5th Place Solution - An ensemble of 53 OOFs

TLDR: Mostly stayed away for the first two weeks, participated more or less normally for the next 10-12 days, had many plans for the last few days but ran out of time as other things came up.

I had three solutions which scored > 0.385 on the private LB, the one with the highest CV scored 0.38509 and would have secured 4th place. But I didn't choose it, the one with the best public LB (0.38337) scored 0.38502 and got me fifth place. I watched with bemusement as I dropped about a 100 places from #2 to about #102 or so over the past few days, as the blendy blenders went about systematically overfitting the public LB, and the floaty followers formed a solid block (or few) of several dozens, pushing the rest of us artificially behind for a few days - so there's a certain satisfaction in being on the right side of the shakeup.

Drivers of diversity in the ensemble

1. Model type: While most of my models were XGBs or LGBMs, the ensemble also included CatBoost, HGB, Neural Nets, Random Forests and Logistic Regression. I didn't get around to adding other models like SVMs.
2. Playing around with Hyperparameters: Many people use GOSS and DART variants of LGBMs, but one can also play around with various hyperparameters e.g. depth, number of estimators, max leaves, scoring function, evaluation metric etc.
3. Augmentation with the original data : I used varying number of copies of the original dataset, from 1-6. I avoided 7 and above because after a point, you're biasing away from the competition data towards the original (e.g. 8 copies was literally more than the training data).
4. Using numerical data as categorical: Treating all data as categorical as well as treating numerical features as both numerical and categorical helped, with or without binning.
5. Varying fold numbers: Varying the number of folds in k-fold CV can also add to diversity.

Things I started too late

1. Treating numerical data as categorical and using models other than XGBs and LGBMs (esp. NNs) all boosted the CV and usually LB as well - I should have added more such models.
2. I should also varied more hyperparameters, esp. with XGBs, and added more CatBoost models.

Things I never got around to doing

A long list, but one that I considered early and probably should have prioritized was to repurpose @cdeotte's winning solution from February, since the original dataset had nearly no signal, just like that month.

Finally, I'd like to thank those who shared code and insights, including but not limited to

@ravaghi, @suikeitin, @masayakawamata, @hahahaj, @omidbaghchehsaraei, @pirhosseinlou, @ayushchandramaurya, @patrykszcz, @gauravduttakiit

and congratulate @masayakawamata, @mahog, @hahahaj, @paperxd, our resident Topper-in-chief @cdeotte, and everyone who finished strong. Happy Kaggling!