details.

Final Model

Our final model was a combination of 3 high scoring single models. CatBoost (Public/Private LB of 0.9639/0.9408), LGBM (0.9617/0.9384), and XGB (0.9602/0.9324). These models were diversified because Konstantin built the CAT and LGB while I built the XGB and NN. And we engineered features independently. (In the end we didn't use the NN which had LB 0.9432). XGB notebook posted here.

One final submission was a stack where LGBM was trained on top of the predictions of CAT and XGB and the other final submission was an ensemble with equal weights. Both submissions were post processed by taking all predictions from a single client (credit card) and replacing them with that client's average prediction. This PP increased LB by 0.001.

How to Find UIDs

We found UIDs in two different ways. (Specific details here).

- Wrote a script that finds UIDs here
- Train our models to find UIDs here and here

If you remember, Konstantin's original public FE kernel here without UIDs achieves local validation AUC = 0.9245 and public LB 0.9485. His new FE kernel here achieves local validation AUC = 0.9377 and public LB 0.9617 by finding and using UIDs. Soon I will post my XGB kernel which finds UIDs with even less human assistance and proves to beat all other methods of finding UIDs. (XGB posted here). The purpose of producing UIDs by a script was for EDA, special validation tests, and post process. We did not add the script's UIDs to our models. Machine learning did better finding them on its own.

EDA

EDA was daunting in this competition. There were so many columns to analyze and their meanings were obscured. For the first 150 columns, we used Alijs's great EDA here. For the remaining 300 columns, we used my V and ID EDA here. We reduced the number of V columns with 3 tricks. First groups of V columns were found that shared similar NAN structure, next we used 1 of 3 methods:

- We applied PCA on each group individually
- We selected a maximum sized subset of uncorrelated columns from each group
- We replaced the entire group with all columns averaged.

Afterward, these reduced groups were further evaluated using feature selection techniques below.