*DATA CLEAN

1.creat data-process info
   (1) understand the intuition of each feature
   (2) categorize them by leakage information, joint feature and numerical.
   (3) calculate missing percentage
   (4) set up methods of outlier disposal

2.load data and data-process info.

3.set up sample ratio and select sample from raw_data.

4.drop null data:
   (1) delete the row whose 'desc' is null.
   (2)'loan_status', delete 'Does not meet the credit policy' and null in this feature.
   (3) delete the rows whose features are all null.

5.convert str number to float:
   (1) convert 'revol_util', 'sec_app_revol_util' from XX% to a value between 0 and 1.

*FEATURE CLEAN

1.replace single feature with joint feature:
   (1) replace 'annual_inc','dti','verification_status','revol_bal' by joint if exists.
   (2) For some features about two applicants, add their values with 'sec_app_'
   (3) delete initial joint features

2.remove 'policy code' feature, and drop 'earliest_cr_line', 'sec_app_earliest_cr_line' features for now

3.add TfidfVectorizer of 'desc' to raw_data

4. process the feature according to the instruction in col_info
   (1) delete leakage information
   (2) deal with outlier and discretization

5.remove features if they contain 50% missing values
   delete ['mths_since_last_record', 'mths_since_last_major_derog',
           'mths_since_recent_bc_dlq', 'mths_since_recent_revol_delinq']

7.merge none, any to other in 'home_ownership'

*FEATURE ENGINEERING

1.transfer zipcode to 'mean_household_income' by first three number.

2.transfer 'issue_d' to 'confi_ind'

3.transfer 'state address' to' price_level'

4.compute extra feature:
  (1) the ratio of satisfactory accounts
  (2) the ratio between the number of revolving trades with balance > 0
      and the number of currently active revolving trades
  (3) loan amount / annual income
  (4) loan amount / annual income
5.OneHotEncode categorical feature

*Model Training

1.standerdize numerical features

2.test and train split

3.use GridSearchCV find best parameter of each model
  Logistic, XgBoost, Neutral Network, Random Forest, Decision Tree

*Improvement:

1.use selected feature to fit selected model.
  (1) select the model with best AUC
  (2) rank the features by importance in Random Forest, and select top 25.

2.the intuition of CountVectorization() is more meaningful than TfidfVectorization() in NLP

3. discretization