Data

Target distribution at last submission

Target (scaled_categorized app rating)	1	2	3	4	5
Counts	8169	115338	16130	634	153
Percenatge	5.82%	82.14%	11.49%	0.45%	0.11%

Changes I made since last submission:

1. Regroup target classes, 1,2,3 stay the same, 4 &5 are combined into one class '4+',so new distribution in table

Target distribution after regrouping

Target (scaled_categorized app rating)	1	2	3	4+
Counts	8169	115338	16130	787
Percenatge	5.82%	82.14%	11.49%	0.56%

2. Resampling (combining Oversampler and Undersampler)

First I used RandomUndersampler downscale classes 1-3 to size 2058, and then I used Oversampler(RandomOverSampler and SMOTENC) upsampled class '4+' to size 882. There is no noticeable change of performance when change from RandomOversampler to SMOTENC.

Note: 882=787*(1+40%)
2058=882*7/3

Target distribution after resampling

Target (scaled_categorized app rating)	1	2	3	4+
Counts	2058	2058	2058	882
Percenatge	29.2%	29.2%	29.2%	12.5%

3. Train_test are split using stratification by setting "stratifyy=y"

4. All classifier have parameter "class_weight='balanced'"

Reports

- Base model (Tree) without correction for class imbalance:

	precision	recall	f1-score	support
0	0.10	0.38	0.15	1634
1	0.88	0.60	0.72	23068
2	0.30	0.50	0.37	3226
3	0.15	0.31	0.20	157
accuracy			0.58	28085
macro avg	0.35	0.45	0.36	28085
weighted avg	0.76	0.58	0.64	28085

- Base Model after Corection for Imbalance:

	precision	recall	f1-score	support
0	0.08	0.53	0.15	1634
1	0.89	0.42	0.57	23068
2	0.25	0.47	0.32	3226
3	0.09	0.52	0.15	157
accuracy			0.43	28085
macro avg	0.33	0.48	0.30	28085
weighted avg	0.77	0.43	0.52	28085

- RandomizedSearchCV Best RandomForest Model with Correction for Class Imbalance

	precision	recall	f1-score	support
0	0.09	0.52	0.15	1634
1	0.90	0.43	0.59	23068
2	0.26	0.54	0.35	3226
3	0.10	0.56	0.18	157
accuracy			0.45	28085
macro avg	0.34	0.51	0.32	28085
weighted avg	0.78	0.45	0.53	28085