10.1 Preprocessing Housing Price Dataset

2:30 talks about pandas-profiling (which is to print a better report, but not a must in our course)

5:00 the data type in pandas profiling figure might be wrong (it determines if it's categorical simply based on the type of the data, say string. But some strings can be ordinal data e.g. Excellent/Good/Fair/Poor, so be careful about that)

In the lecture (actually related to chapter 09):

1.PR(Precision-recall) curve (x-axis is precision and y-axis is recall): AP score is the area under the PR(Precision-recall) curve

~ AP score (Average precision) is independent of the threshold because it actually goes through every single threshold to plot the curve but other metrics are subject to threshold changing (E.g. Recall)

2.receiver operating characteristic (ROC) curve: more popular

it considers all possible thresholds for a given classifier given by predict\_proba but instead of precision and recall it plots false positive rate (FPR) and true positive rate (TPR or recall).

https://ubc-cs.github.io/cpsc330/lectures/09\_classification-metrics.html#receiver-operating-characteristic-roc-curve

In here we see that when we lower our threshold, recall increases, however, when we keep lowering the threshold to a low point, we see that we are having more false positive. (more towards Top left is better)

~AUC (Area under the curve): evaluating the tanking of positive examples.

1.0: all positive points have a higher score than all negative points

0.5: means random chance

Lecture 10 - regression\_metrics

Alpha in Ridge(): higher = simpler model, lower = more complex model, which is exactly the reverse of C in linearRegression

1. Mean Squared Error(MSE)

A perfect predictions would have MSE = 0

2. Root mean squared error (RMSE)

sqrt(MSE)

3. R-square

4. MAPE: percent\_errors = (pred\_train - y\_train) / y\_train \* 100.0