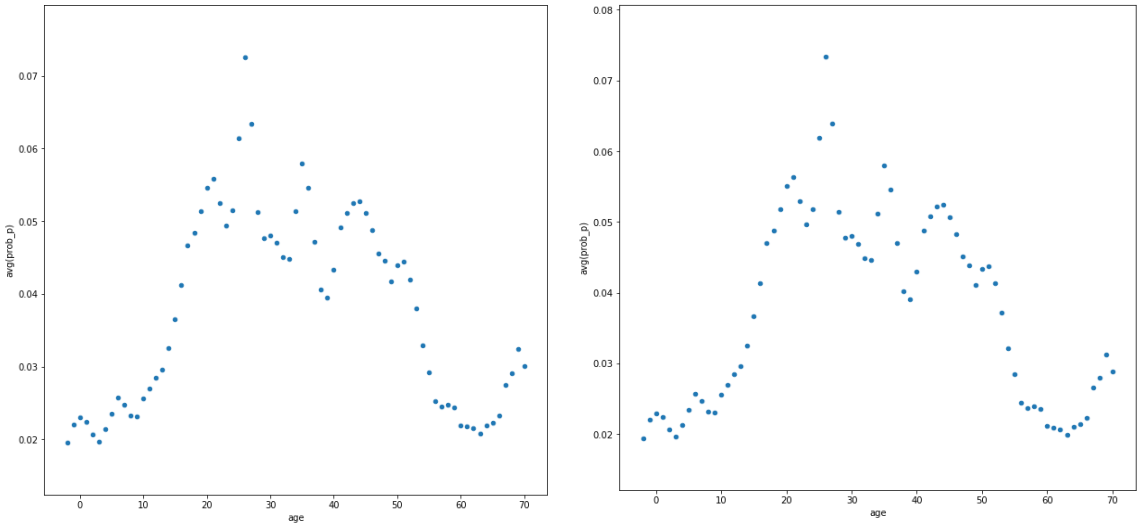
Week 10 big data assignment

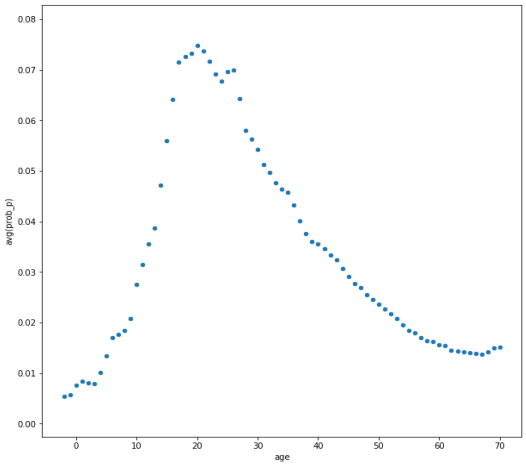
Group 4: Mengrui Zhang, Rohit Gala

1. The reason why we take two datasets as input to rescale sample since we want to create a new normalized sample and we need to keep we normalized training set and test set under same measure. As a result, we need to pass both test set itself and training set to normalize the test set. Similarly, we also pass two training sets to normalize the training set.

2. For MLP model, we need at least a three-depth layer. For example, a [8,3,3] layers. If we use a layer only contains two level-input layer and output layer, the MLP model will perform like the logistic model.

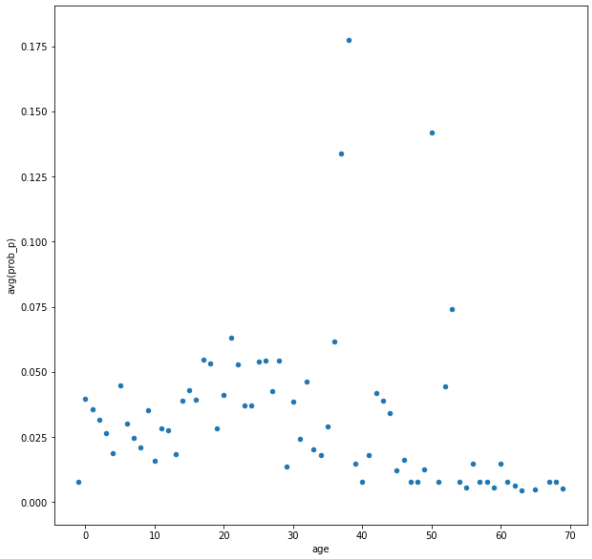


The graphs above are average prob_p for both MLP and logistic model. We can find that this two are identical, and not similar to the ITEM model. And if we add an intermediate layer, then our MLP model will perform more similar to ITEM model.



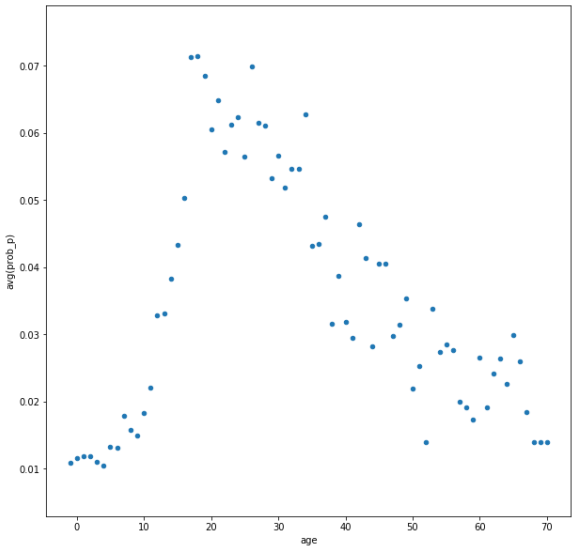
3. The sample size may extremely influence the MLP performance, especially when sample size is not big enough. For example, if we take the sample size as 1000, the result for MLP model is ridiculous.

```
MLP, 0.16210733644463807, 2.281259321294527, 0.163567873751, 0.167543239712
7A, 0.1895659708315342, 0.19666370926955482, 0.191748801476, 0.186534318272
MLP, 0.16210733644463807, 2.281259321294527, 0.163567873751, 0.167543239712
7A, 0.1895659708315342, 0.19666370926955482, 0.191748801476, 0.186534318272
```



When we increase our sample size to 3000, the MLP model gives a reasonable result, however its performance is still not good.

```
MLP, 0.1755648192853413, 0.20945216097185895, 0.1788274707, 0.180946047013
7A, 0.19117470175474285, 0.19689742078999548, 0.191803955493, 0.190586534009
MLP, 0.1755648192853413, 0.20945216097185895, 0.1788274707, 0.180946047013
7A, 0.19117470175474285, 0.19689742078999548, 0.191803955493, 0.190586534009
```

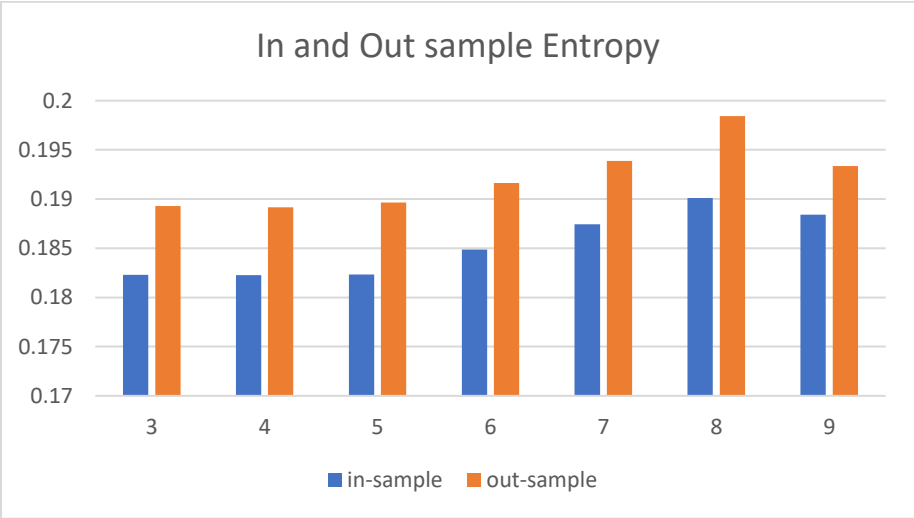


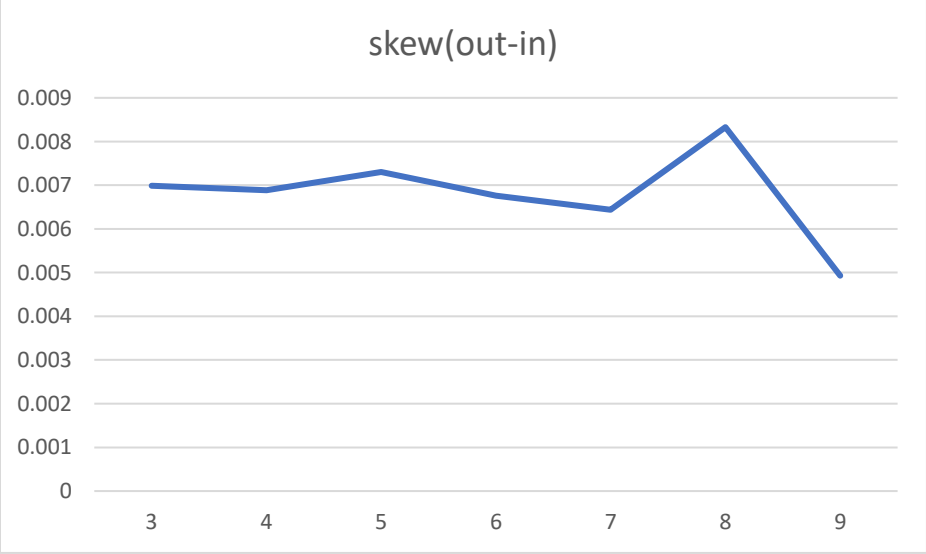
And when we increase the sample size to 7000, the performance of MLP model exceed than that of logistic model.

```
MLP, 0.19163378779134488, 0.19720473184351167, 0.193209711166, 0.192632705834
7A, 0.1960452186063877, 0.1981427664101968, 0.196656664802, 0.195875726113
MLP, 0.19163378779134488, 0.19720473184351167, 0.193209711166, 0.192632705834
7A, 0.1960452186063877, 0.1981427664101968, 0.196656664802, 0.195875726113
```

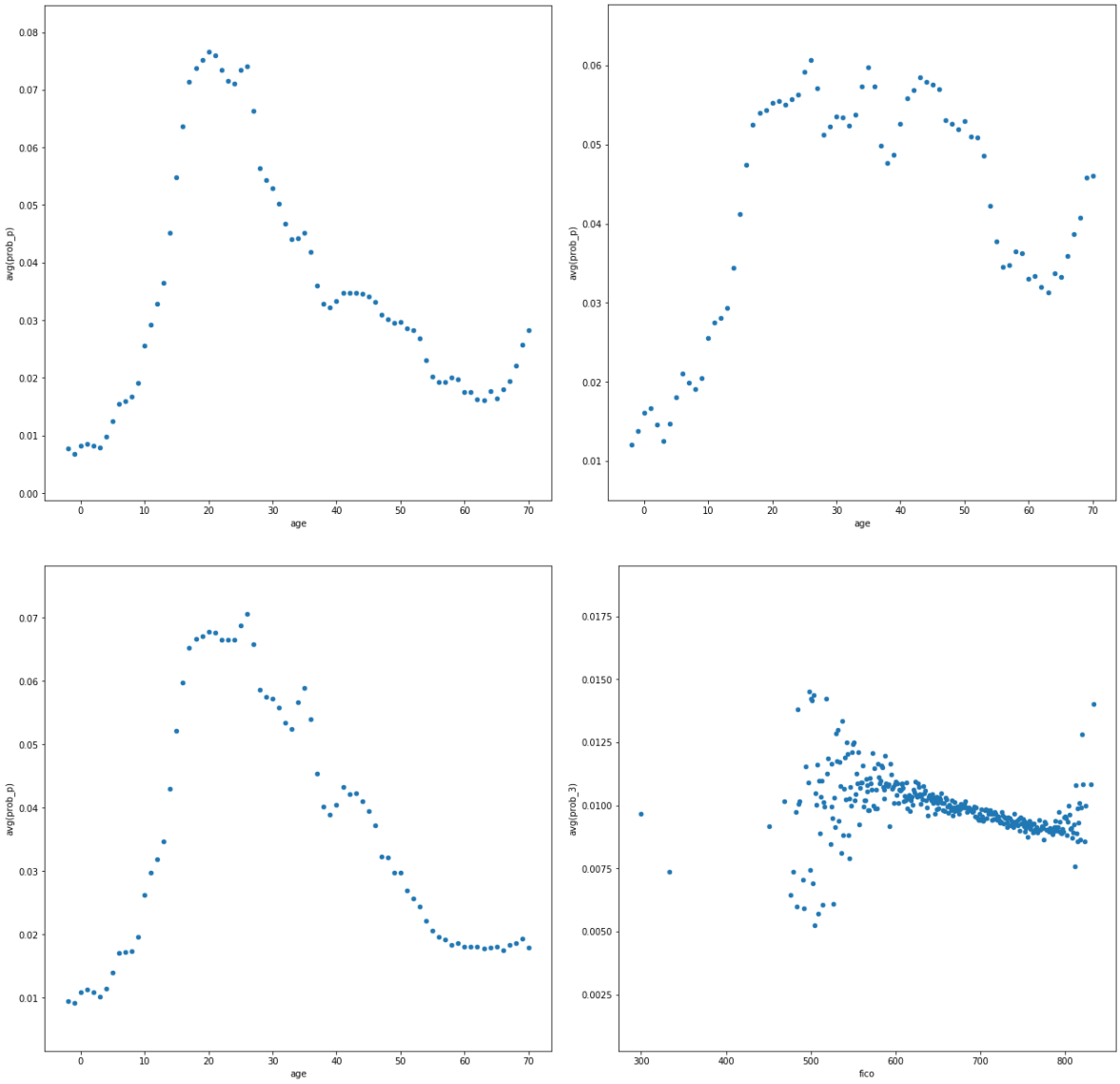
As a result, in order to get an optimized result for MLP model, we should at least support a perhaps 10k sample records.

4. I try to do the in-sample error and out-sample error, and it happens that our model classifies all points into one group, then the error is constant. So that I try to plot skew of entropy as the error.





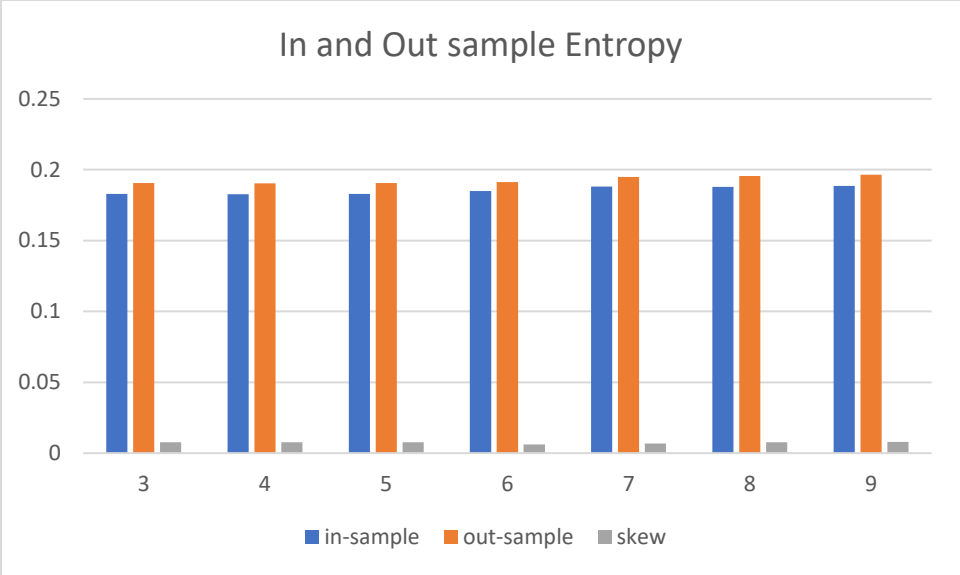
A very interesting phenomena happens here, even number layers performs very bad and odd number layers performs somehow good.

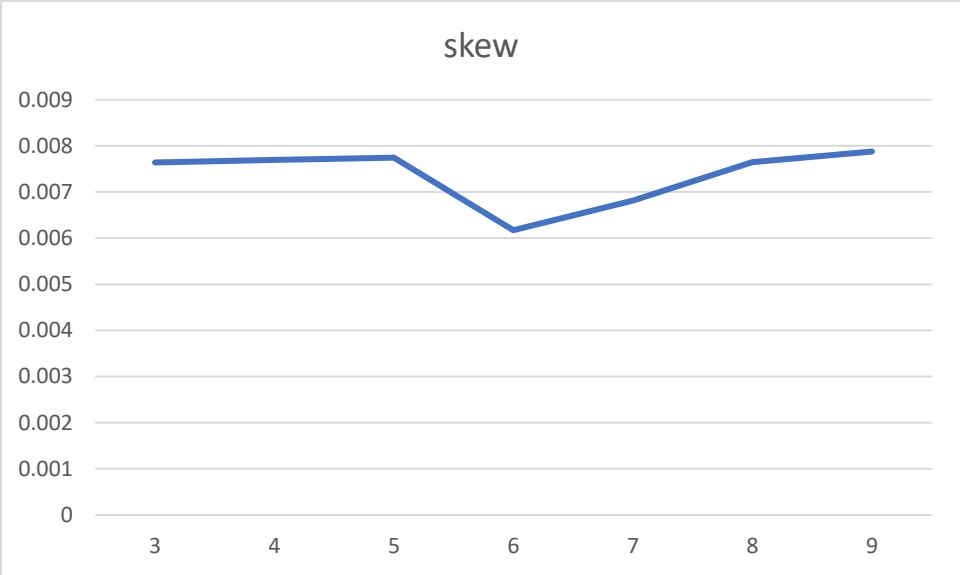


Left-up 7 layers, right-up 8 layers, left-bottom 9 layers, right-bottom 9 layers

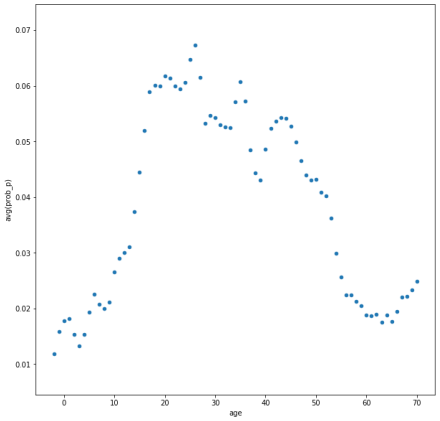
We may find that the graph with 7 and 9 depths still look like ITEM model, but the graph with 8 depth looks ridiculous; however, the fico graph under 9-depth is terrible. As a result, the best model may be 7-depth model.

5. 80K as sample size.

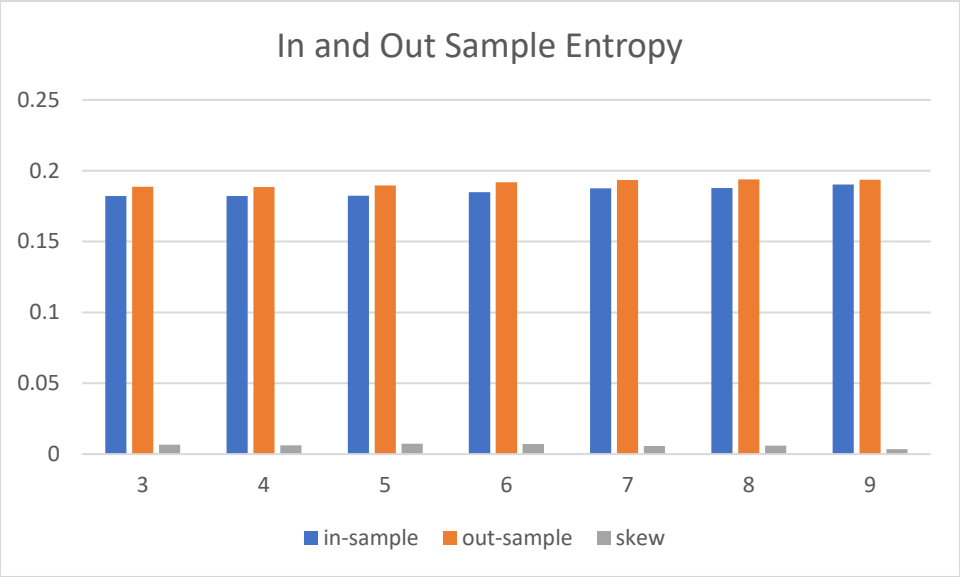


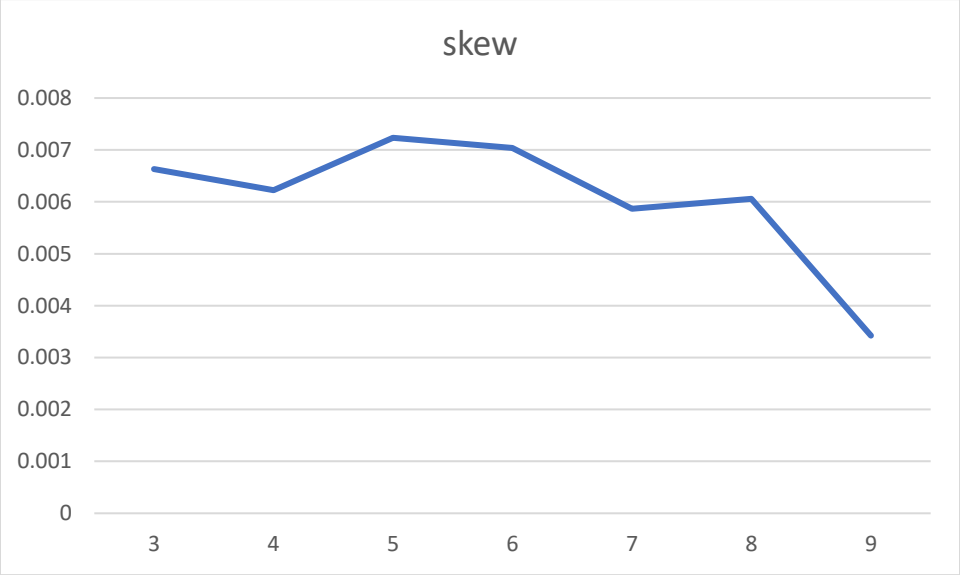


For this smaller sample size, a smaller depth may be better. As a result, it seems that 6-depth model is best. And it seems that it make sense since the graph below is not too overfitting or underfitting.



Use 150K as sample size.





We can conclude that the large the sample size is, the more potential depths may be supported, however, there is also potential overfitting problem, we don't detect.

6. Since I don't find a way to change the ITEM model's graph, the only important features we can not easily replicate other than the small peak around age 8, is that a sharp decrease around fico 750

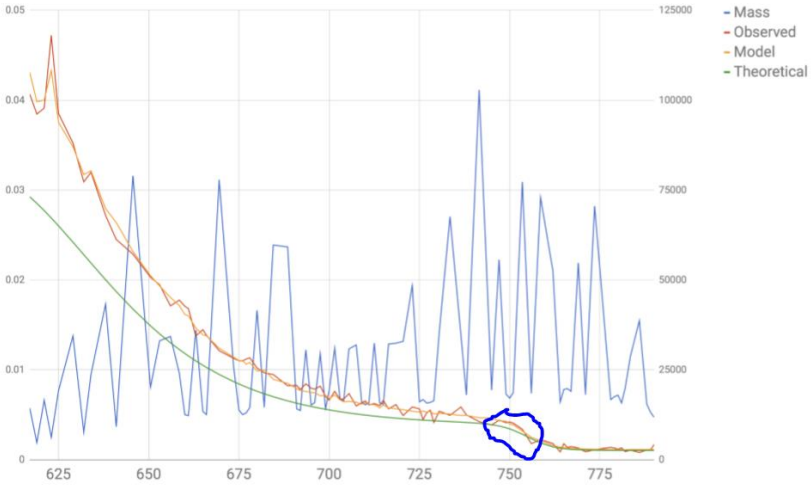
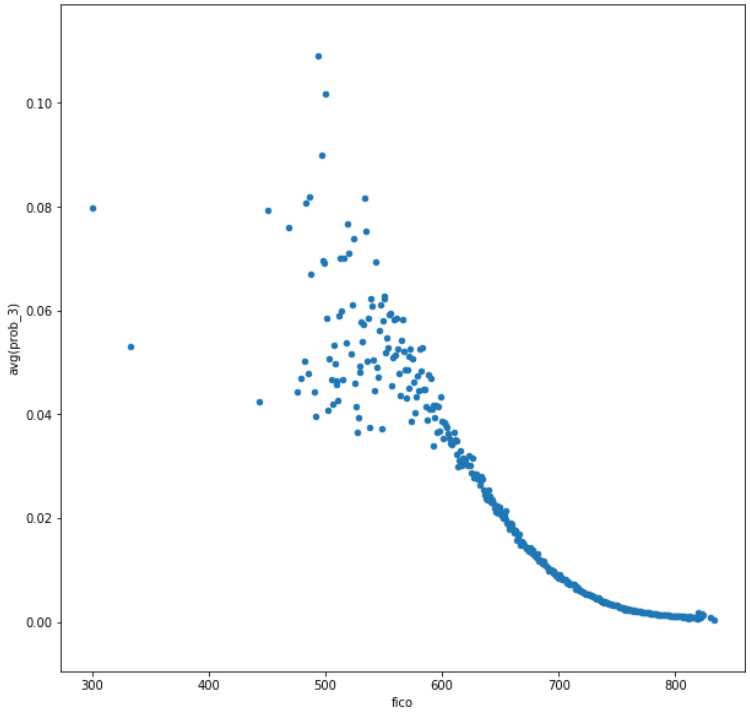
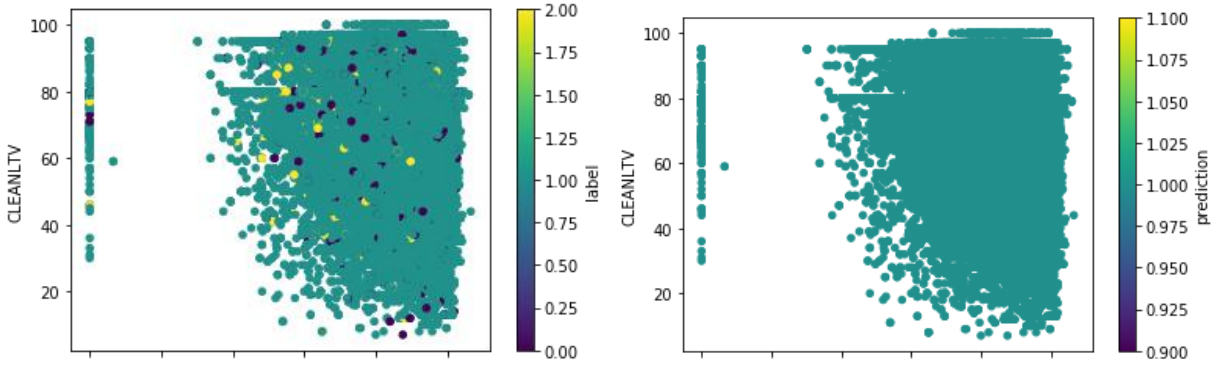


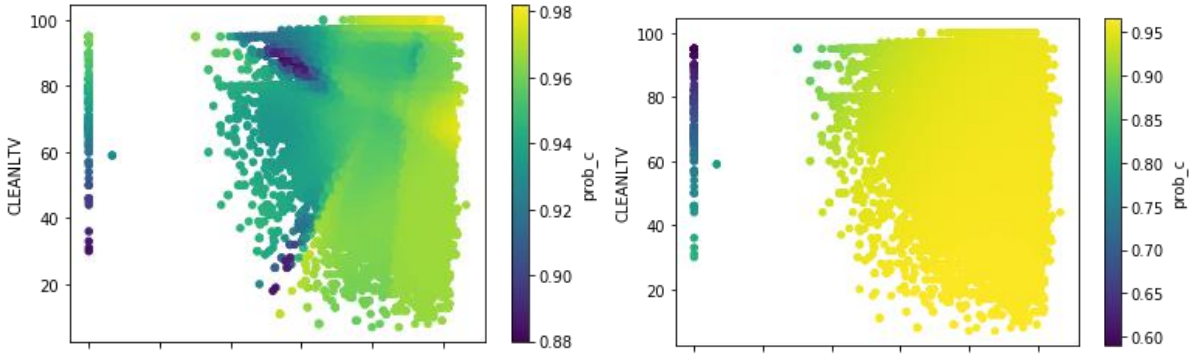
Figure: C to 3 by Fico (TrueModel)



7. For this part, see the graph below.



These two graphs show the classification of our model and true sample. We can find our models-both MLP and logistic model, classify all points into one group. As a result, it may hurt, since our true sample does have two other labels. In other word, for this high skewness sample, we may need some control to help us make a better predict or classification.



Left: MLP, Right: Logistic

But if we look insight into the probability the two models give, we may find a very interesting result. We can find that color changing in logistic graph is linear for whole graph, however it has different pattern in MLP model. As a result, we may think that, MLP can deal with a more difficult model, for it has different pattern for different place. In other word, the changing of color may not follow only one trend. But it also give us an alert, that how to control or penalize it to avoid overfitting.