3rd place - Ridge and CV are all you need

TLDR:

- Slowly built an ensemble of 60+ models over the month
- Searched for a good ensembler in the first 2 weeks
- Relied on the various XGB hyperparameters in public notebooks
- Trust CV

This month I focused on having a wide variety of different models (like XGB vs NN) instead of the same model with different FE (there is still some FE in here though). Here are some of the tricks I used:

Ridge as ensembler

At the beginning of the comp, I used HC as the ensembler. But it had an issue: adding a model often decreased the CV score. Then I tried genetic algorithms (GA). It worked pretty well up to about 20 models, but it had the same issue as HC after going beyond that. After searching for a while, I settled on Ridge. It was very fast (about 1 minute for my 60 OOFs) and adding a model almost always improved the CV score, so I used it to the end.

Some notes on using Ridge:

- The oofs were stacked in the shape of (n_samples, n_models * 7)
- One-hot encode the target and use multi-target regression

Turn classification into regression

I trained a few XGBs and NNs using multi-target regression with MSE loss by one-hot encoding the target. I would have trained more regression models like this but ran out of time.

XGB with product features, label encoded

Late into the competition, I decided to use product features to add diversity, but I accidentally label encoded the new features. It worked surprisingly well and boosted CV by 0.0001.

Diverse libraries

During my quest for the most amount of unique models, I found some pretty cool libraries:

- pytabkit
- pytorch-tabular

- rtdl_num_embeddings

Without these libraries I would probably not be in top 10

Conclusion

I first want to thank all the people who have helped me reach this far, including but not limited to: @cdeotte, @optimistix, @siukeitin, @ravi20076, @masayakawamata, @omidbaghchehsaraei, @ravaghi, @yekenot (I probably missed a lot of people). Through these past 3 months, I have managed to get 3 top 10 finishes, some Kaggle Swag, and an important learning: "CV and Ridge is all you need".

Additional notes

My final list of models:

- 3 "all-embedding" NNs based on here
- 1 Autogluon
- 5 Catboosts
- 1 Gandalf NN from pytorch-tabular
- 2 GRNs based on here
- Some LAMA NNs
- 5 LGBMs
- 1 LNN based on here
- 3 Logistic Regression models based on here
- 1 normal NN from here
- OOFs from here
- 1 RealMLP from pytabkit
- 3 Pytorch NNs with rtdl_num_embeddings
- 1 SAINT model
- 1 TabTransformer model from here
- 3 TabMs from pytabkit
- 20 XGBs

- 1 ExtraTrees model
- 1 GradientBoostedTreesLearner from ydf

I had also managed to select my 2 best submissions on the private LB, and funnily enough, my best CV was the second best on private LB, and my best public LB was my best private. Never seen that happen before.