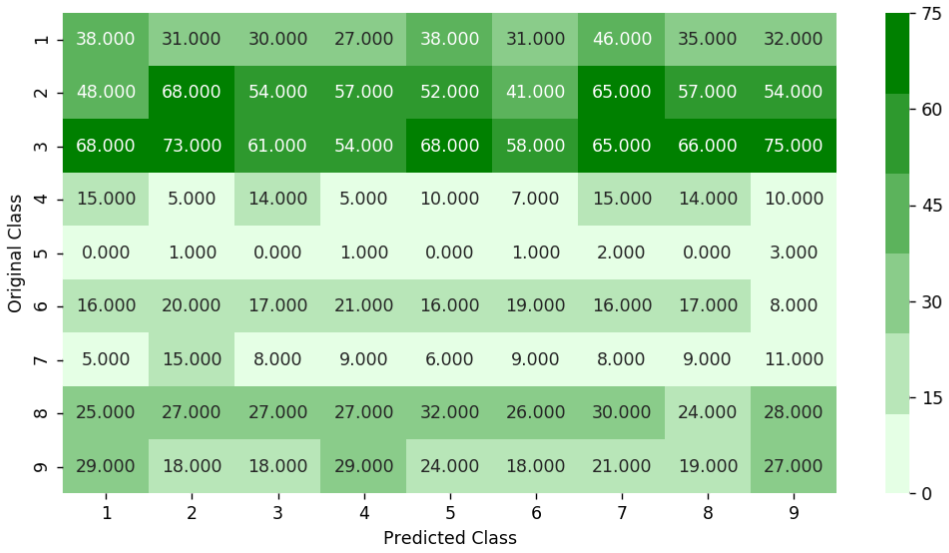
Machine Learning Model

Created	@Jun 10, 2021 12:31 PM
Tags	

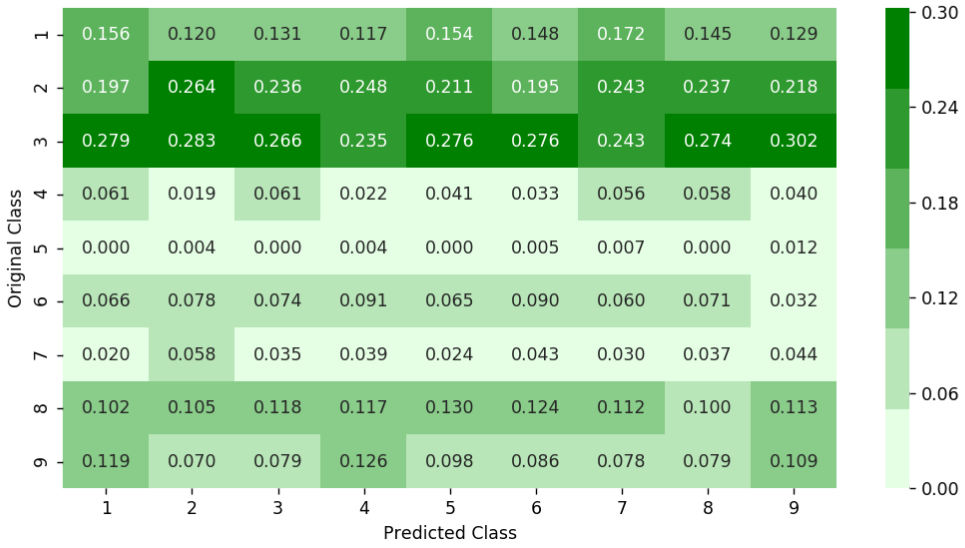
Random Model

Log loss on Cross Validation Data using Random Model 2.46
Log loss on Test Data using Random Model 2.48
Accuracy 11.49

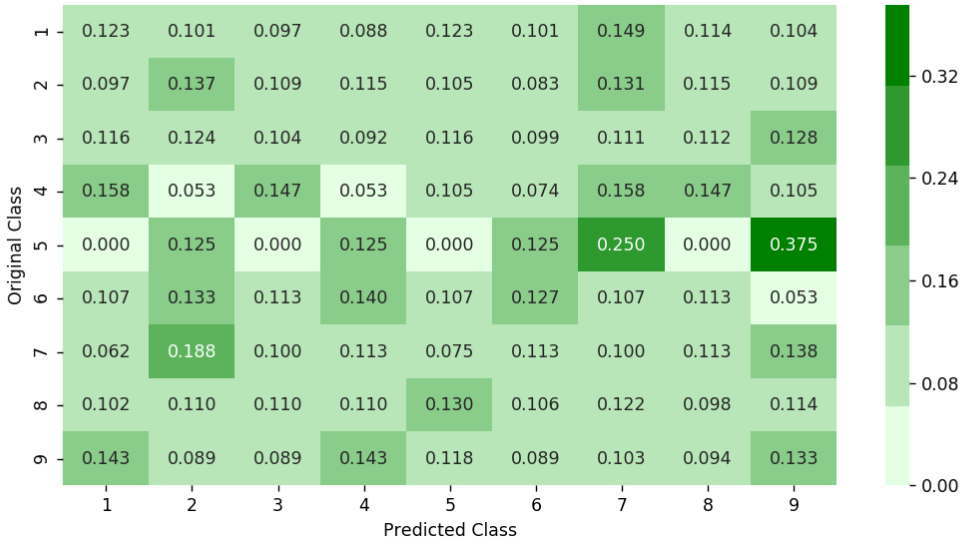
Confusion Matrix



Precision Matrix



Recall Matrix

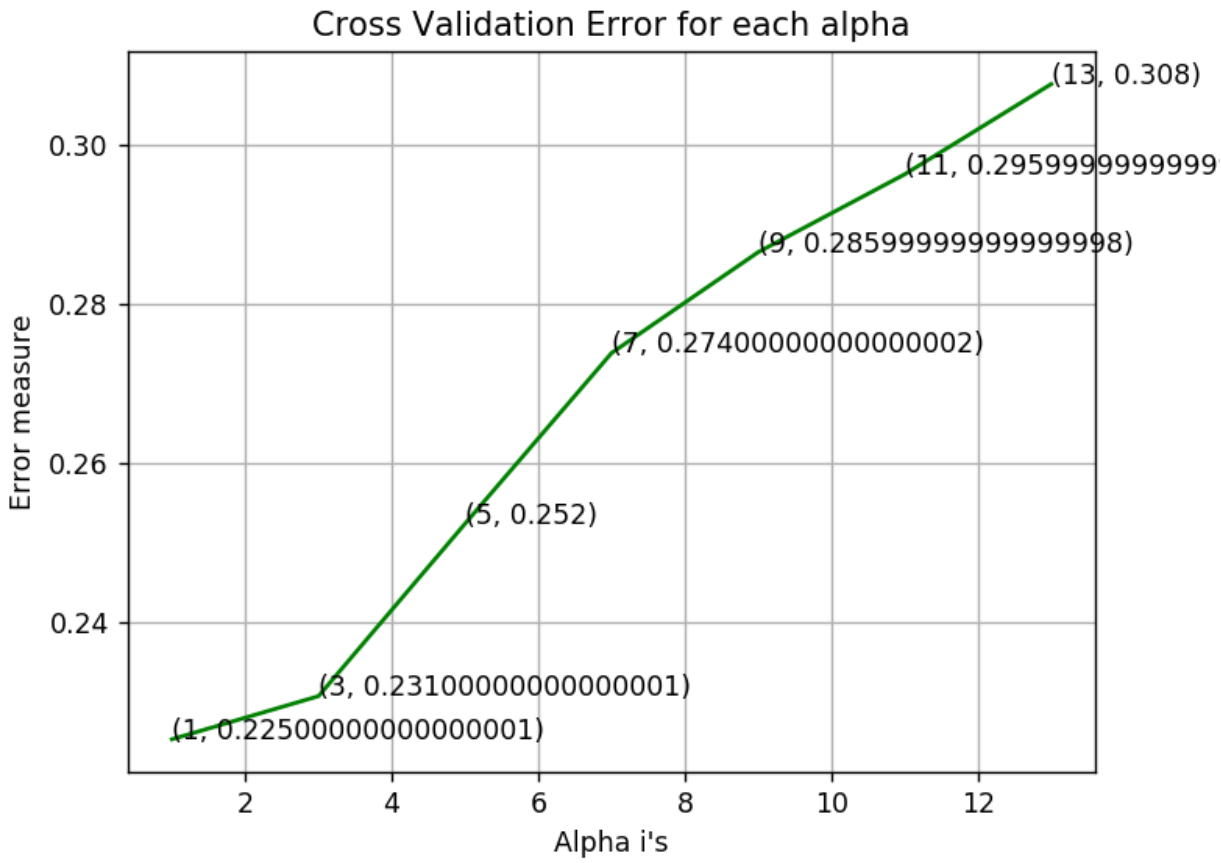


Bytes file

K Nearest Neighbor Classification

Hyperparameter Search

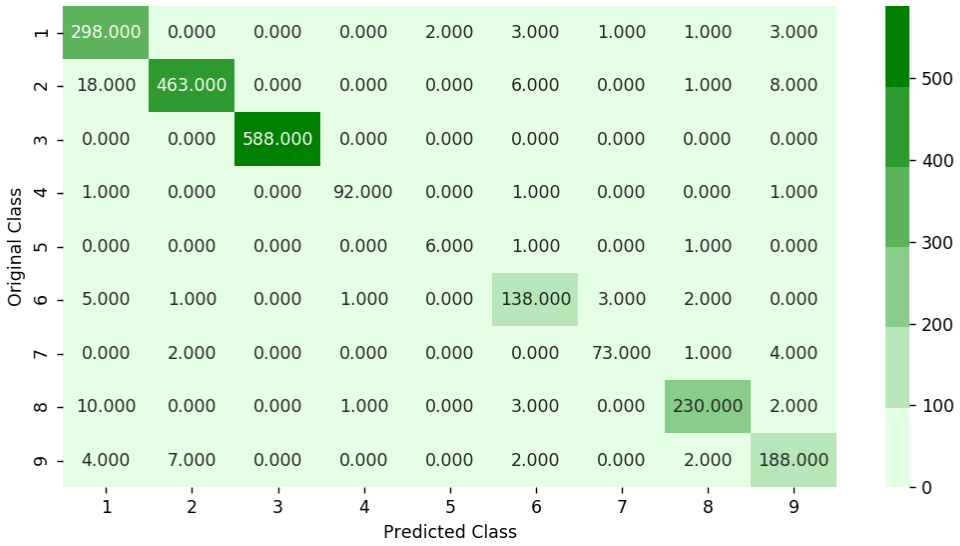
log_loss for k = 1 is 0.225386237304
log_loss for k = 3 is 0.230795229168
log_loss for k = 5 is 0.252421408646
log_loss for k = 7 is 0.273827486888
log_loss for k = 9 is 0.286469181555
log_loss for k = 11 is 0.29623391147
log_loss for k = 13 is 0.307551203154



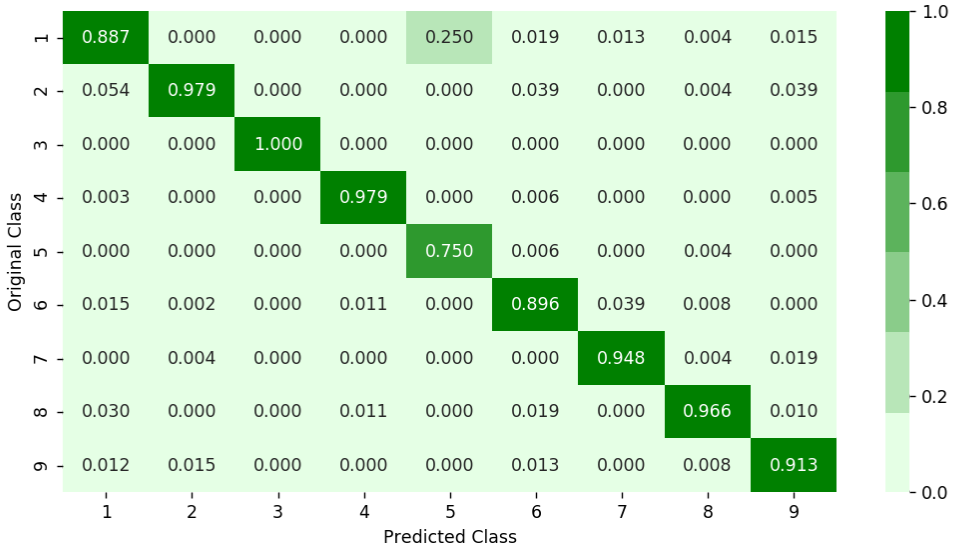
Results from the Best Model

For values of best alpha = 1 The train log loss is: 0.08
For values of best alpha = 1 The cross validation log loss is: 0.23
For values of best alpha = 1 The test log loss is: 0.24
Accuracy 95.49

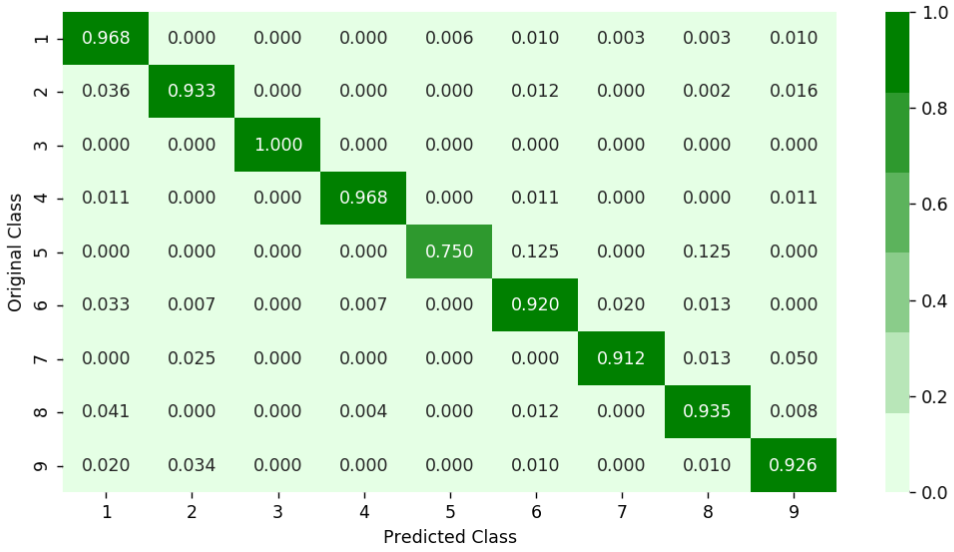
Confusion Matrix



Precision Matrix



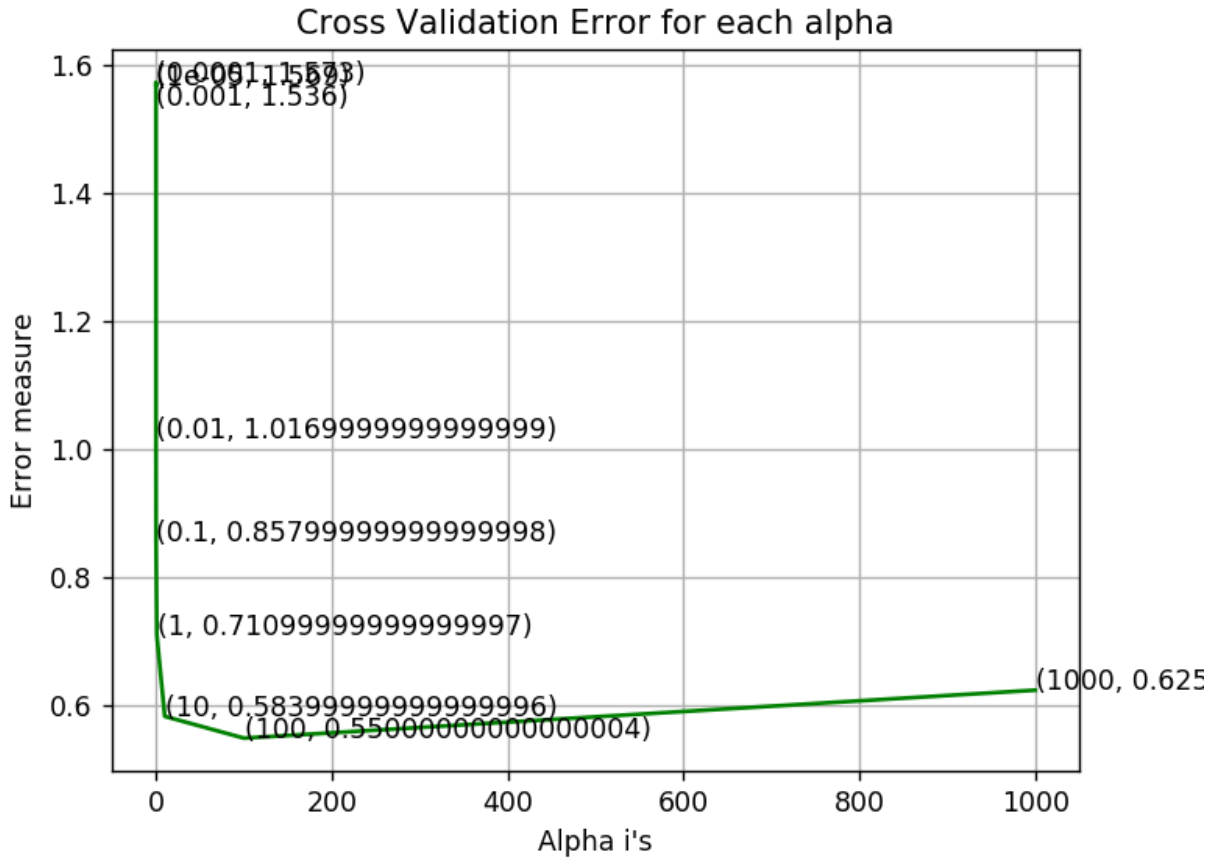
Recall Matrix



Logistic Regression

Hyperparameter Search

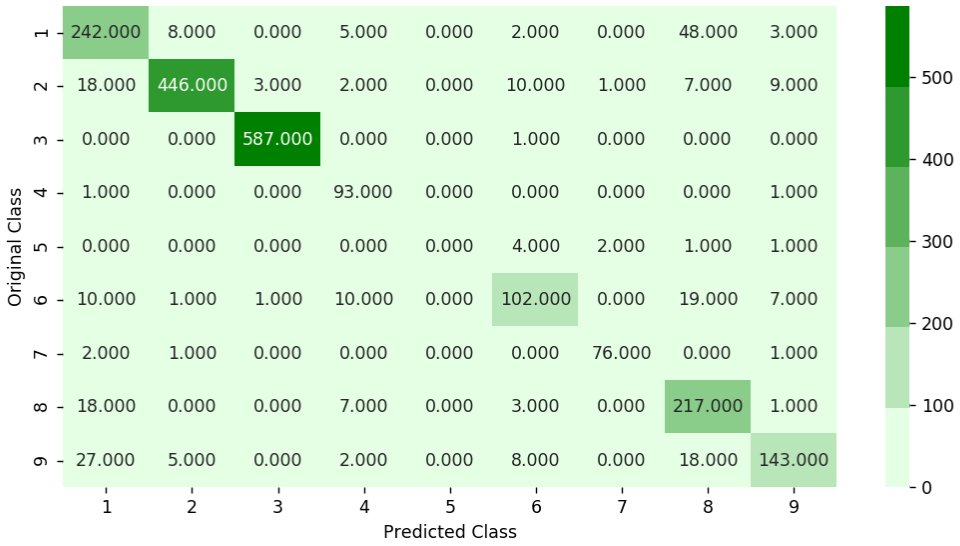
log_loss for c = 1e-05 is 1.56916911178
log_loss for c = 0.0001 is 1.57336384417
log_loss for c = 0.001 is 1.53598598273
log_loss for c = 0.01 is 1.01720972418
log_loss for c = 0.1 is 0.857766083873
log_loss for c = 1 is 0.711154393309
log_loss for c = 10 is 0.583929522635
log_loss for c = 100 is 0.549929846589
log_loss for c = 1000 is 0.624746769121



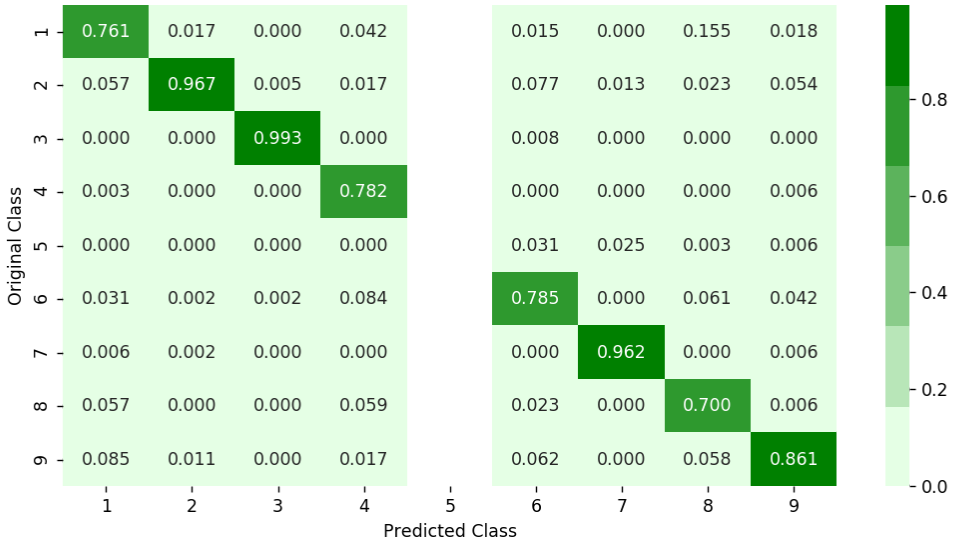
Results from the Best Model

log loss for train data 0.50
log loss for cv data 0.55
log loss for test data 0.53
Number of misclassified points 87.67

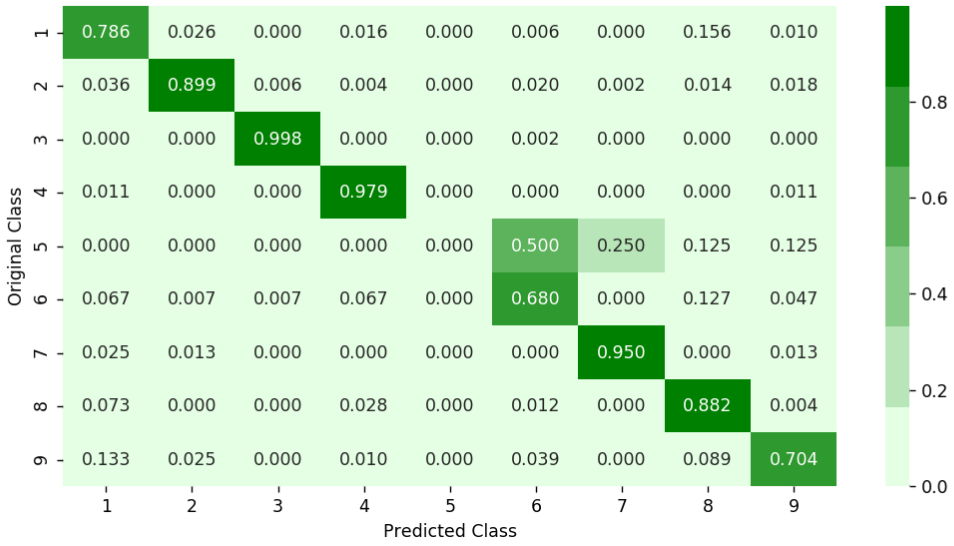
Confusion Matrix



Precision Matrix



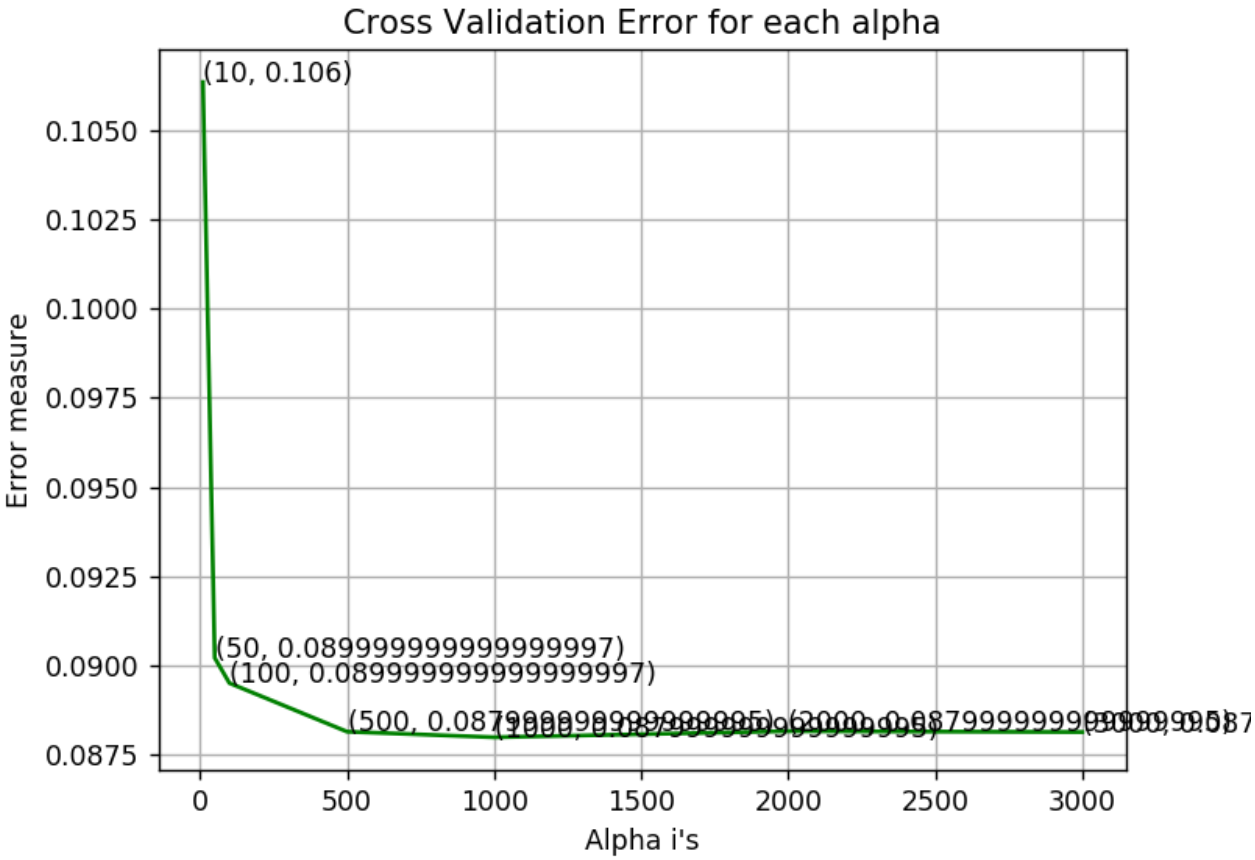
Recall Matrix



Random Forest Classifier

Hyperparameter Search

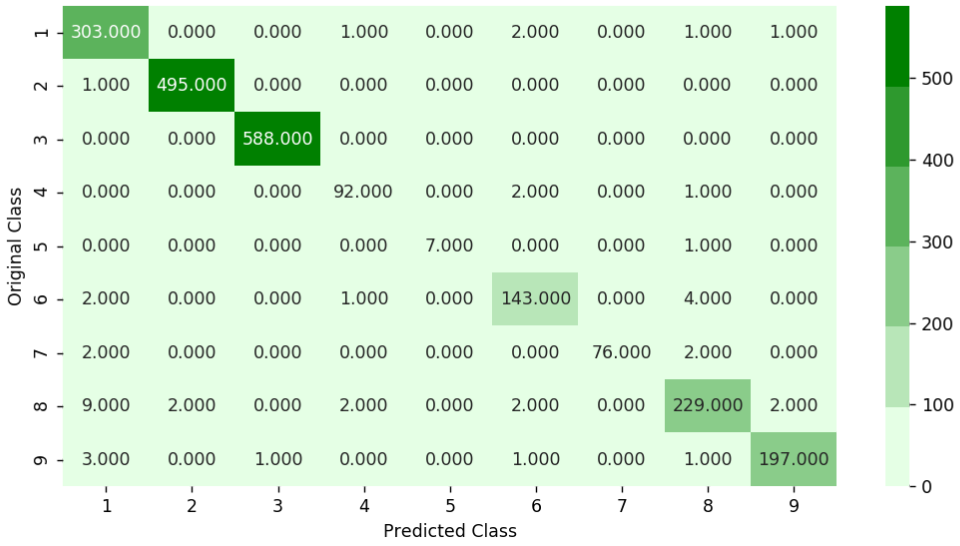
log_loss for c = 10 is 0.106357709164
log_loss for c = 50 is 0.0902124124145
log_loss for c = 100 is 0.0895043339776
log_loss for c = 500 is 0.0881420869288
log_loss for c = 1000 is 0.0879849524621
log_loss for c = 2000 is 0.0881566647295
log_loss for c = 3000 is 0.0881318948443



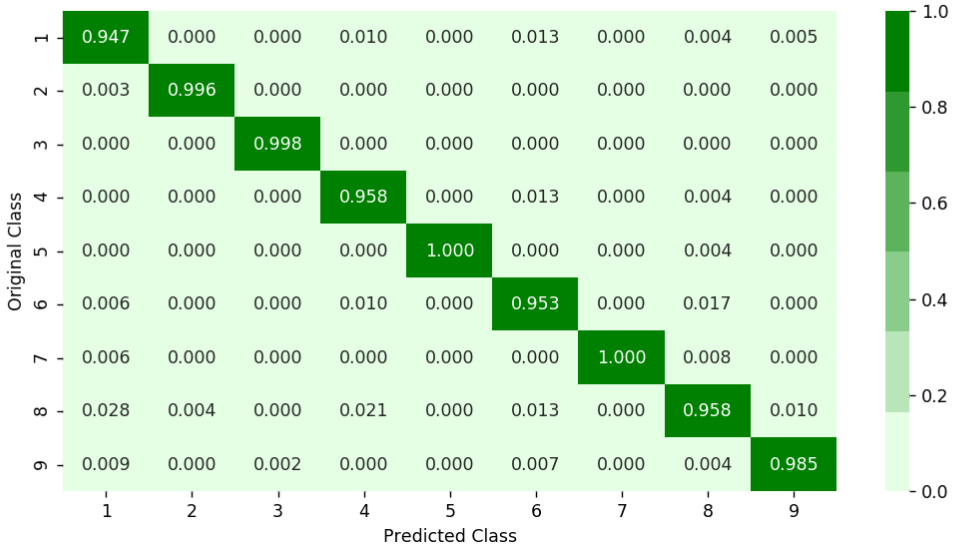
Results from the Best model

For values of best alpha = 1000 The train log loss is: 0.031
For values of best alpha = 1000 The cross validation log loss is: 0.09
For values of best alpha = 1000 The test log loss is: 0.08
Accuracy 96.76

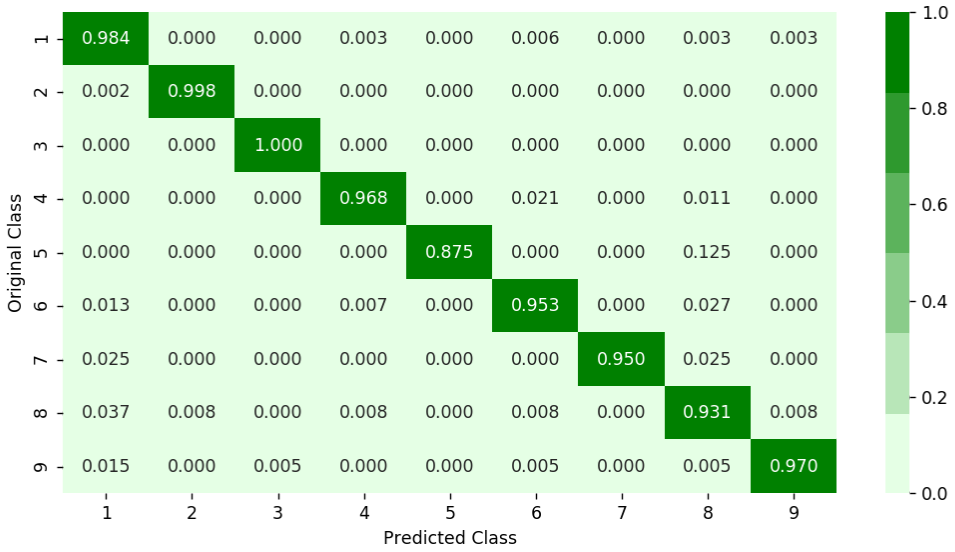
Confusion Matrix



Precision Matrix



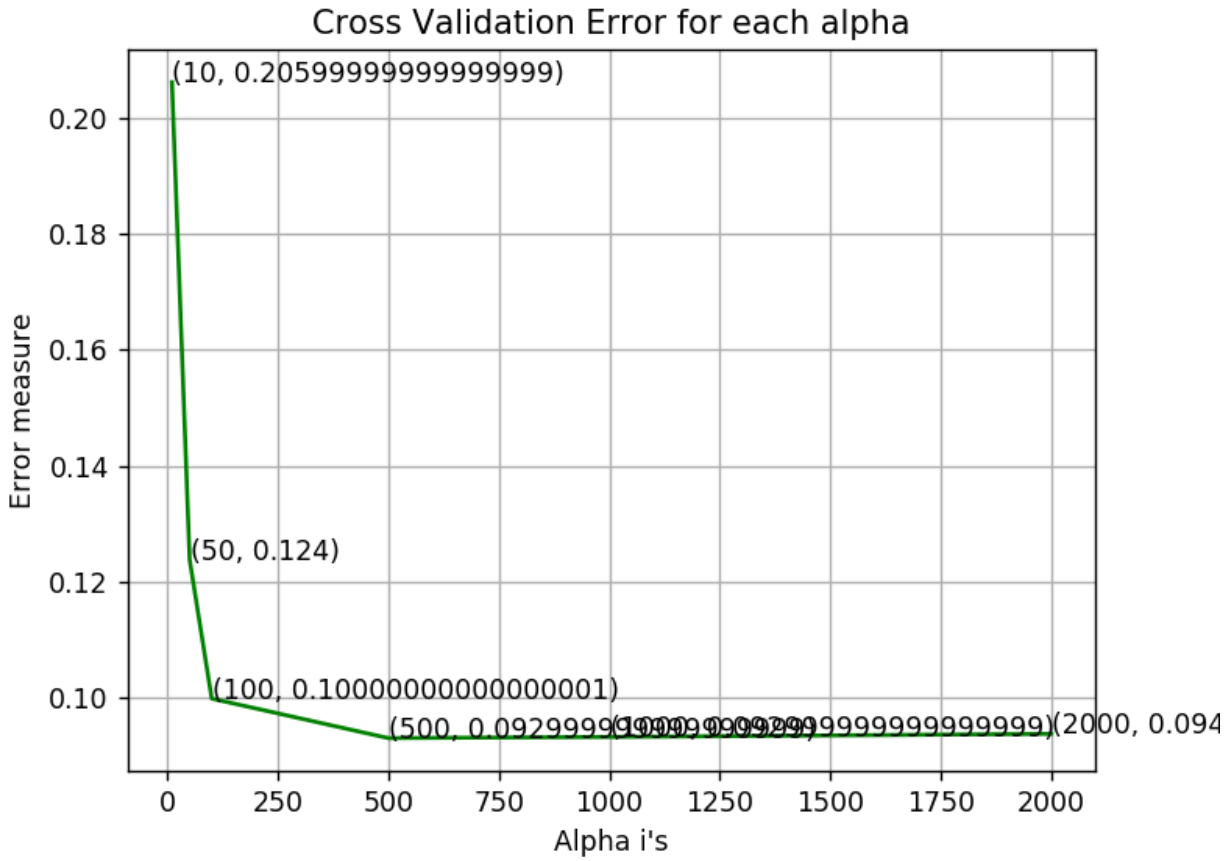
Recall Matrix



XgBoost Classification

Hyperparameter Search

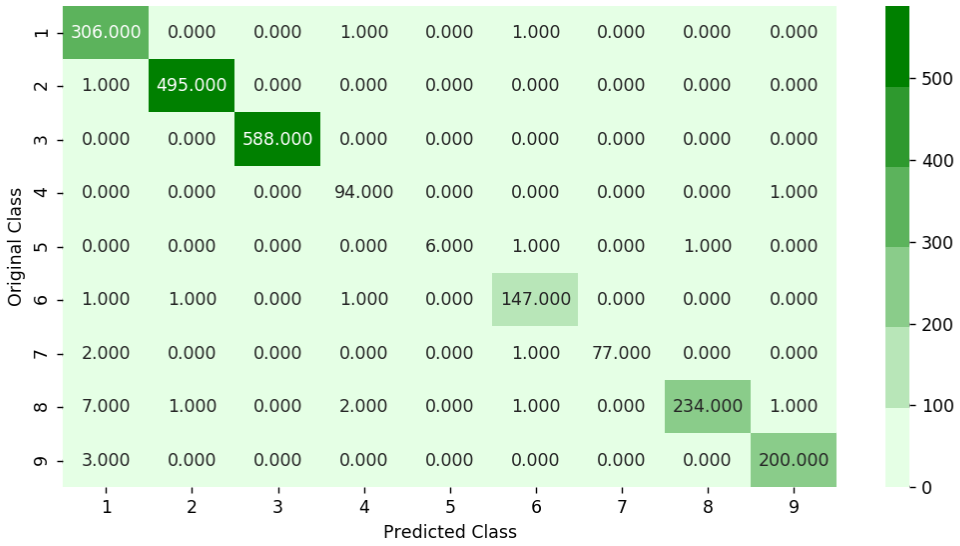
log_loss for c = 10 is 0.20615980494
log_loss for c = 50 is 0.123888382365
log_loss for c = 100 is 0.099919437112
log_loss for c = 500 is 0.0931035681289
log_loss for c = 1000 is 0.0933084876012
log_loss for c = 2000 is 0.0938395690309



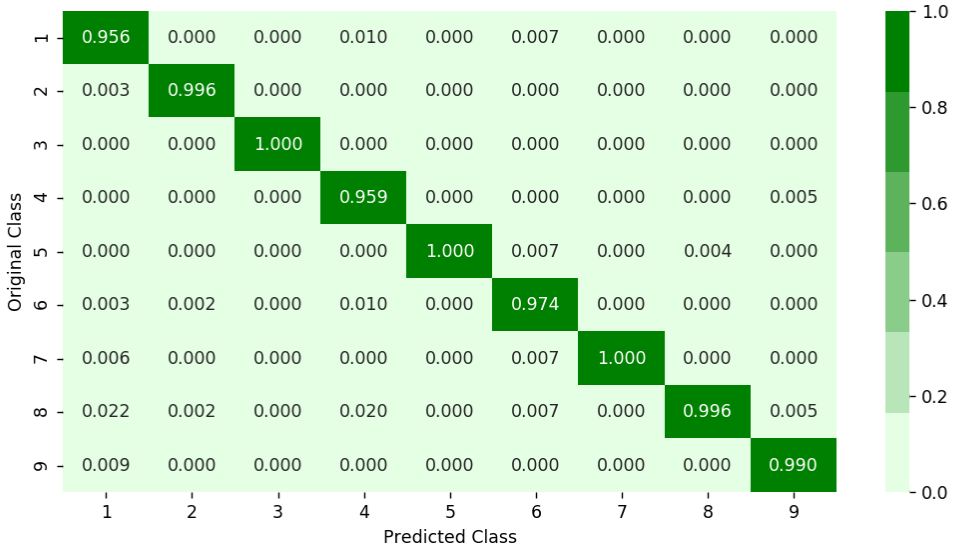
Results from the Best Model

For values of best alpha = 500 The train log loss is: 0.022
For values of best alpha = 500 The cross validation log loss is: 0.09
For values of best alpha = 500 The test log loss is: 0.08
Accuracy 98.67

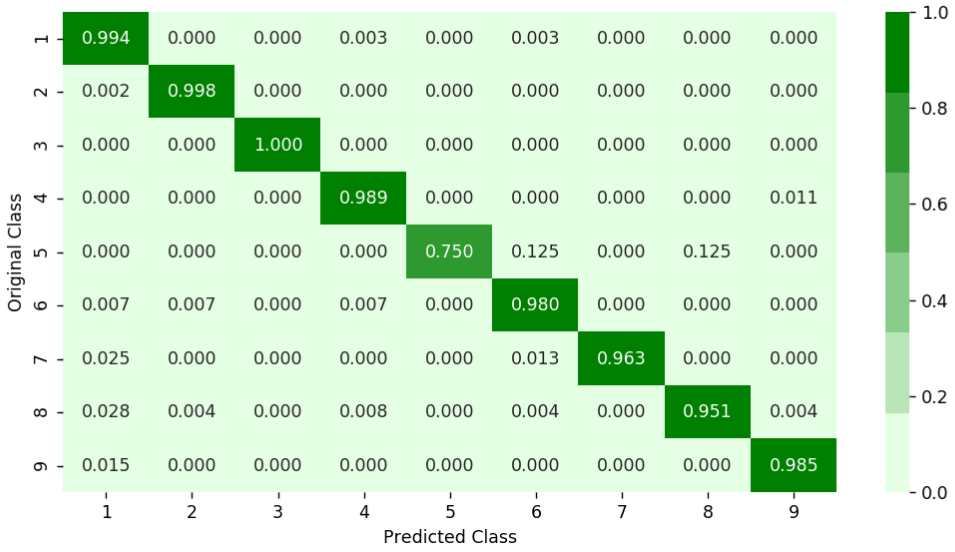
Confusion Matrix



Precision Matrix



Recall Matrix



XgBoost Classification 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:   26.5s
[Parallel(n_jobs=-1)]: Done   9 tasks      | elapsed:  5.8min
[Parallel(n_jobs=-1)]: Done  19 out of  30 | elapsed:  9.3min remaining:  5.4min
[Parallel(n_jobs=-1)]: Done  23 out of  30 | elapsed: 10.1min remaining:  3.1min
[Parallel(n_jobs=-1)]: Done  27 out of  30 | elapsed: 14.0min remaining:  1.6min
[Parallel(n_jobs=-1)]: Done  30 out of  30 | elapsed: 14.2min 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_est
imators': [100, 200, 500, 1000, 2000], 'max_depth': [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': 500, 'max_depth': 5, 'learning_rate': 0.05, 'colsample_bytree': 0.5}

Results from the Best Parameter Model

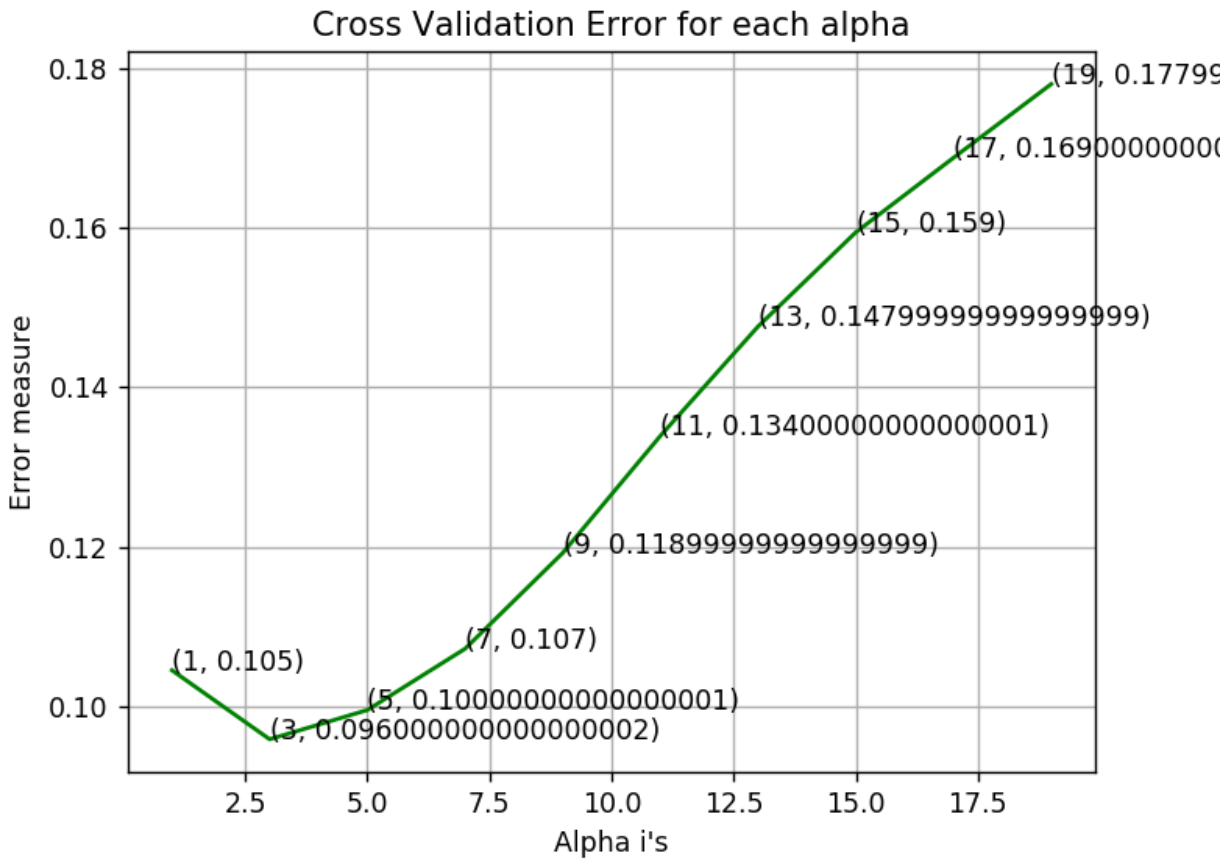
train loss 0.022
cv loss 0.09
test loss 0.08
Accuracy 98.67

ASM file

K-Nearest Neighbors

Hyperparameter search

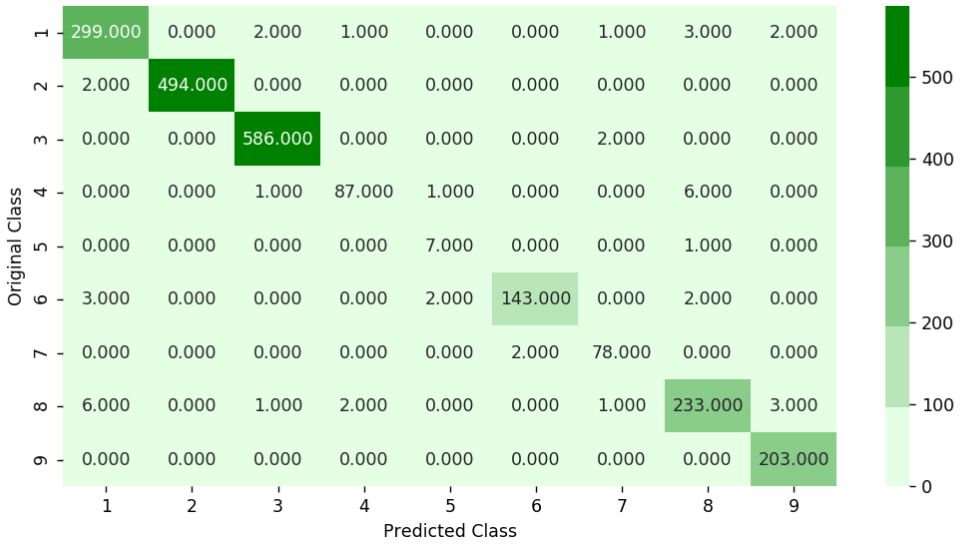
log_loss for k = 1 is 0.104531321344
log_loss for k = 3 is 0.0958800580948
log_loss for k = 5 is 0.0995466557335
log_loss for k = 7 is 0.107227274345
log_loss for k = 9 is 0.119239543547
log_loss for k = 11 is 0.133926642781
log_loss for k = 13 is 0.147643793967
log_loss for k = 15 is 0.159439699615
log_loss for k = 17 is 0.16878376444
log_loss for k = 19 is 0.178020728839



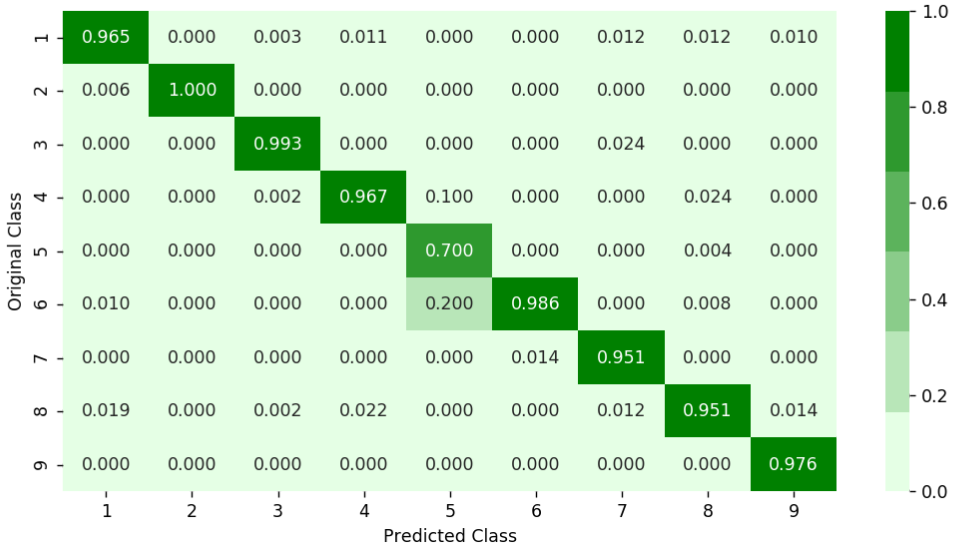
Results from the Best Model

log loss for train data 0.048
log loss for cv data 0.096
log loss for test data 0.090
Accuracy 97.98

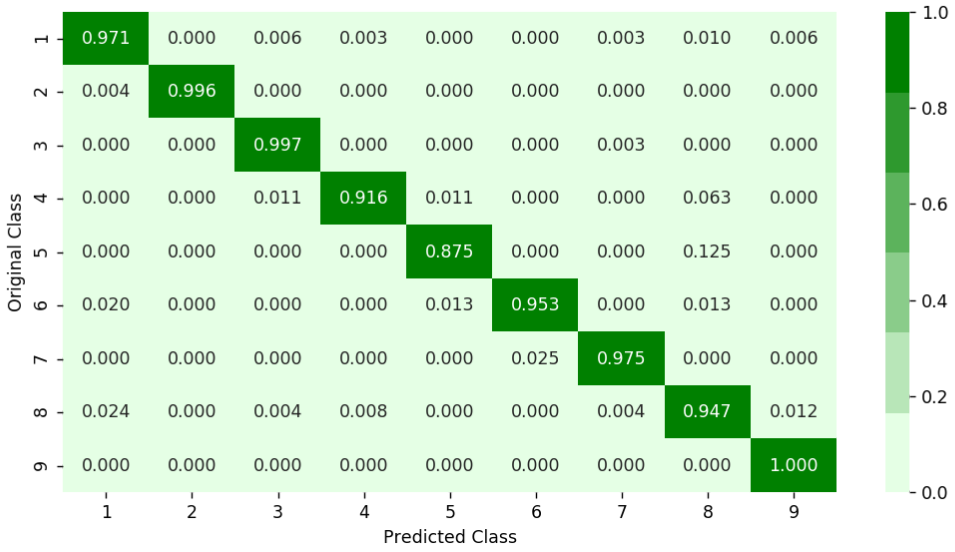
Confusion Matrix



Precision Matrix



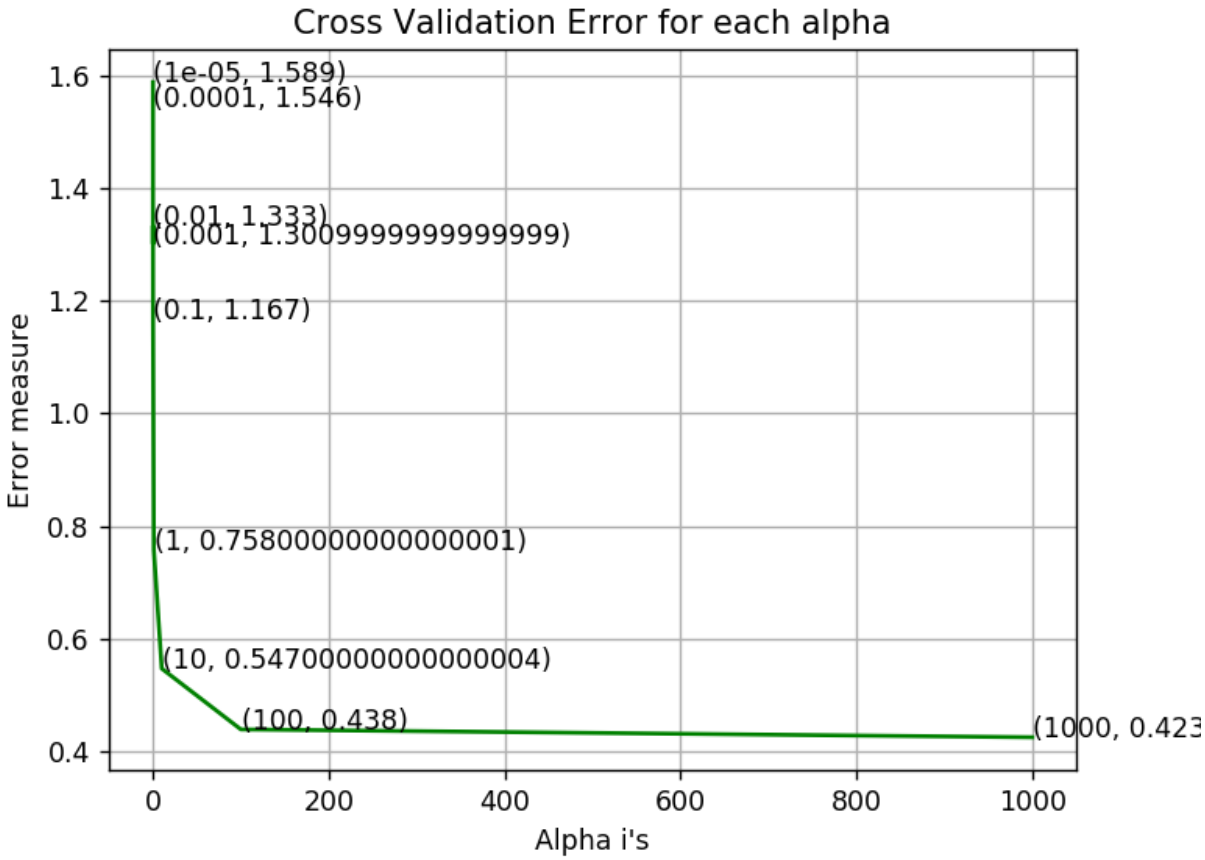
Recall Matrix



Logistic Regression

Hyperparameter search

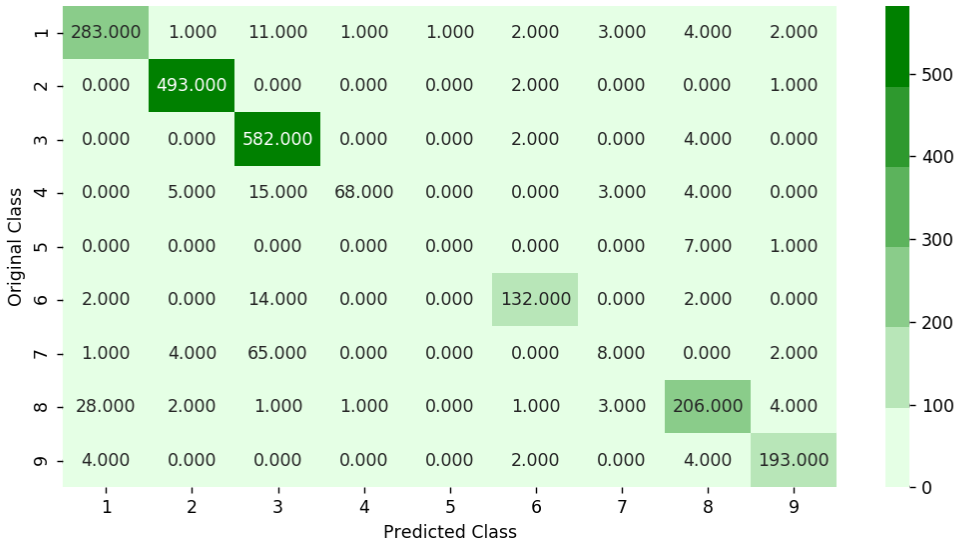
log_loss for c = 1e-05 is 1.58867274165
log_loss for c = 0.0001 is 1.54560797884
log_loss for c = 0.001 is 1.30137786807
log_loss for c = 0.01 is 1.33317456931
log_loss for c = 0.1 is 1.16705751378
log_loss for c = 1 is 0.757667807779
log_loss for c = 10 is 0.546533939819
log_loss for c = 100 is 0.438414998062
log_loss for c = 1000 is 0.424423536526



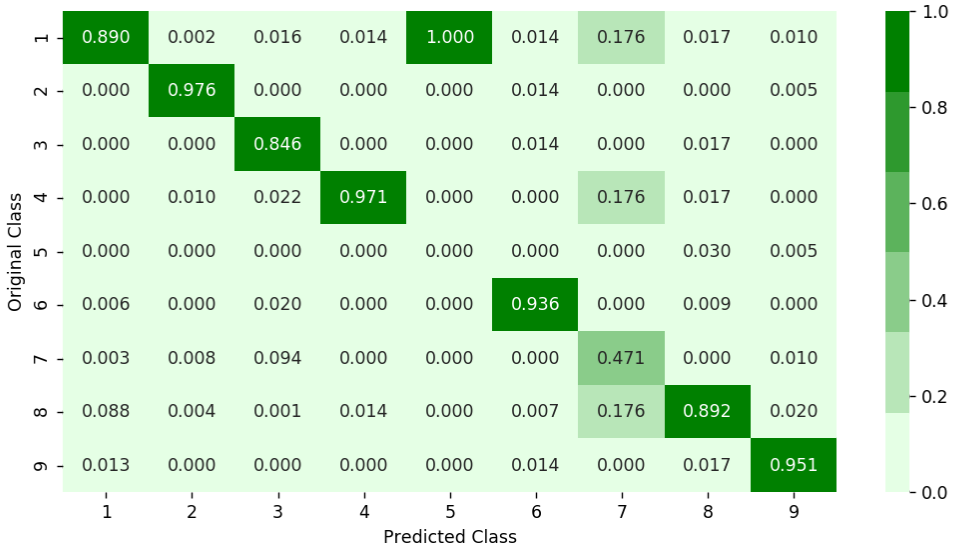
Results from the Best Model

log loss for train data 0.40
log loss for cv data 0.42
log loss for test data 0.42
Number of misclassified points 90.39

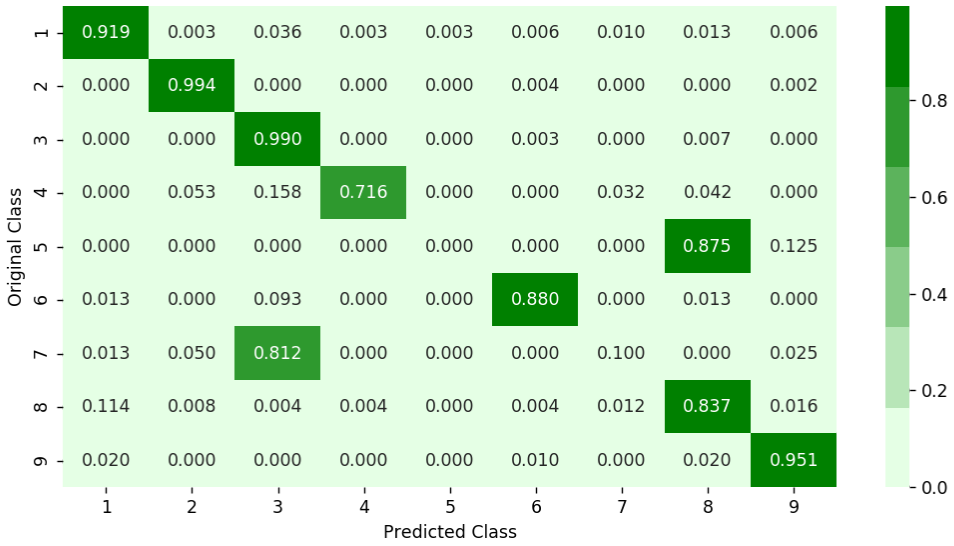
Confusion Matrix



Precision Matrix



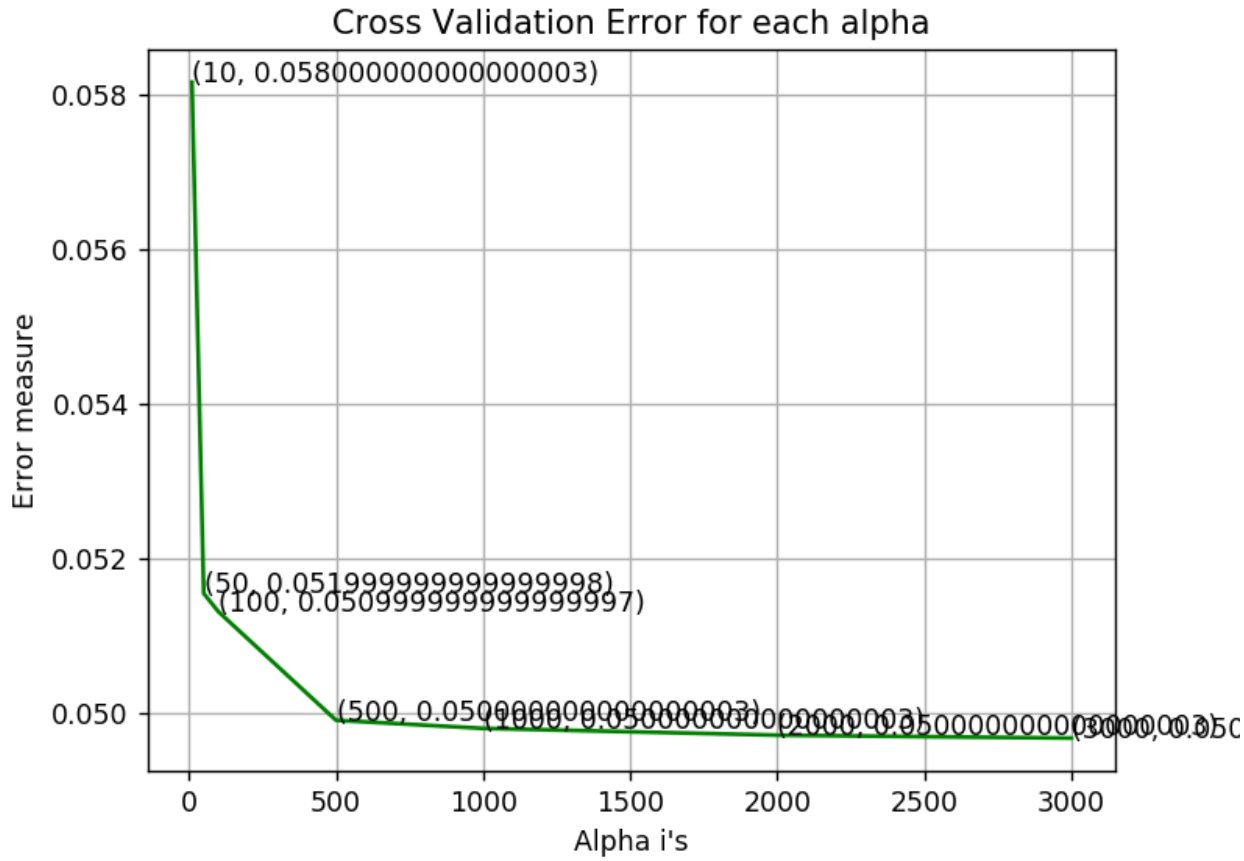
Recall Matrix



Random Forest

Hyperparameter search

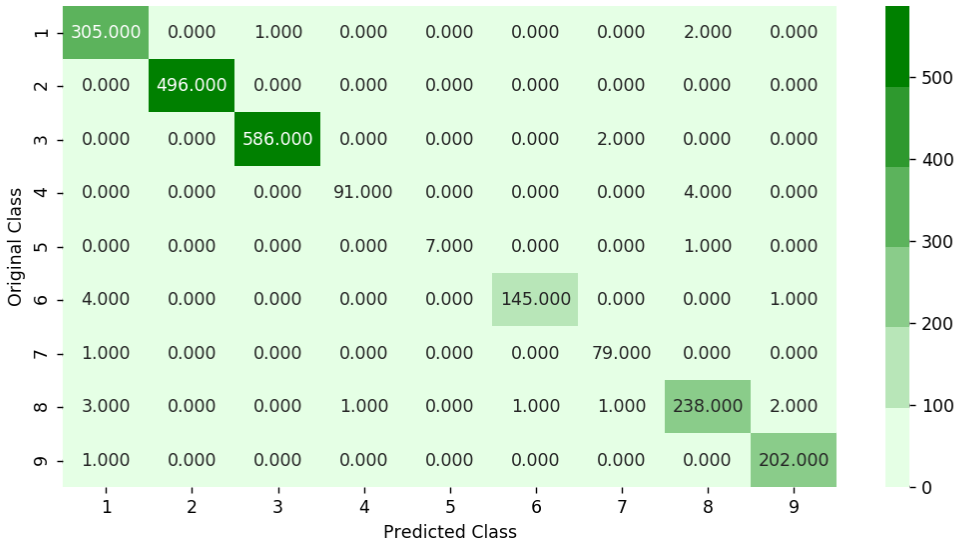
log_loss for c = 10 is 0.0581657906023
log_loss for c = 50 is 0.0515443148419
log_loss for c = 100 is 0.0513084973231
log_loss for c = 500 is 0.0499021761479
log_loss for c = 1000 is 0.0497972474298
log_loss for c = 2000 is 0.0497091690815
log_loss for c = 3000 is 0.0496706817633



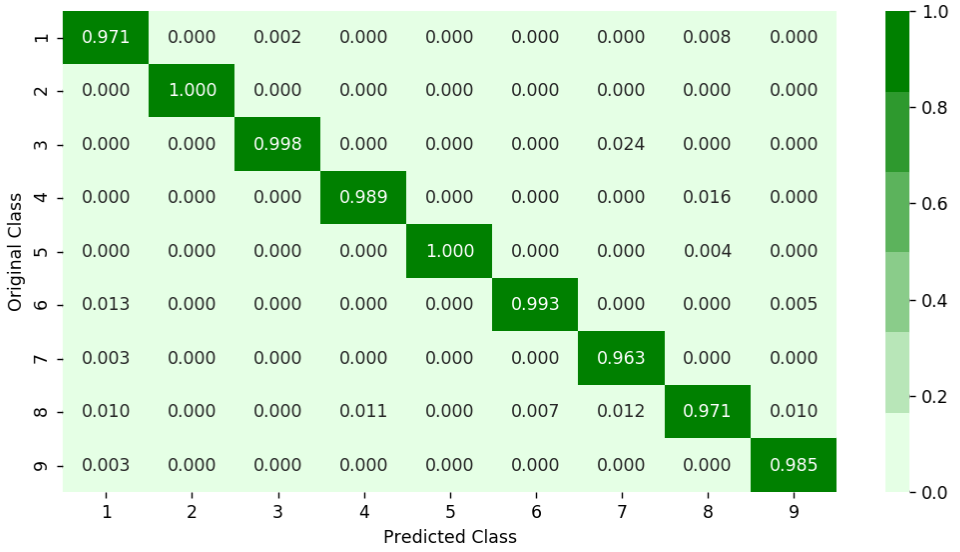
Results from the Best Model

log loss for train data 0.012
log loss for cv data 0.050
log loss for test data 0.057
Accuracy 98.85

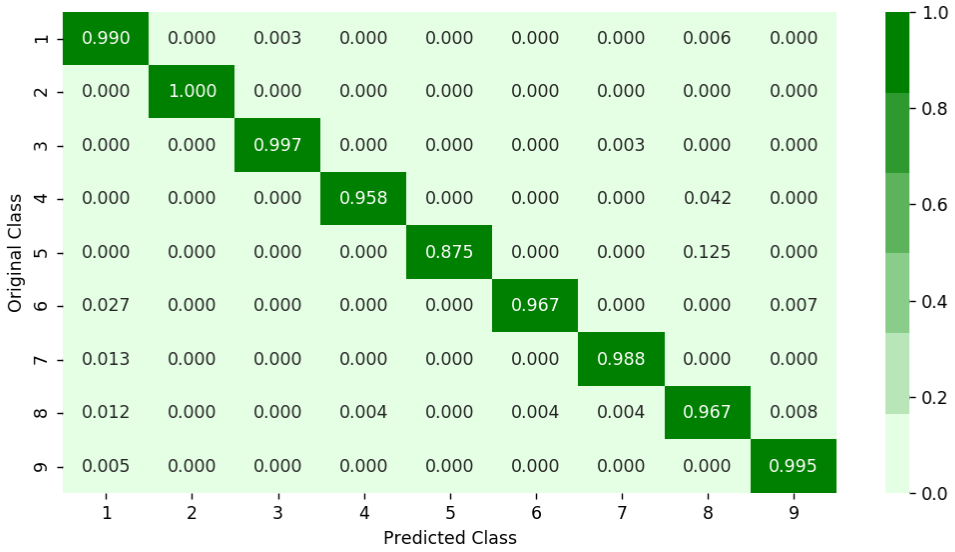
Confusion Matrix



Precision Matrix



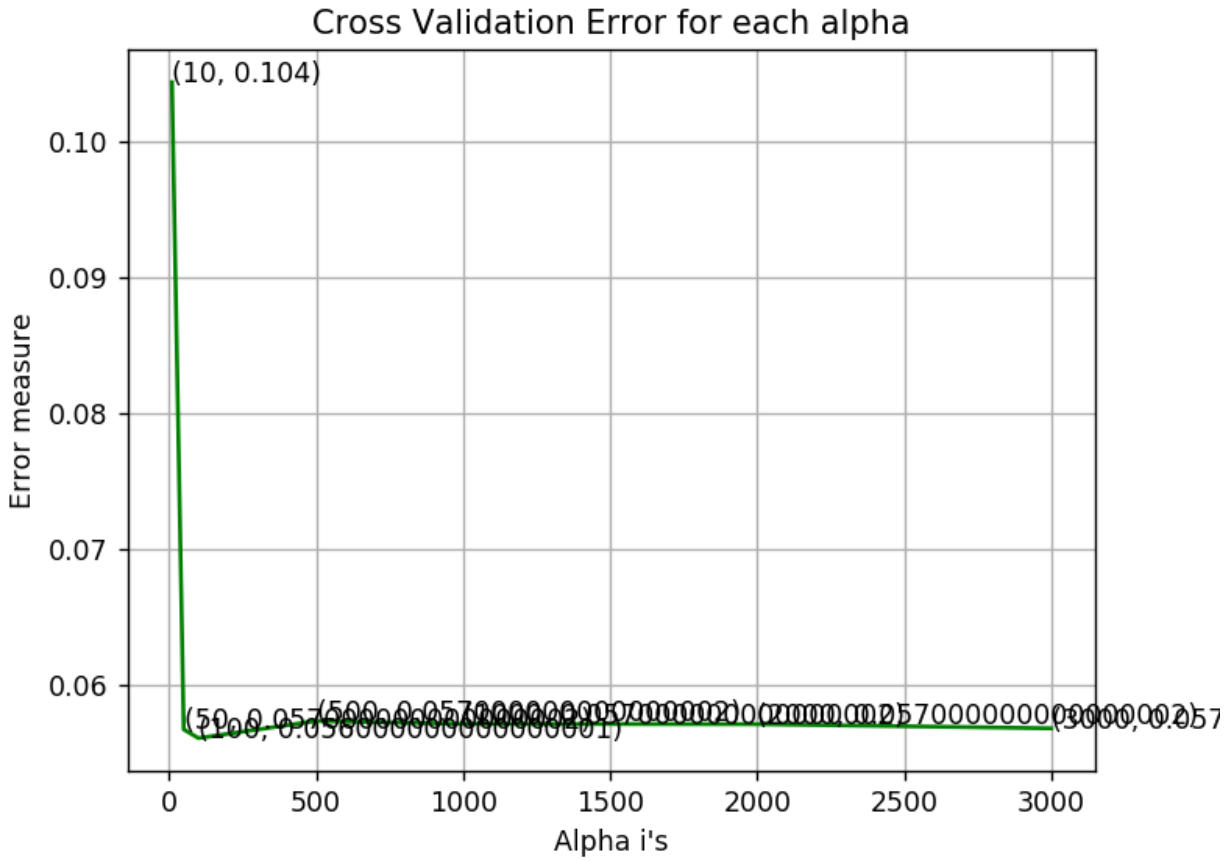
Recall Matrix



XgBoost Classifier

Hyperparameter search

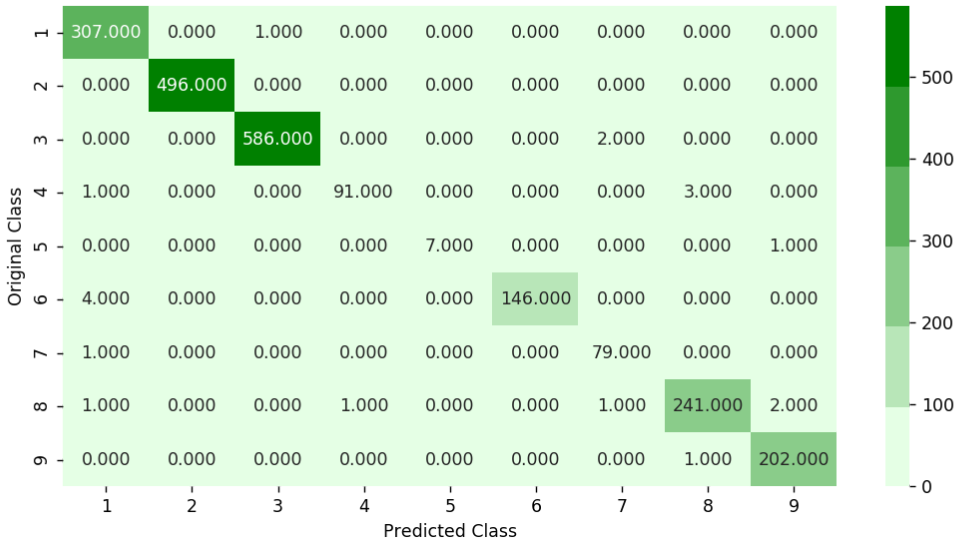
log_loss for c = 10 is 0.104344888454
log_loss for c = 50 is 0.0567190635611
log_loss for c = 100 is 0.056075038646
log_loss for c = 500 is 0.057336051683
log_loss for c = 1000 is 0.0571265109903
log_loss for c = 2000 is 0.057103406781
log_loss for c = 3000 is 0.0567993215778



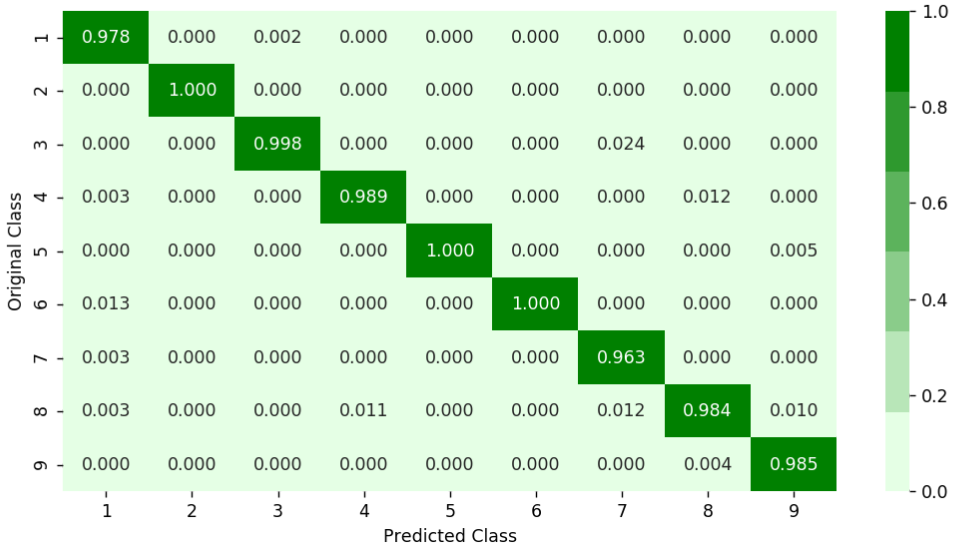
Results from the Best Model

For values of best alpha = 100 The train log loss is: 0.012
For values of best alpha = 100 The cross validation log loss is: 0.056
For values of best alpha = 100 The test log loss is: 0.049
Accuracy 99.13

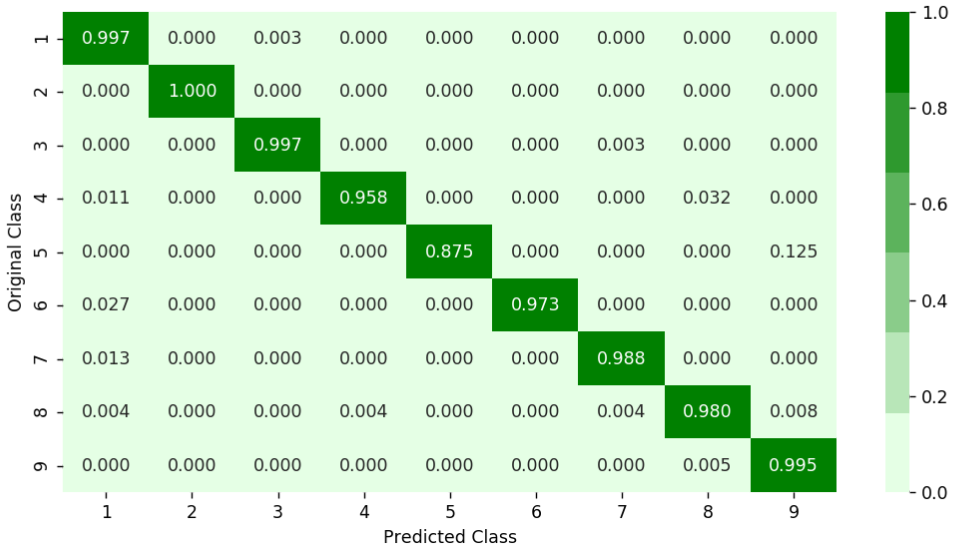
Confusion Matrix



Precision Matrix



Recall Matrix



XgBoost Classifier with best hyperparameters

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


[Parallel(n_jobs=-1)]: Done   2 tasks      | elapsed:    8.1s
[Parallel(n_jobs=-1)]: Done   9 tasks      | elapsed:   32.8s
[Parallel(n_jobs=-1)]: Done  19 out of  30 | elapsed:  1.1min remaining:   39.3s
[Parallel(n_jobs=-1)]: Done  23 out of  30 | elapsed:  1.3min remaining:   23.0s
[Parallel(n_jobs=-1)]: Done  27 out of  30 | elapsed:  1.4min remaining:    9.2s
[Parallel(n_jobs=-1)]: Done  30 out of  30 | elapsed:  2.3min 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_est
imators': [100, 200, 500, 1000, 2000], 'max_depth': [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': 200, 'max_depth': 5, 'learning_rate': 0.15, 'colsample_bytree': 0.5}

Results from the Best Parameter Model

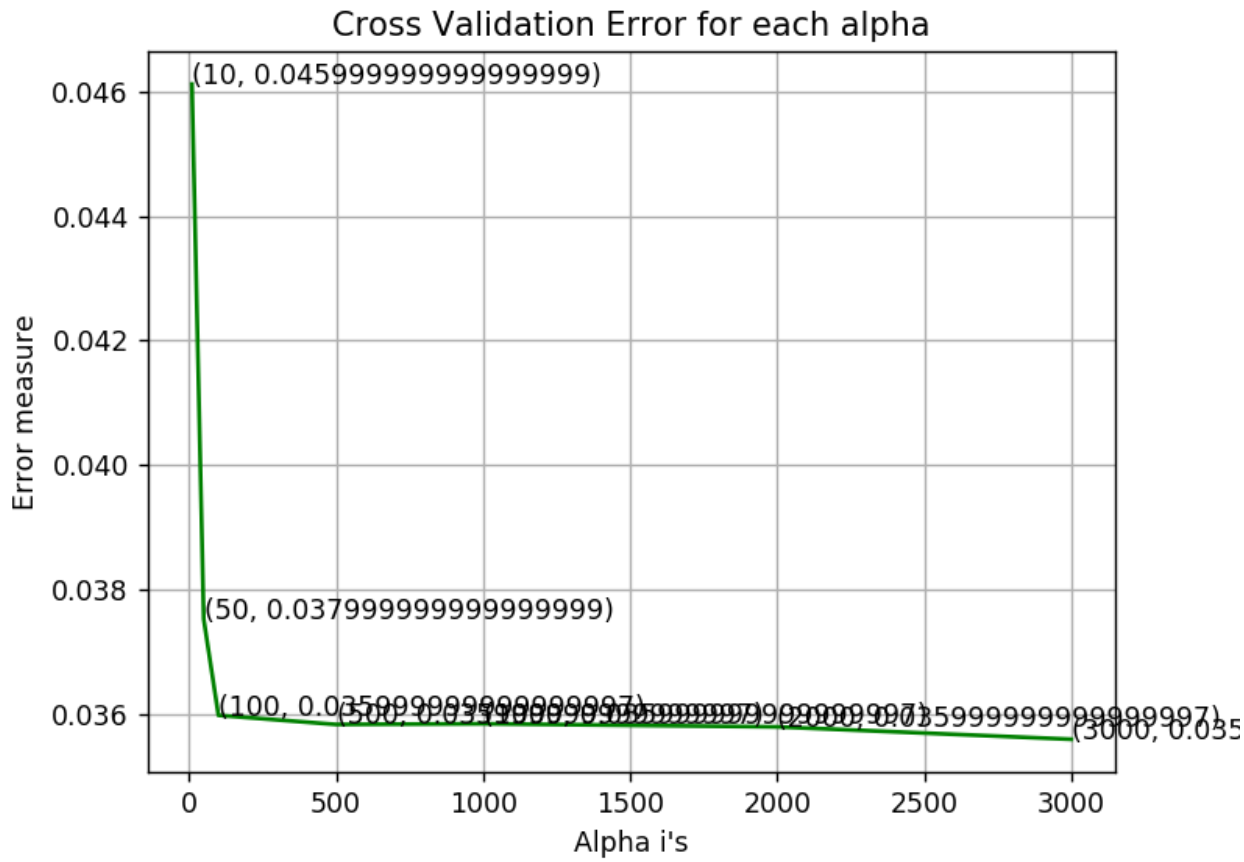
train loss 0.010
cv loss 0.050
test loss 0.048
Accuracy 99.16

Merged features

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

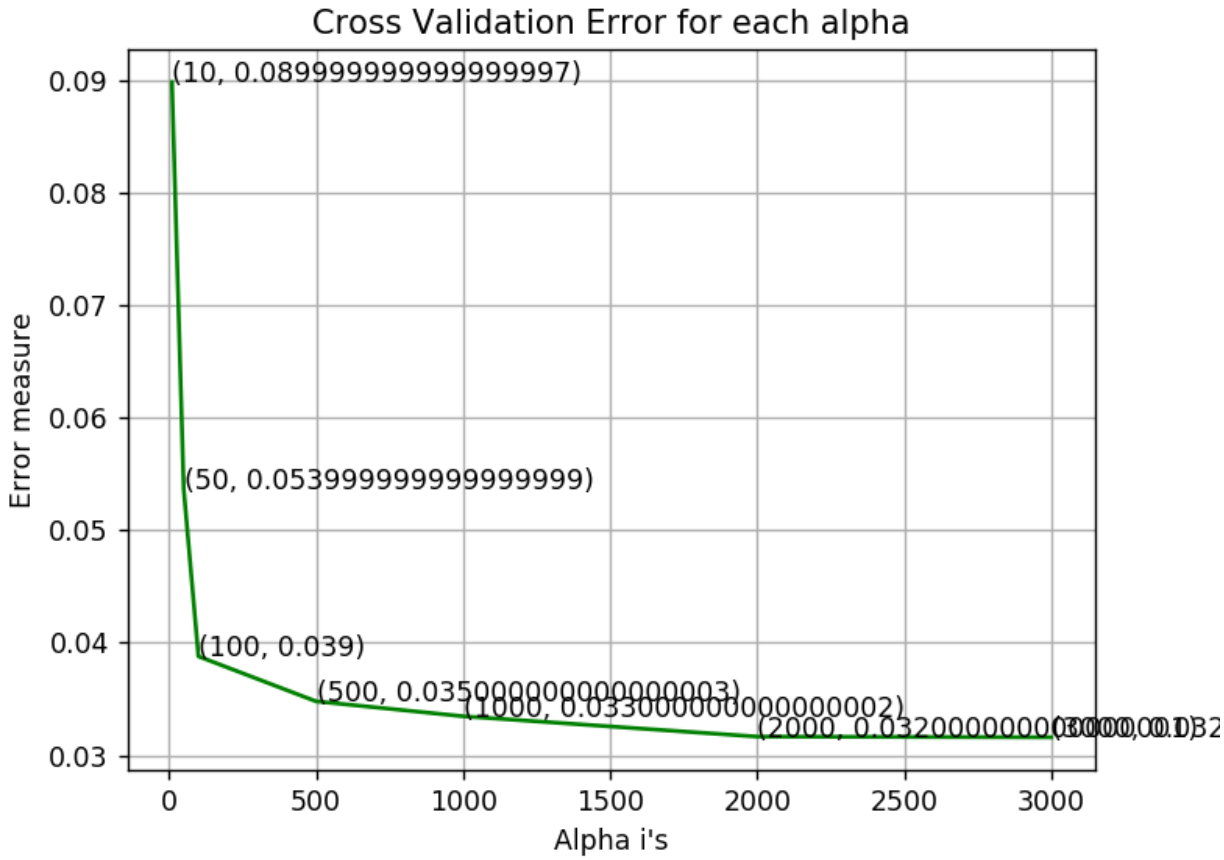
train loss 0.016
cv loss 0.035
test loss 0.040
Accuracy 98.92

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_est
imators': [100, 200, 500, 1000, 2000], 'max_depth': [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