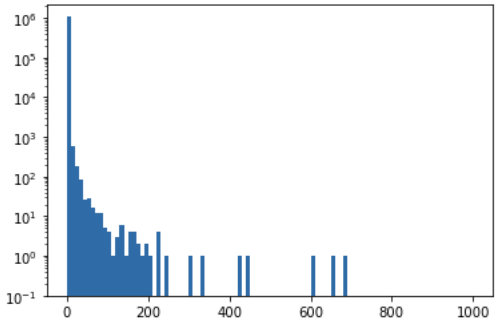
New York Property Data Model Modification Notes

Part I. Add special treatment for distances of 1

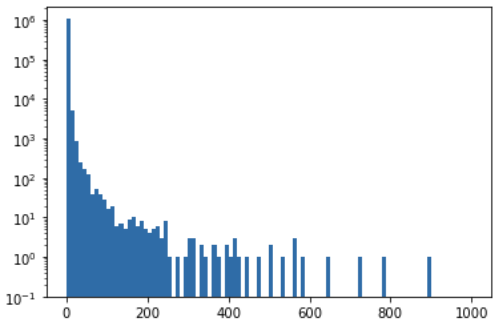
Treated the four variables' distances of 1 as missing value (nan) and ran the model. Returned the top 100 records and saved as file NY_top_100_Pt_1.csv.

Part II. Modifications on the fraud model

- Change the number of PC (from 6 to 7)
  - Considered changing number of PCs between 4 and 9:
    - PC of 9 explained variable ratio: [0.53169854, 0.78121611, 0.84502019, 0.8854793, 0.91587336, 0.9373666, 0.95148805, 0.96451998, 0.97015733]
  - Since the original choice of 6 already has an explained variable ratio of 0.9373, I decided to only increase the number of PC by 1 and proceed with PC of 7 and an explained variable ratio of 0.9515.
- Change the power of p's in score 1 and score 2:
  - For score 1, I decided to change it to the power of 4 as the histogram shows more condensed distribution with clear gap between low and high values.



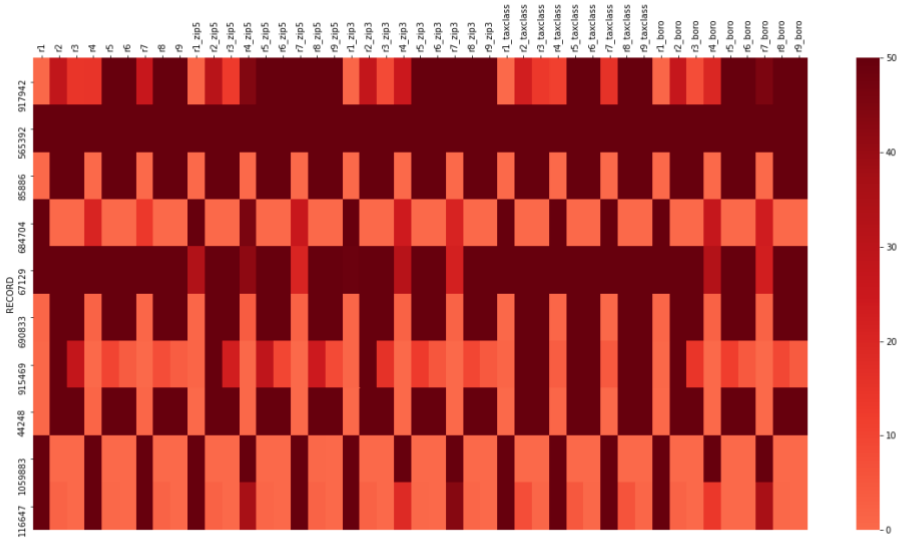
  - For score 2, I have changed it to power of 1 in order to contrast with score 1. The histogram reveals a more even distribution with less clear gap for middle scores but still maintains clear divides for higher scores.



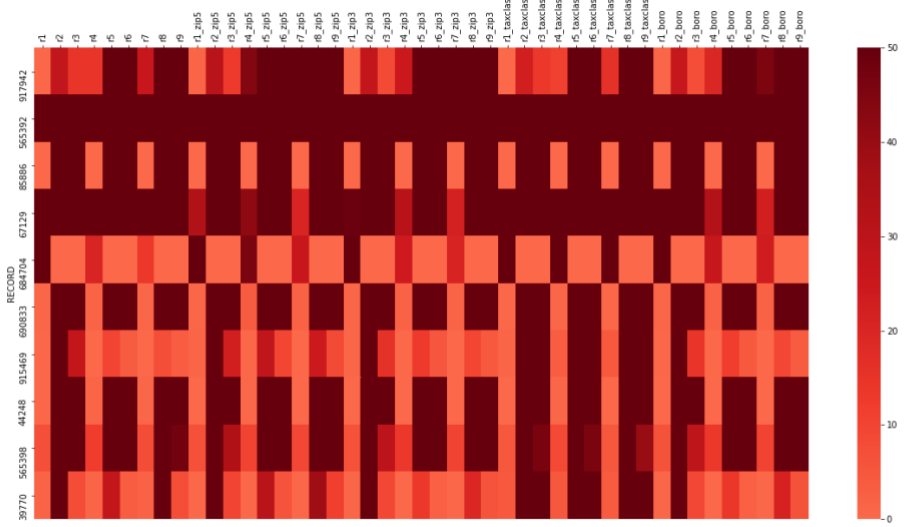
- Modify the autoencoder
  - The autoencoder has only been modified for shape from 6 to 7 to fit the new number of PC.
- Find the percentage difference in the top 100 records

- After comparing top 100 record number between the two models, the overlapping percentage is 95%.

- Explain why the top 10 new records are unusual (New Top 10 & Old Top 10)
  - First, I took a look at the z-scaled heatmap for top 10 records. As shown in the graph, there does not appear to have a consistent trend for unusual values. Therefore, each case should be investigated individually.



  - For record 917942 (ranked #1), record 565392 (ranked #2), record 684704 (ranked #4), record 67129 (ranked #5), 915469 (ranked #7), and 1059883 (ranked #9), the main abnormality is that all those properties don't have any building frontage or depth (BLDFRONT and BLDDEPTH all have values of 0).
  - For record 684704 (ranked #4), 915469 (ranked #7), and 1059883 (ranked #9), additional abnormality is in total and land market value as these three properties have values of 0. Record 684704 is a residential vacant land, record 915469 is government property while record 1059883 is an easement.
  - For record 85886 (ranked #3) and record 690833 (ranked #6), the abnormality is property types since both properties are parks and recreations with high market value but extremely small building sizes (similar to other records identified before but values are not 0).
  - For record 116647 (ranked #10), the abnormality is in lot sizes since the property is apartment with extremely small lot size values (LTFRONT and LOTDEPTH values are extremely small).



- In addition, there are two different records from top 10 of the Part I results. The records are 565398 (ranked #9) and 39770 (ranked #10). The abnormality in record 565398 is from its extremely high market value while having no building sizes, along with it being a government property. On the other hand, the abnormality in record 39770 is from its missing values and low market values while having huge lot sizes.