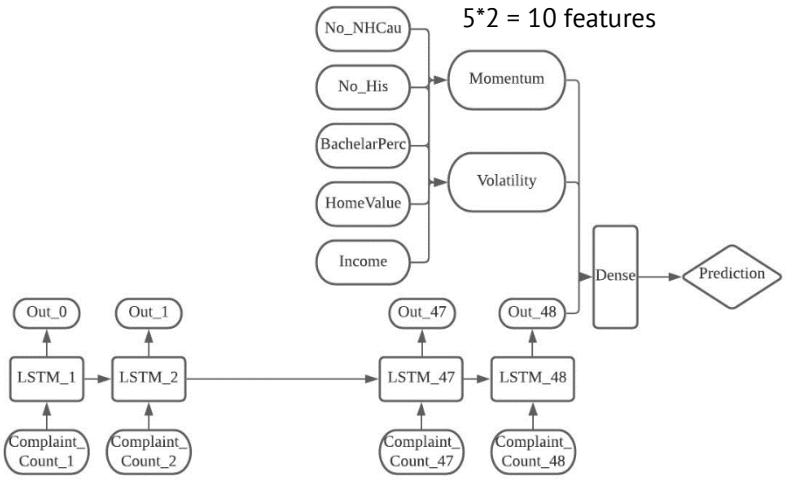


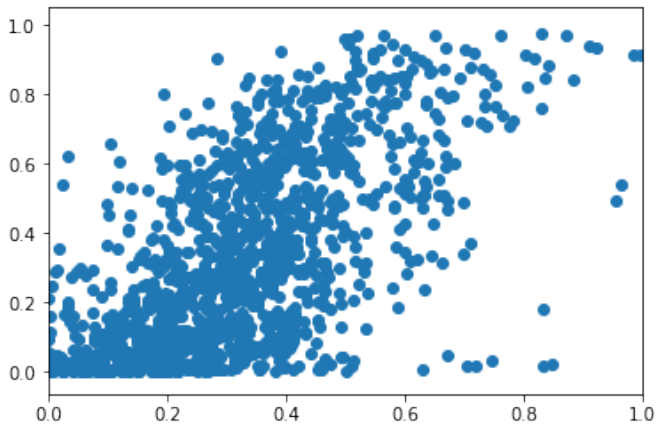
Figure 1: Hybrid model tensor structure

We will introduce new methodologies on predicting gentrification process, by using the minimum percentile value $\boldsymbol{k}$ that could make this tract pass the second gentrification test (see next section for definition of first test, second test and threshold). And we will work on two datasets:

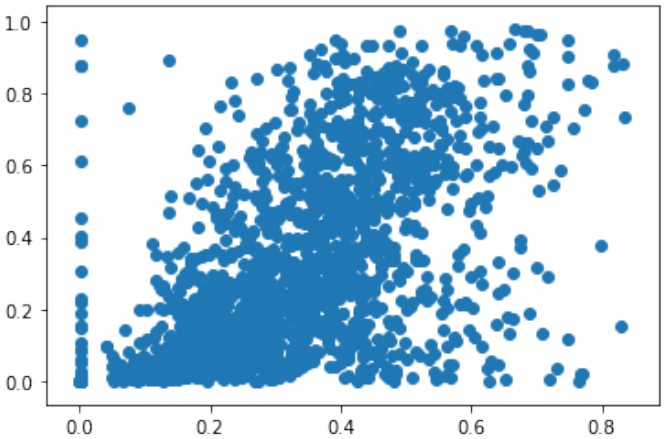
- **Datasets A:** One model is only on tracts that passed the first test of the gentrification testing
- **Dataset B:** Another model is all on tracts even it didn't pass the first test.

The model on first dataset is useful for predicting whether a tract is going to gentrified in the next year. And the model on second dataset is also useful in predicting the residential home value movement because the second gentrification test is a simple reflection on the home price and education level. If the predicted threshold is large, we could foresee that this tract will have highest home value in NYC area in the following year.

Because there are only 317 tracts that occurred in all data and passed the first test, we use SVM model for model A and use LSTM-DNN model for problem model B



Prediction result on testing dataset only on tracts that passed first test using Support Vector Machine regressor



Prediction result on testing dataset on all tracts using LSTM-DNN model

Section 1: Gentrification Test and Quantify Gentrification Process:

By definition from Governing website, a tract is eligible to gentrify if it passed two independent tests:

The first test:

- The tract had a population of at least 500 residents at the beginning and end of a decade and was located within a central city.
- The tract's median household income was in the bottom 40th percentile when compared to all tracts within its metro area at the beginning of the decade.
- The tract's median home value was in the bottom 40th percentile when compared to all tracts within its metro area at the beginning of the decade.

The second test: $F_{66}(Geoid)$

- An increase in a tract's educational attainment, as measured by the percentage of residents age 25 and over holding bachelor's degrees, was in the **top third percentile** of all tracts within a metro area.
- A tract's median home value increased when adjusted for inflation.
- The percentage increase in a tract's inflation-adjusted median home value was in **the top third percentile** of all tracts within a metro area.

Hence, if a tract passes the first test, then it is eligible to be defined as gentrified if the percentile for the home value and bachelor percentage is greater than 66. And we can say the second test is a test with threshold 66.67. And mathematically, we state $F_{66.67}(Geoid) = 1$ if one tract can pass the second test with threshold 66.67. (top 66.67 percentile in bachelor percentage and median home value)

The we can quantify the gentrification process by defining **a new property $\boldsymbol{k}$** for one tract $Geoid$:

$$\boldsymbol{min_{0\leq k\leq 1}\, F_k(Geoid) = 1}$$

The higher the k is, the sooner this tract is gentrified. And if one tract has $k \geq \frac{2}{3}$, we can say this track is gentrified. And we can calculate a value k for each tract at each year using this definition. We can also assign a $\boldsymbol{k}$ value to tracts even if it did

not pass the first test, then this ***k* simply represent the end year percentile position for home value and education.** That is the reason why we say the **model B** is useful for predicting the home price movement in the next year.

Section 2: Data Processing and Exploration

2.1: Census Data

Download:

We used the census API to download the **Census data** dataset from 2009 to 2018, in order to download additional features to calculate percentage of residents that are their age of 25+ and are holding bachelor's degrees. These are the additional variables I downloaded from the API

B15001_017E	Estimate!!Total!!Male!!25 to 34 years!!Bachelor's degree
B15001_025E	Estimate!!Total!!Male!!35 to 44 years!!Bachelor's degree
B15001_033E	Estimate!!Total!!Male!!45 to 64 years!!Bachelor's degree
B15001_041E	Estimate!!Total!!Male!!65 years and over!!Bachelor's degree
B15001_058E	Estimate!!Total!!Female!!25 to 34 years!!Bachelor's degree
B15001_066E	Estimate!!Total!!Female!!35 to 44 years!!Bachelor's degree
B15001_074E	Estimate!!Total!!Female!!45 to 64 years!!Bachelor's degree
B15001_082E	Estimate!!Total!!Female!!65 years and over!!Bachelor's degree

Processing:

We sum up all those 8 features for each year and each geoid, then divided by the total population of the tract in that year, which successfully gives us the percentage of population 25+ and having bachelor's degree. Then we merged the 10 yearly census files into one file. And we also removed the tracts that have any missing data across the past ten years.

In summary, we have 3867 different tracts that has no missing values in the past ten years. ('mycensus_all.csv')

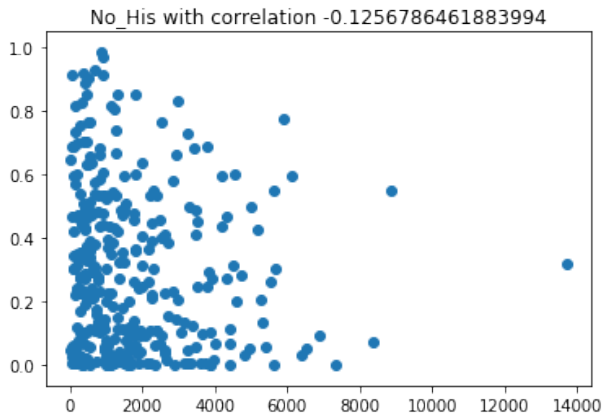
And if we apply the gentrification test, there are **612 tracts** pass the first test. There are **90 tracts** passed the second test, which are the gentrified tracts.

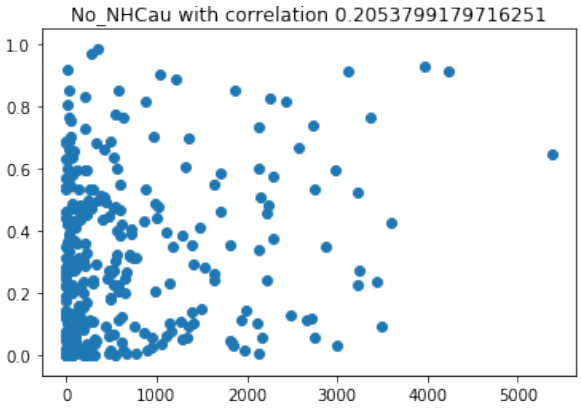
geoid	year	Populatic	No_NHIn	No_NHCa	No_NHBl	No_NHAs	No_NHHa	No_NHOt	No_NHMu	No_His	MedianIr	MedianH	NAME	county	state	tract	BachelarPerc
34003001000	2009	6600	0	5600	87	651	0	37	66	159	127757	731500	Census Tr	3	34	1000	0.233485
34003002100	2009	2008	7	1327	43	485	0	0	0	146	167917	1000001	Census Tr	3	34	2100	0.331673
34003002200	2009	5196	0	3531	18	1202	0	22	47	376	89615	535700	Census Tr	3	34	2200	0.144534
34003002300	2009	6158	0	4262	105	1713	0	0	35	43	93640	662300	Census Tr	3	34	2300	0.225398
34003003100	2009	4900	0	1860	242	1719	0	0	67	1012	92386	382900	Census Tr	3	34	3100	0.204898
34003003200	2009	3897	0	1770	299	913	0	76	40	799	63365	356600	Census Tr	3	34	3200	0.221966
34003003300	2009	6715	0	3140	370	1884	0	0	43	1278	92903	379000	Census Tr	3	34	3300	0.159792
34003003401	2009	2715	0	1418	130	536	0	0	124	507	103083	399700	Census Tr	3	34	3401	0.182689
34003003402	2009	3379	0	2254	130	705	0	0	12	278	89719	408000	Census Tr	3	34	3402	0.17372
34003003500	2009	3969	0	1127	666	967	0	0	112	1097	54092	416300	Census Tr	3	34	3500	0.175107
34003005000	2009	5962	0	4616	0	575	0	0	0	771	63903	442500	Census Tr	3	34	5000	0.122274
34003006100	2009	6188	0	3767	99	716	0	10	35	1561	61425	486100	Census Tr	3	34	6100	0.184228

And I generated the gentrification process property for each tract for each year from 2010 to 2019. (the earliest gentrification property k is available is at year 2010, as I need to calculate the increase in home value from 2009 to 2010). The gentrification property of each tract at each year is generated using the formula mentioned above, even if the tract did not pass the first gentrification test (calculate k for all 3867 tracts).

$$k: \boldsymbol{min_{0 \le k \le 1}} \boldsymbol{F_k(Geoid)} = \mathbf{1}$$

Then I can look at the relation between different features and the gentrification process of each tract, by looking at the scatter plot of the feature value and the gentrification property of the tracts that has passed the first test (612 tracts).

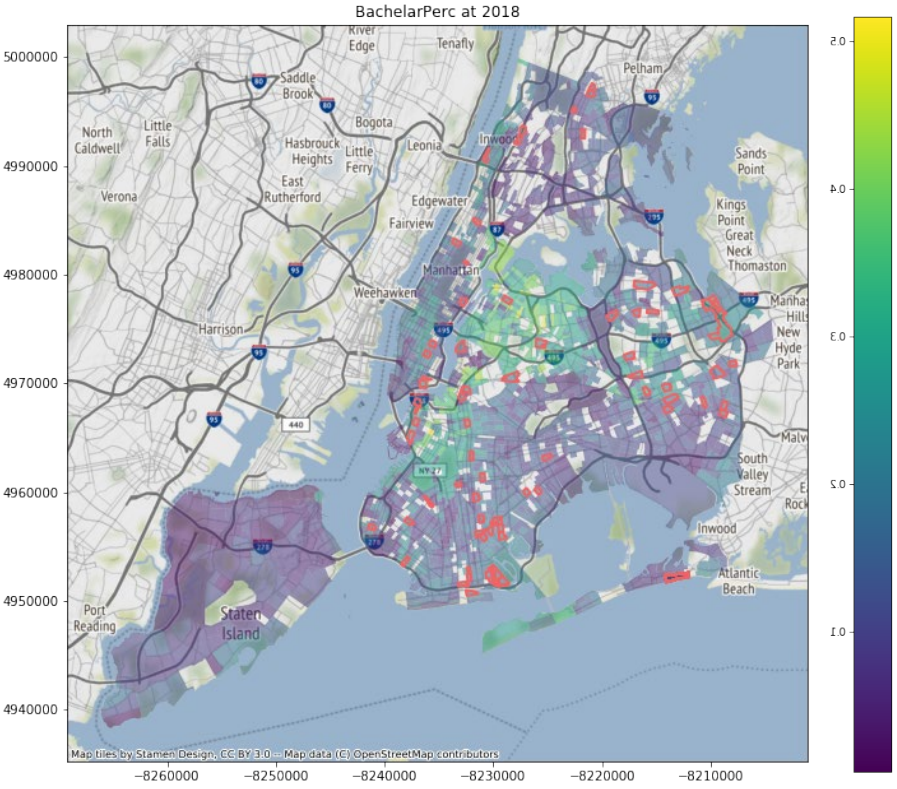


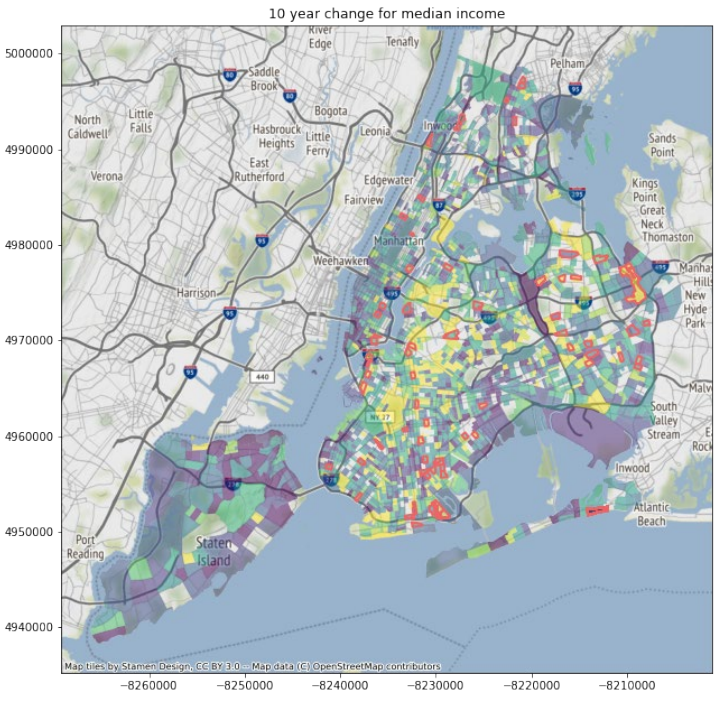


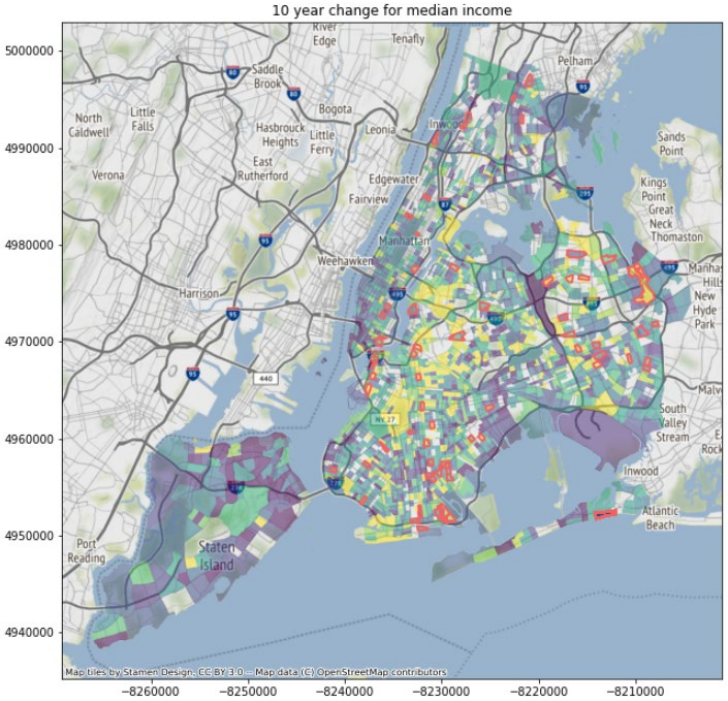
And from my observation, the feature "**Proportion of population of non-Hispanic Caucasians**" and "**Proportion of population of Hispanic**" has the biggest absolute correlation with the gentrification process. That's why I decided to include it into my model as my feature.

2.2: NYC tracts shapefile

I downloaded the shapefile for the NYC tracts from https://data.cityofnewyork.us/City-Government/2010-Census-Tracts/fxpq-c8ku . Because it only contains the 2724 tracts shapefiles from NYC city, I will not able to visualize all 3867 tracts that is contained in the "mycensus_all.csv".







In the graphs above we compared income, house price and college education level changes from 2009 to 2018 (represented by the colour contour). We also highlighted gentrified areas (circled by the red line.) in order to see how their patterns differ from non-gentrified ones. We did not plot tracts with missing values in any of the years.

From image above we can observe that the **bachelor's degree percentage can be a good indicator** in forecasting as all the gentrified features are next to the region a high percentage of the population are having bachelor's degree. It is evident that Brooklyn and Queens has seen a large increase in both median home value and median income. This matches with our observations that most gentrified tracts are in these areas. We can also observe that college education levels increased in those areas too. However, it is worth mentioning that income, house prices and \% of college graduates increased in almost every part in New York, although not as significantly. Hence the **home price and household income** can also be a good indicator in determining the gentrification process in the next year.

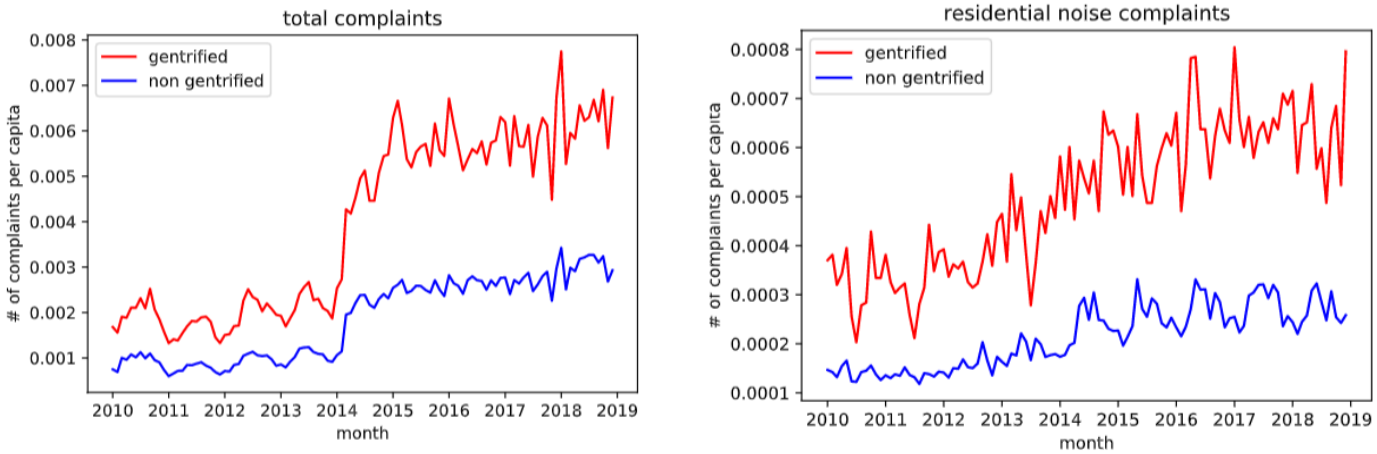
2.3: 311 Calls Complains

We simply used the 311 call dataset given by the competition. We removed the records that do not have the location value (**Latitude | Longitude**). This left us 3394922 records of 311 calls in the past 10 years. And to convert Latitude and longitude in the 311 call to the specific geoid, we downloaded the "census_block_loc.csv" from Kaggle data competition.

	Latitude	Longitude	BlockCode	County	State
0	40.48	-74.280000	340230076002012	Middlesex	NJ
1	40.48	-74.276834	340230076005000	Middlesex	NJ
2	40.48	-74.273668	340230076003018	Middlesex	NJ
3	40.48	-74.270503	340230076003004	Middlesex	NJ

Figure 2: census_black_loc.csv

So for each 311 call, I looked at the Lat and Long of that call location, and try to file the most close BlockCode in the "census_block_loc" file, and then remove the last 4 digits of the block-code to generate the tracts "Geoid" for that piece of phone call. Then I used "Numba" packages in python to accelerate this calculation process and apply it to all 3million+ records of 311 call in past 10 years. I also calculated what are the neighbour tracts of these 311 calls by looking at the 2nd 3rd and 4th closest tracts geoid to this 311-phone call. There are **2164 unique tracts geoid** occurred in the past 10 years in this dataset only.



However, if we merge 311 call and "mycensus" files by geoid, there are then only **1744 common tracts** geoid showed up in both files. And **317 common tracts** passed the first test, **73 tracts** passed the second test and eligible for gentrification.

We compare the total number of complaints between gentrified and non-gentrified tracts. To make the comparison fair, we normalise the number of complaints in gentrified and non-gentrified areas by their respective total populations. We find that the per-capita number of complaints is significantly higher in gentrified areas, especially after gentrification. We notice a sharp increase in complaints in the gentrified areas around the time most of them were considered gentrified-year 2014. This can also be attributed to other factors - for example, the data in newer years has a lot fewer missing values, and so the number of complaints could have been under-reported. The biggest difference between gentrified and non-gentrified areas was noticed in residential noise complaints.

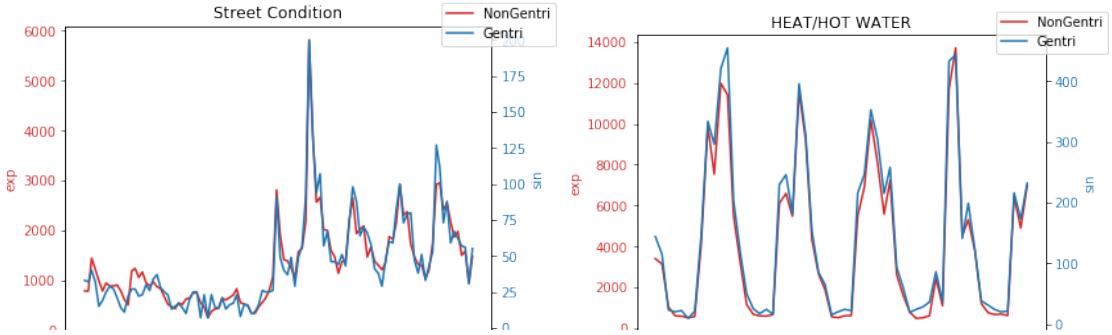


Figure 3: features has same distribution over gentrified and nongentrified tracts

Then we further explore the difference in different types of noise complaints. We specifically extract the complaint type “noise - residential” and then extract another complaint type that contained keyword ‘noise’. Then we count the monthly number of complaints on ‘Noise-residential’ and ‘Noise-others” in the gentrified and non-gentrified tracts. If we scale the number of complains to match each other, we could see Noise Residential significantly differ from other features, while most of other complaint type has exactly the same seasonality and property in gentrified and non-gentrified region.

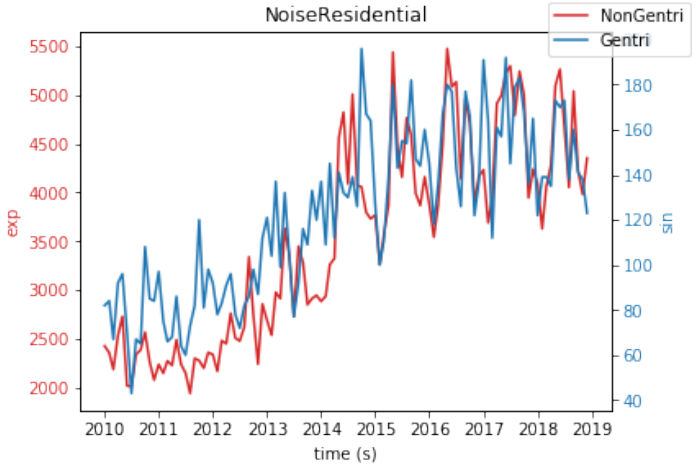


Figure 4: Noise residential has differed property in different types of tracts

Hence, I will use the monthly timeseries of number of complaints on “Noise Residential” and “Noise others” in each tract as a feature to make predictions.

Section 3: Support Vector Machine Model

From basic features, which are calculated monthly or yearly, we generated synthetic static features, which are the momentum, mean and standard deviation of the base feature. Two Ordinary Least Squares tests on 317 gentrified tracts (left) and all 1744 tracts (right) are performed to validate that feature importance of these synthetic features.

OLS Regression Results

Dep. Variable:	threshold	**R-squared:**	0.599
Model:	OLS	**Adj. R-squared:**	0.595
Method:	Least Squares	**F-statistic:**	130.1
Date:	Sun, 25 Oct 2020	**Prob (F-statistic):**	5.97e-295
Time:	14:17:11	**Log-Likelihood:**	586.29
No. Observations:	1585	**AIC:**	-1135.
Df Residuals:	1566	**BIC:**	-1033.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2322	0.016	14.769	0.000	0.201	0.263
noiseother momentum	105.3319	581.059	0.181	0.856	-1034.403	1245.067
noiseother std	-99.3193	43.673	-2.274	0.023	-184.982	-13.656
noiseother mean	68.1538	38.022	1.792	0.073	-6.425	142.733
noiseother begin/end change	1.737e-05	0.000	0.122	0.903	-0.000	0.000
noiseresid momentum	687.9690	313.754	2.193	0.028	72.546	1303.392
noiseresid std	98.4616	30.224	3.258	0.001	39.177	157.746
noiseresid mean	-44.5070	24.012	-1.854	0.064	-91.605	2.591
noiseresid begin/end change	-2.986e-05	0.000	-0.295	0.768	-0.000	0.000
No_NHCau momentum	3.258e-05	0.000	0.305	0.760	-0.000	0.000
No_NHCau std	0.0002	5.39e-05	4.089	0.000	0.000	0.000
No_His momentum	-4.087e-05	6.3e-05	-0.649	0.516	-0.000	8.26e-05
No_His std	5.218e-06	3.73e-05	0.140	0.889	-6.79e-05	7.84e-05
BachelarPerc momentum	20.4271	0.720	28.387	0.000	19.016	21.839
BachelarPerc std	0.2919	0.513	0.568	0.570	-0.715	1.299
MedianHomeValue momentum	1.113e-09	2.47e-10	4.511	0.000	6.29e-10	1.6e-09
MedianHomeValue std	-4.419e-10	7.51e-11	-5.884	0.000	-5.89e-10	-2.95e-10
MedianIncome Momentum	6.828e-05	3.06e-06	22.314	0.000	6.23e-05	7.43e-05
MedianIncome std	-5.663e-06	1.99e-06	-2.840	0.005	-9.57e-06	-1.75e-06

Omnibus:	19.033	**Durbin-Watson:**	0.860
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	21.756
Skew:	-0.204	**Prob(JB):**	1.89e-05
Kurtosis:	3.405	**Cond. No.**	1.39e+13

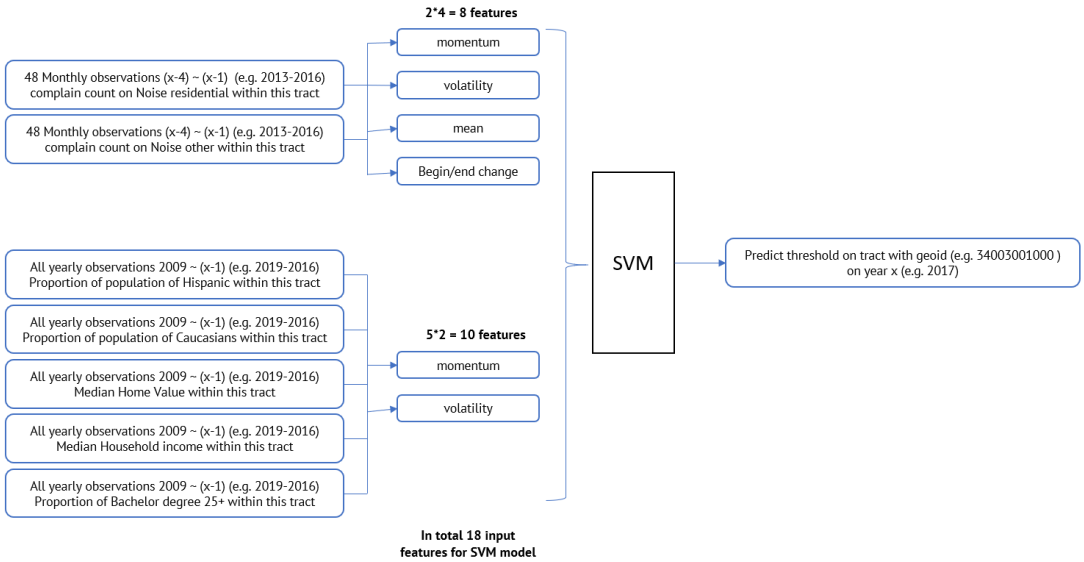
OLS Regression Results

Dep. Variable:	threshold	**R-squared:**	0.388
Model:	OLS	**Adj. R-squared:**	0.387
Method:	Least Squares	**F-statistic:**	306.6
Date:	Sun, 25 Oct 2020	**Prob (F-statistic):**	0.00
Time:	14:12:07	**Log-Likelihood:**	1619.7
No. Observations:	8720	**AIC:**	-3201.
Df Residuals:	8701	**BIC:**	-3067.
Df Model:	18		
Covariance Type:	nonrobust		

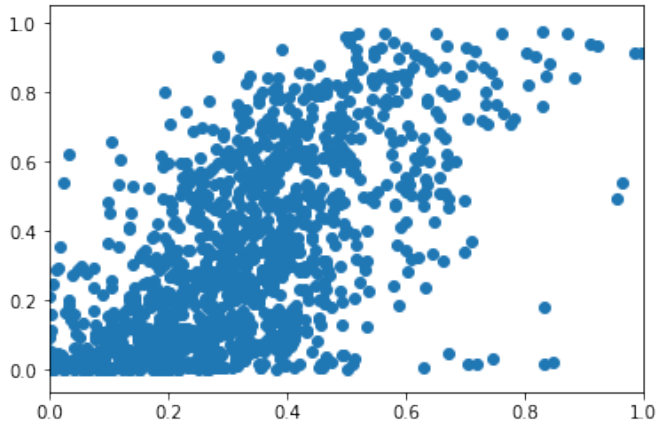
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2866	0.007	42.526	0.000	0.273	0.300
noiseother momentum	88.9072	216.230	0.411	0.681	-334.955	512.770
noiseother std	-9.1495	15.282	-0.599	0.549	-39.106	20.807
noiseother mean	54.6044	11.763	4.642	0.000	31.546	77.662
noiseother begin/end change	-1.875e-05	4.43e-05	-0.423	0.672	-0.000	6.81e-05
noiseresid momentum	-307.2743	208.803	-1.472	0.141	-716.578	102.030
noiseresid std	-6.4892	14.409	-0.450	0.652	-34.735	21.756
noiseresid mean	-15.2687	11.130	-1.372	0.170	-37.086	6.549
noiseresid begin/end change	8.796e-05	4.72e-05	1.865	0.062	-4.51e-06	0.000
No_NHCau momentum	0.0003	3.32e-05	8.415	0.000	0.000	0.000
No_NHCau std	7.687e-05	1.79e-05	4.285	0.000	4.17e-05	0.000
No_His momentum	0.0001	3.78e-05	3.802	0.000	6.97e-05	0.000
No_His std	-0.0001	2.08e-05	-5.136	0.000	-0.000	-6.61e-05
BachelarPerc momentum	19.4320	0.299	65.081	0.000	18.847	20.017
BachelarPerc std	-0.2928	0.213	-1.375	0.169	-0.710	0.125
MedianHomeValue momentum	1.256e-09	2.33e-10	5.382	0.000	7.99e-10	1.71e-09
MedianHomeValue std	-4.979e-10	6.83e-11	-7.293	0.000	-6.32e-10	-3.64e-10
MedianIncome Momentum	7.601e-09	1.36e-09	5.596	0.000	4.94e-09	1.03e-08
MedianIncome std	7.008e-10	4e-10	1.750	0.080	-8.41e-11	1.49e-09

Omnibus:	2.817	**Durbin-Watson:**	0.740
Prob(Omnibus):	0.244	**Jarque-Bera (JB):**	2.850
Skew:	0.040	**Prob(JB):**	0.241
Kurtosis:	2.961	**Cond. No.**	6.04e+12

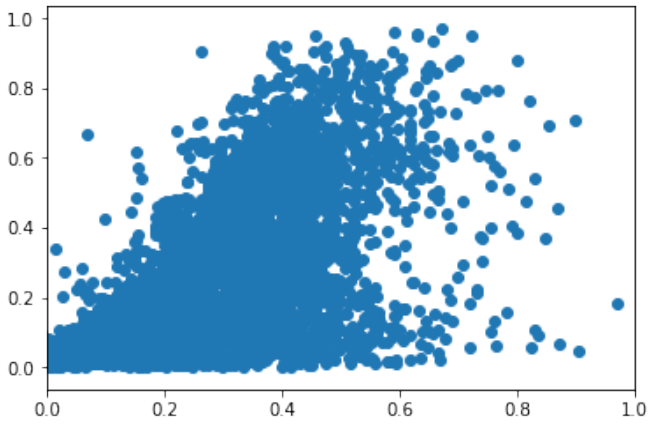
From these 2 OLS reports, the R-squared values are 0.599 and 0.388, which is significant enough to continue to a SVM regression model.



The predictions of the regressor against the true value is plotted. They is a clear positive correlation between our prediction and the true values.



Prediction result on testing dataset only on tracts that passed first test using Support Vector Machine regressor



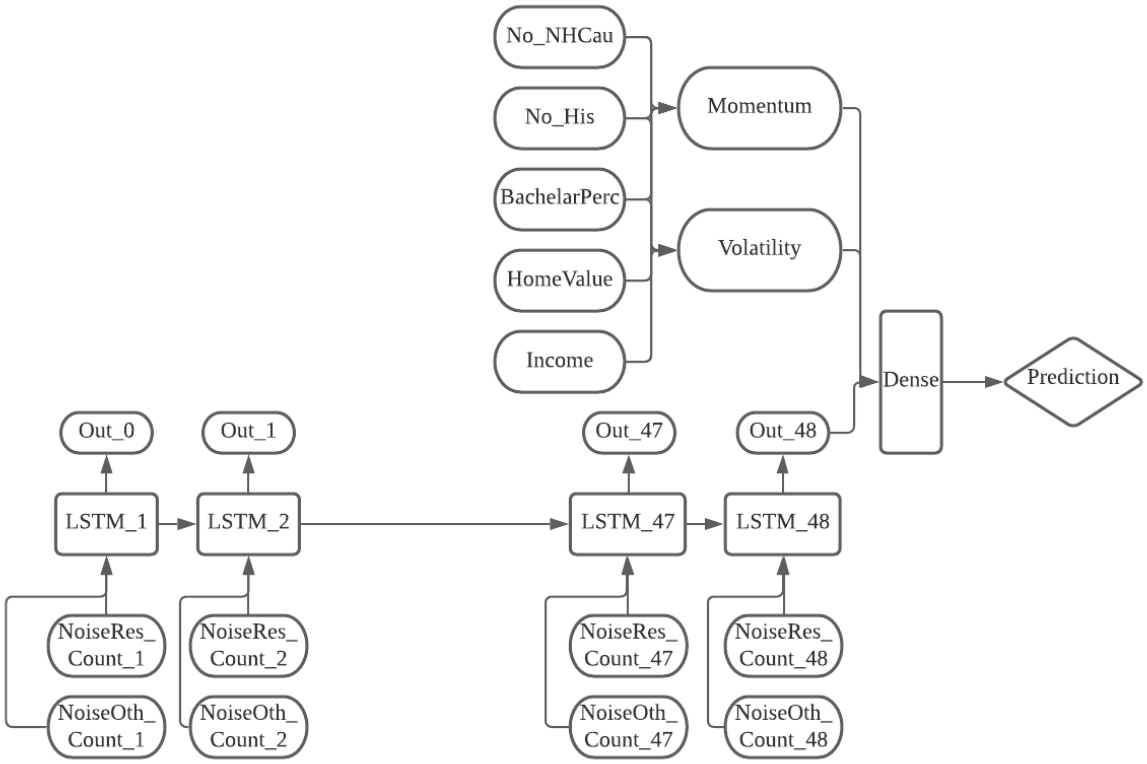
Prediction result on testing dataset on all tracts using Support Vector Machine regressor

Section 4: LSTM-DNN Model

Another approach to the prediction task is a LSTM-based Neural network. Looking at our data set, the number of complaints may be calculated monthly to have a time-series structure, and we may train a LSTM on them.

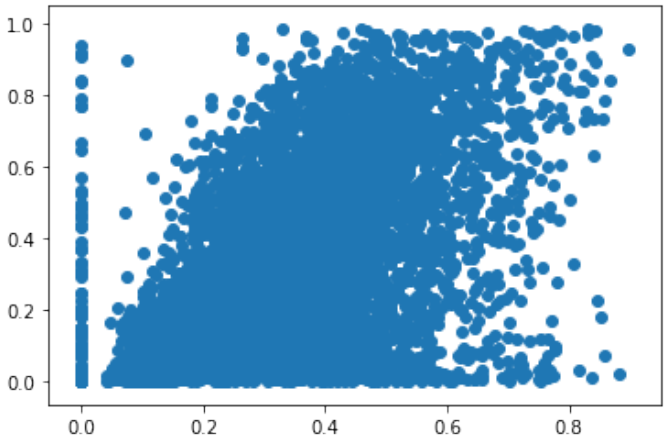
Given data on 1744 tracts from 2010 to 2018, instead of using the entire 8 years as a single sample to predict how gentrified the tract is at the end of 2018, we decided to slice them into shorter time-series with a length of 48 months, to predict the gentrification threshold at the end of this 48-month period. This means a single tract can be used as 5 correlated samples instead of a single sample, and the dimension of our data at this step becomes (1744*5, 48, 2). A train-test-split puts 20% of these rows into a test set.

We also want to include the static data without a time structure. This is why we came up with the following architecture:

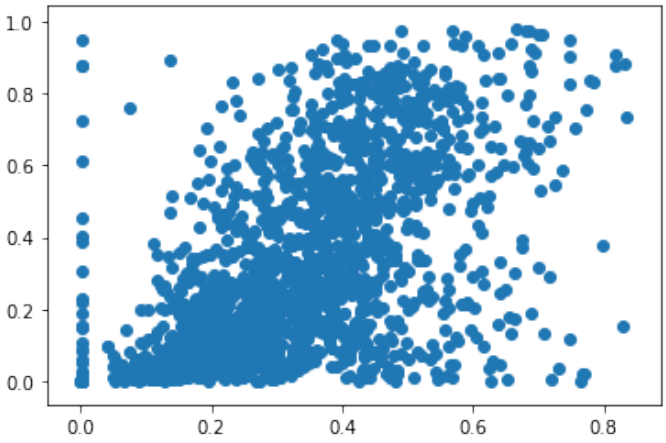


The LSTM has 2 inputs, 10 hidden cells and 1 layer. The last output of the LSTM is concatenated with the momentum and volatility of 5 other features, before going through a single dense layer. An arctan function is applied onto the output of this dense layer, which maps the output to range 0 – 1. This output is compared with the correct gentrification threshold, trained with Mean Square Loss.

Scatter plots of target against prediction for training and testing data set are plotted. Linear regressions give slope of 0.89 and 0.91, and R-square of 0.57 and 0.58.



Prediction result on training dataset on all tracts using LSTM-DNN model



Prediction result on testing dataset on all tracts using LSTM-DNN model

Section 5: Conclusion

In conclusion, we visualized the gentrification process of NYC, studied the feature importance and correlation to the gentrification process, achieved 0.636 R-square on gentrification process prediction with SVM, and 0.581 R-square on home price prediction with a neural network.