Project thoughts

* Two data sets, one for transactions and one for identities associated with transactions. Not all transactions of associated identities. Each transaction id in both data frames are unique (within the respective data frame)
* Identities has ~144,000 training data points and 41 features:
  + TransactionID -> links the identity with a transaction
  + Id\_01 THROUGH Id\_38 -> Network connection and digital signature information. All this information is purposefully anonymous
  + Device\_Type -> ‘mobile’ or desktop’ (as well as NA)
  + Device\_Info -> over 1800 different entries specific to device
* Transactions has ~590,00 transactions and 393 features (does not include isFraud)
  + TransactionID -> for matching with identities data
  + TransactionDT -> time of transaction (in reference to an consistent but arbitrary point)
  + TransactionAMT -> amount of transaction in US dollars
  + ProuductCD -> the product associated with the transaction (5 options), probably more like differentiating services than
  + Card1 THROUGH card6 -> card information
    - Card1 -> number between 1 and 20000
    - Card2 -> number between 1 and 600
    - Card3 -> strange distribution, mostly 150, some 185 scattering of other numbers
    - Card4 -> discover, visa, master card, American express or NA
    - Card5 -> another strange distribution between 100 and 240
    - Card6 -> credit, debit, card charge, credit or debit, NA
* Amounts with more than two decimals are probably made with foreign money
* distances are between (but not limited to) billing address, mailing address, zip code, IP address, phone area, etc
* addr1 as billing region, addr2 as billing country
* some transactions are from the same card or from the same account, or both the same
* create feature where purchaser and recipient email domain are different
* With xgboost, if there is signal in the missingness of the data, this is fit by the model
* NEED GOOD CROSS VALIDATION STRATAGY
* Sklearn TimeSeriesSplit
* Issue is that relationships are changing and the test set is all after the training set in time series
* “I do CV using time series split and use entire train for final training, the number of rounds is the one of the latest\largest fold times a multiplier that i tweaked after some tests.”
* sample\_weights
* gpu\_hist for tree\_method
* The Pemaildomain tells a lot about the isFraud. 95% transaction made by protonmail.com, isFraud=1 Also, 40% transaction made by protonmail.com in Remaildomain, isFraud=1
* During labeling the categorical values, protonmail.com tends to come up in others. Instead use this as another label. This gained me +0.120.

Columns that dummify

* ProductCD (okay)
* Card4 (okay)
* Card6 (okay)
* **P\_emaildomin (deal with!) -> manually separate into buckets based on risk**
* **R\_emaildomain (deal with!) -> meanually separate into buckets based on rist**
* **Id\_30 (deal with!) -> browser version, separate into old and new**
* **Id\_31 (deal with!) -> browser version, separate into old and new**
* **Id\_33 (deal with!) -> screen\_resolution, ?**
* Device\_type (okay) ->
* **Device\_Info(dealwith!) -> ???**

Links

* Weight more recent observations - <https://datascience.stackexchange.com/questions/9488/xgboost-give-more-importance-to-recent-samples>