08/11/2021

1. (May have to run on imbalanced dataset)
2. Re-run initial/raw XGB with RandomForestClassifier instead
3. Instead
   1. We did some kind of feature engineering (dropping columns)
      1. We also downsampled bc we saw imbalanced dataset
   2. Then we did some more feature engineering (cleaning + added columns)
      1. Saw NO improvement
      2. Decided to use feature importance to pick the top features for next dataset
   3. Used simplified dataset (drop almost all of the columns)
      1. Saw NO improvement
   4. Tune the models
      1. Saw improvement?
   5. This is the best our model can do
   6. 08/16/2021:
      1. Include the total payment amount in the dataset
         1. Will need to drop before running models if we run again
      2. 08/19/2021:
      3. Run with 6 features and tune
         1. Shared in Slack
      4. Think about description/proposal (informal) for calculation/results
         1. ROI (or other calc) for Downsampled and the data that was cut out before the downsampling and feature engineering occurred
         2. Calculate ROI for Full Dataset (will need to predict on full dataset if possible)
      5. 8/23/2021
         1. Implement our project ideas using code
            1. Ali – Work on both proposals
            2. Viviann – Work on her proposal
            3. Theodore – Analysis on false pos/neg and run SVM on both
      6. 8/31/2021
         1. Split up data based on Grades, Purpose, and Loan Amount
            1. Roi = (total amount – funded amount)/funded amount

Should be run prior to models

Baseline is 0.03

Also run the roi on the train set

This will be the FINAL baseline

* + - * 1. Ali – revisit the baseline roi calculation to see why it is so high

Also use new roi calc for for-loop simulation comparison

* + - * 1. Vivann – revisit her calculation for expected amount
        2. Theodore – work on adding roi to dataset

Check downsample to see if roi is still low

* + 1. 9/1/2021
       1. Theodore – create another dataset using the prediction set
       2. Viviann – look through data exploration
       3. Ali – analysis for roi
       4. For after running the model
          1. sum(df[pnl])/sum(df['funded\_amnt'])
    2. 9/7/2021
       1. Ali – familiarize with code & share code with group
       2. Viviann and Theodore – look at Ali’s code

1. Predict the model to check if our model is viable
2. Create current loan dataset
   1. Use this