Combined file

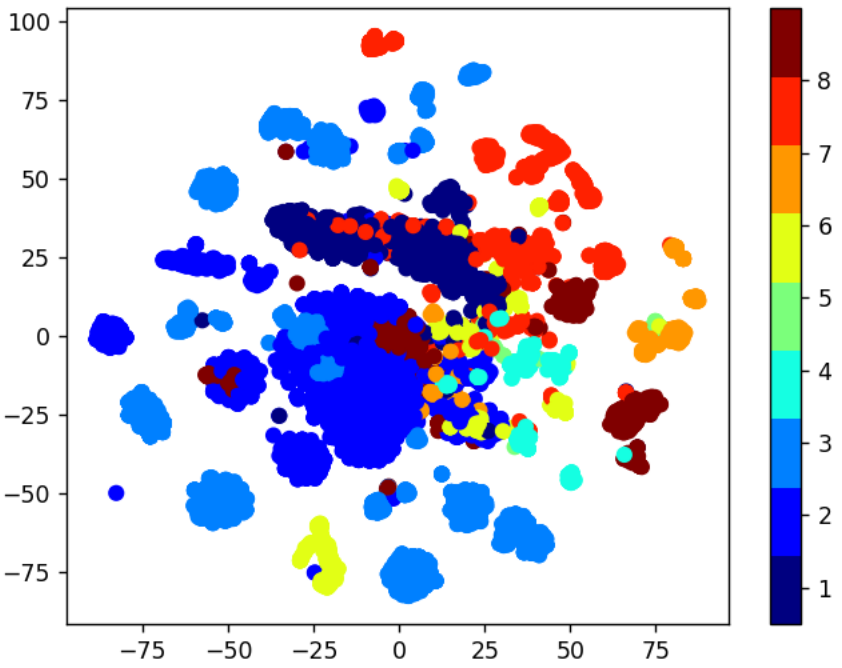
Created	@Jun 10, 2021 11:36 AM
Tags	

Exploratory Data Analysis

Merging both asm and byte file features

ID	0	1	2	3	4	5	6	7	8	9	...	edx
0	0.262806	0.005498	0.001567	0.002067	0.002048	0.001835	0.002058	0.002946	0.002638	0.003531	...	0.0154
1	0.017358	0.011737	0.004033	0.003876	0.005303	0.003873	0.004747	0.006984	0.008267	0.000394	...	0.0049
2	0.040827	0.013434	0.001429	0.001315	0.005464	0.00528	0.005078	0.002155	0.008104	0.002707	...	0.0000
3	0.009209	0.001708	0.000404	0.000441	0.00077	0.000354	0.00031	0.000481	0.000959	0.000521	...	0.0003
4	0.008629	0.001	0.000168	0.000234	0.000342	0.000232	0.000148	0.000229	0.000376	0.000246	...	0.0003

Multivariate Analysis on final features



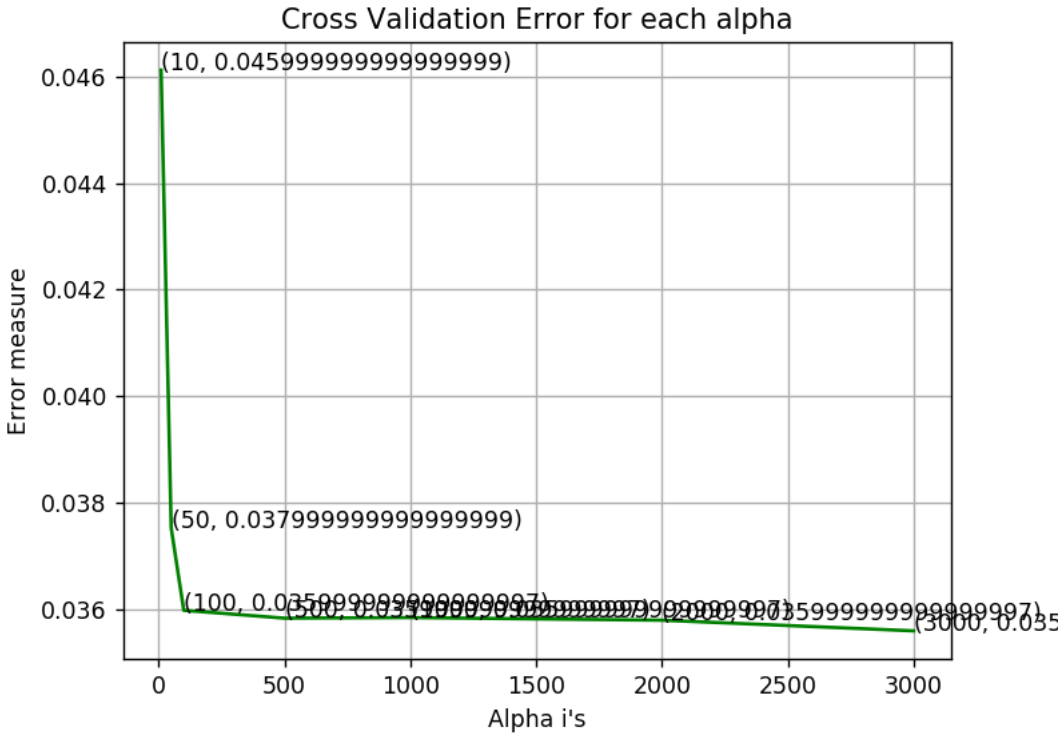
Train and Test Split

Machine Learning Model

Random Forest Classifier

Hyperparameter search

log_loss for c = 10 is 0.0461221662017
log_loss for c = 50 is 0.0375229563452
log_loss for c = 100 is 0.0359765822455
log_loss for c = 500 is 0.0358291883873
log_loss for c = 1000 is 0.0358403093496
log_loss for c = 2000 is 0.0357908022178
log_loss for c = 3000 is 0.0355909487962



Results from the Best model

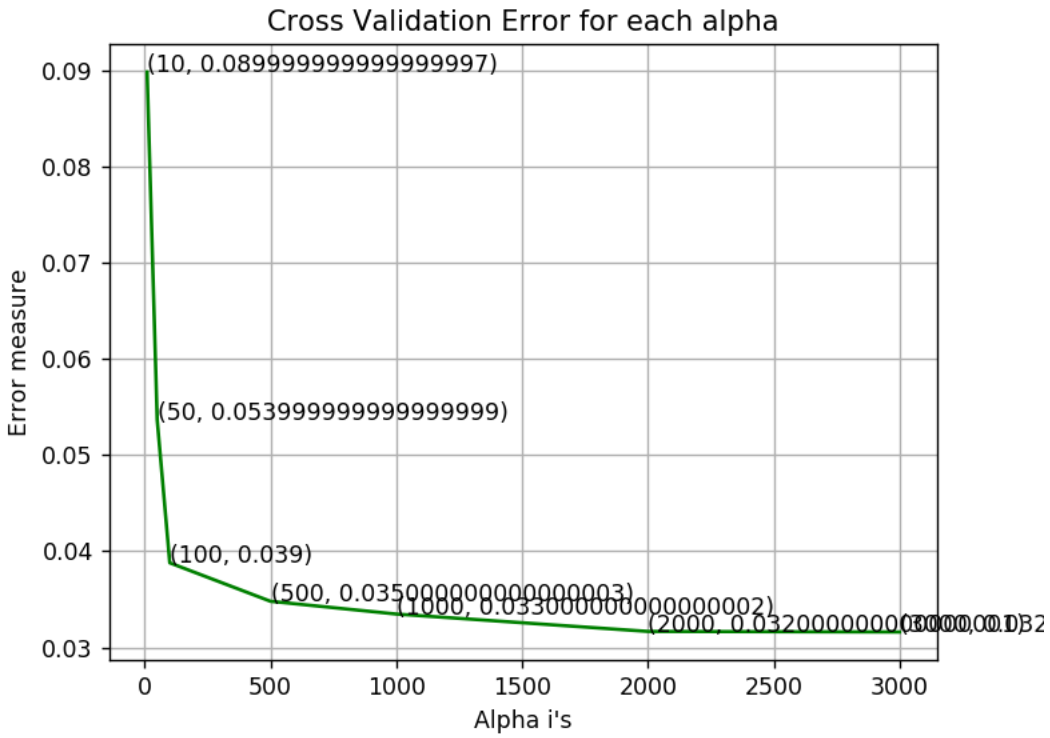
train loss 0.016
cv loss 0.035
test loss 0.040 Accuracy 98.92

XgBoost Classifier

Hyperparameter search

log_loss for c = 10 is 0.0898979446265
log_loss for c = 50 is 0.0536946658041
log_loss for c = 100 is 0.0387968186177
log_loss for c = 500 is 0.0347960327293
log_loss for c = 1000 is 0.0334668083237

log_loss for c = 2000 is 0.0316569078846
log_loss for c = 3000 is 0.0315972694477



Results from the Best model

train log loss is: 0.011
cross validation log loss is: 0.032
test log loss is: 0.032 Accuracy 99.35

XgBoost Classifier with best hyper parameters using Random search

```
Fitting 3 folds for each of 10 candidates, totalling 30 fits


[Parallel(n_jobs=-1)]: Done   2 tasks      | elapsed:  1.1min
[Parallel(n_jobs=-1)]: Done   9 tasks      | elapsed:  2.2min
[Parallel(n_jobs=-1)]: Done  19 out of  30 | elapsed:  4.5min remaining:  2.6min
[Parallel(n_jobs=-1)]: Done  23 out of  30 | elapsed:  5.8min remaining:  1.8min
[Parallel(n_jobs=-1)]: Done  27 out of  30 | elapsed:  6.7min remaining:   44.5s
[Parallel(n_jobs=-1)]: Done  30 out of  30 | elapsed:  7.4min finished


RandomizedSearchCV(cv=None, error_score='raise',
          estimator=XGBClassifier(base_score=0.5, colsample_bylevel=1, colsample_bytree=1,
       gamma=0, learning_rate=0.1, max_delta_step=0, max_depth=3,
       min_child_weight=1, missing=None, n_estimators=100, nthread=-1,
       objective='binary:logistic', reg_alpha=0, reg_lambda=1,
       scale_pos_weight=1, seed=0, silent=True, subsample=1),
          fit_params=None, iid=True, n_iter=10, n_jobs=-1,
          param_distributions={'learning_rate': [0.01, 0.03, 0.05, 0.1, 0.15, 0.2], 'n_estimators': [100, 200, 500, 1000, 2000], 'max_dep
th': [3, 5, 10], 'colsample_bytree': [0.1, 0.3, 0.5, 1], 'subsample': [0.1, 0.3, 0.5, 1]},
          pre_dispatch='2*n_jobs', random_state=None, refit=True,
          return_train_score=True, scoring=None, verbose=10)
```

Best Parameters

{'subsample': 1, 'n_estimators': 1000, 'max_depth': 10, 'learning_rate': 0.15, 'colsample_bytree': 0.3}

Results from the Best Parameter Model

train loss 0.012
cv loss 0.035
test loss 0.032 Accuracy 99.37