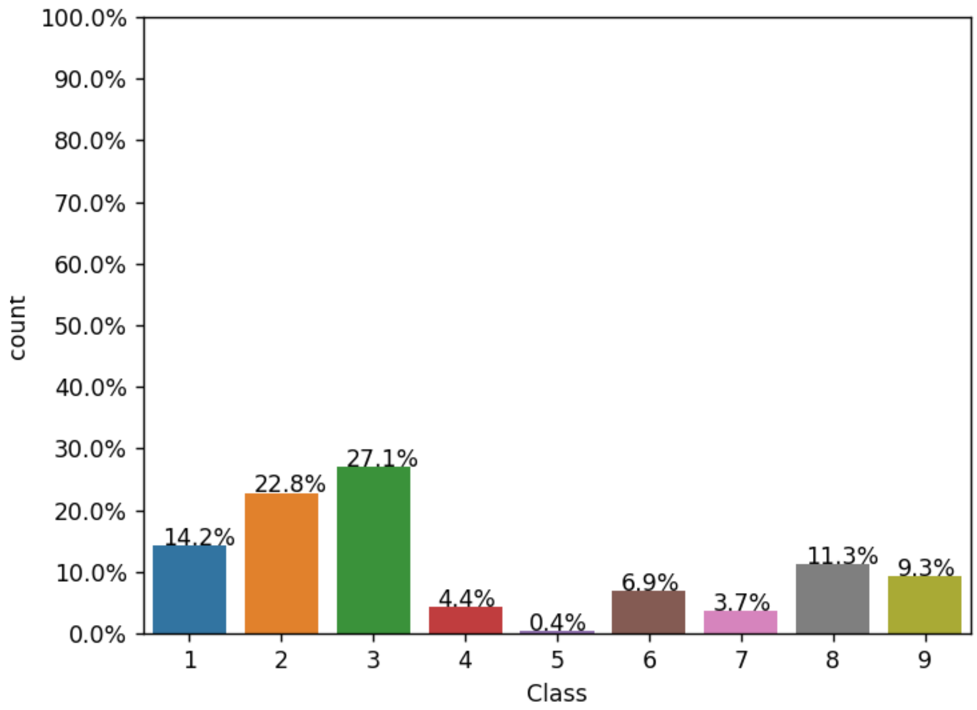
Bytes file

Created	@Jun 10, 2021 11:17 AM
Tags	

Exploratory Data Analysis

Number of data points in each class

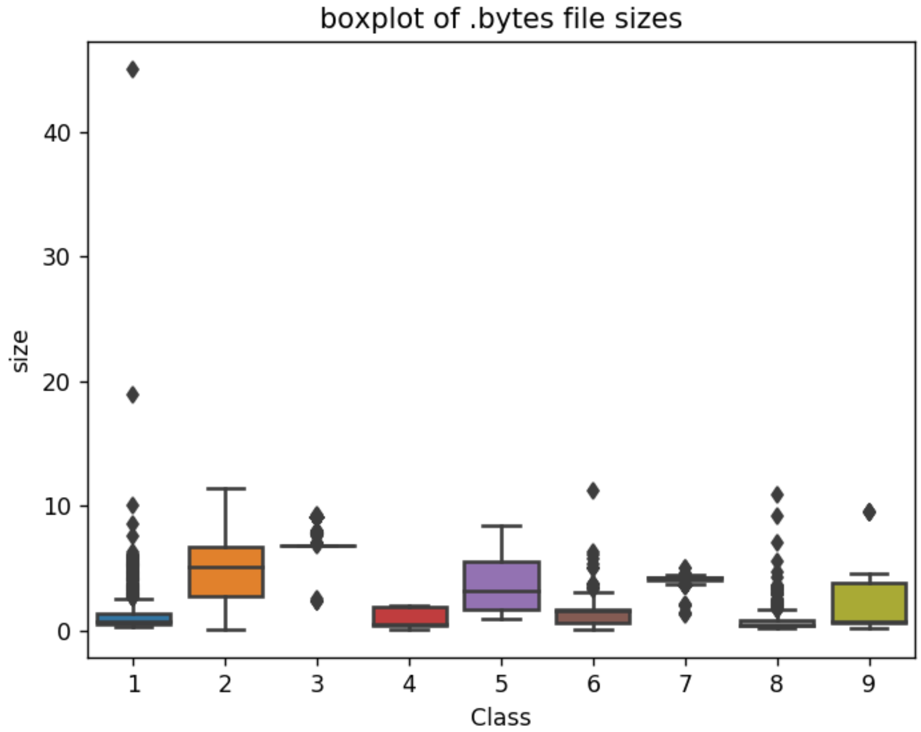


Observation:- Class 5 has less number of data points, **Imbalance data problem.**

File size as feature

ID	File Name	Size
0	01azqd4InC7m9JpocGv5	4.234863
1	01IsoiSMh5gxyDYTI4CB	5.538818
2	01jsnpXSAIgw6aPeDxrU	3.887939
3	01kcPWA9K2BOxQeS5Rju	0.574219
4	01SuzwMJEIXsK7A8dQbl	0.370850

Box plot of file size as feature



Observation:- Class 2, 5 and 9 can be easily distinguished from other classes, using only the file size feature

Copy of Bag of word as feature of the file

ID	File Name	0	1	2	3	4	5	6	7	8	...	f9	fa	fb	fc
0	01azqd4InC7m9JpocGv5	601905	3905	2816	3832	3345	3242	3650	3201	2965	...	3101	3211	3097	2758
1	01IsoiSMh5gxyDYTl4CB	39755	8337	7249	7186	8663	6844	8420	7589	9291	...	439	281	302	7639
2	01jsnpXSAlgw6aPeDxrU	93506	9542	2568	2438	8925	9330	9007	2342	9107	...	2242	2885	2863	2471
3	01kcPWA9K2BOxQeS5Rju	21091	1213	726	817	1257	625	550	523	1078	...	485	462	516	1133
4	01SuzwMJEIXsK7A8dQbl	19764	710	302	433	559	410	262	249	422	...	350	209	239	653

Copy of Combining Bag of Words and File size as Features

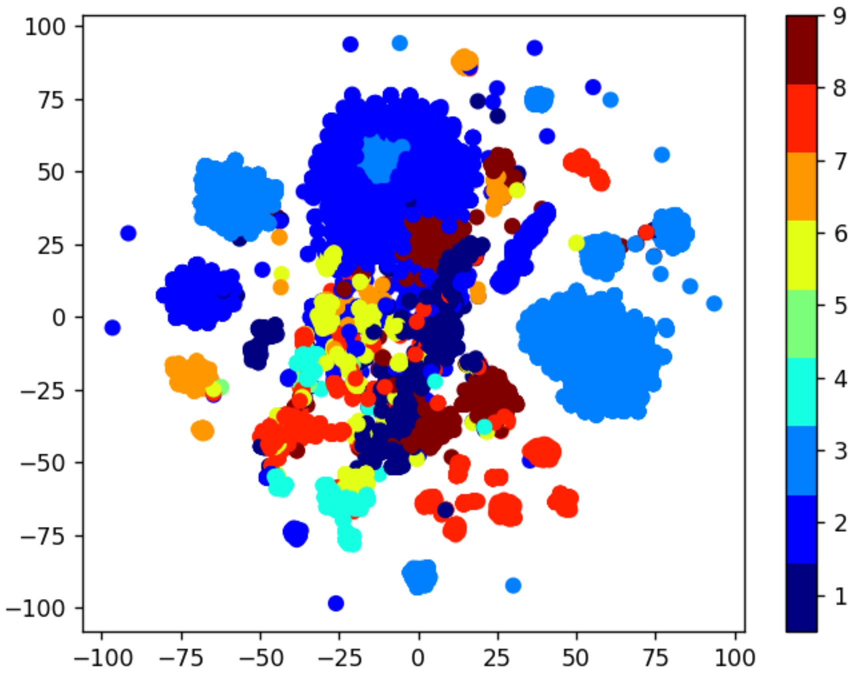
ID	File Name	0	1	2	3	4	5	6	7	8	...	f9	fa	fb	fc
0	01azqd4InC7m9JpocGv5	601905	3905	2816	3832	3345	3242	3650	3201	2965	...	3101	3211	3097	2758
1	01IsoiSMh5gxyDYTl4CB	39755	8337	7249	7186	8663	6844	8420	7589	9291	...	439	281	302	7639
2	01jsnpXSAlgw6aPeDxrU	93506	9542	2568	2438	8925	9330	9007	2342	9107	...	2242	2885	2863	2471
3	01kcPWA9K2BOxQeS5Rju	21091	1213	726	817	1257	625	550	523	1078	...	485	462	516	1133
4	01SuzwMJEIXsK7A8dQbl	19764	710	302	433	559	410	262	249	422	...	350	209	239	653

Copy of Normalizing the Features

ID	File Name	0	1	2	3	4	5	6	7	8
0	01azqd4InC7m9JpocGv5	0.262806	0.005498	0.001567	0.002067	0.002048	0.001835	0.002058	0.002946	0.002638
1	01IsoiSMh5gxyDYTI4CB	0.017358	0.011737	0.004033	0.003876	0.005303	0.003873	0.004747	0.006984	0.008267
2	01jsnpXSAlgw6aPeDxrU	0.040827	0.013434	0.001429	0.001315	0.005464	0.00528	0.005078	0.002155	0.008104
3	01kcPWA9K2BOxQeS5Rju	0.009209	0.001708	0.000404	0.000441	0.00077	0.000354	0.00031	0.000481	0.000959
4	01SuzwMJEIXsK7A8dQbl	0.008629	0.001	0.000168	0.000234	0.000342	0.000232	0.000148	0.000229	0.000376

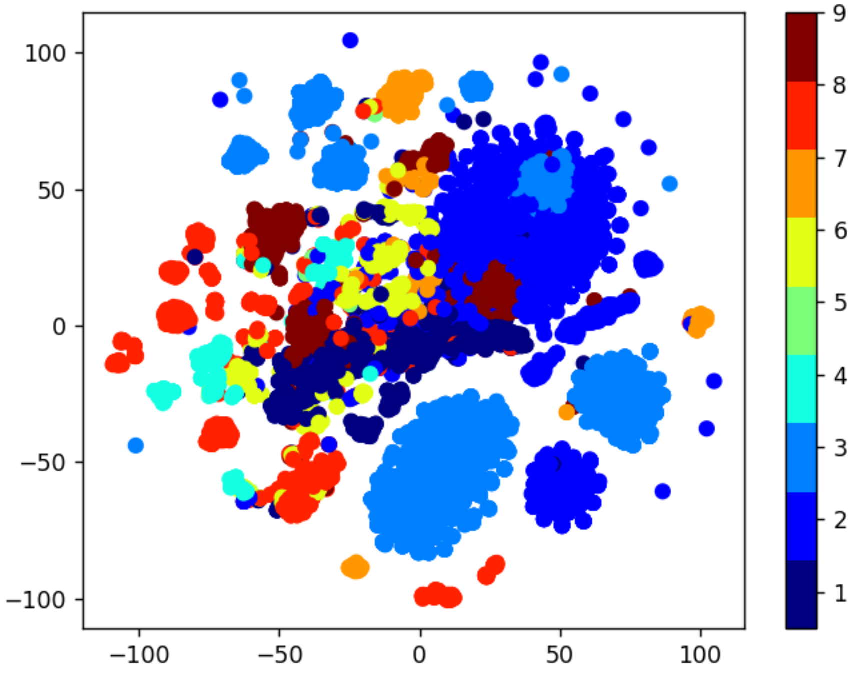
Multivariate analysis of the Features

Perplexity = 50



Observation:- Class 2 and 3 are clearly separated whereas other classes have poor distinctions

Perplexity = 30



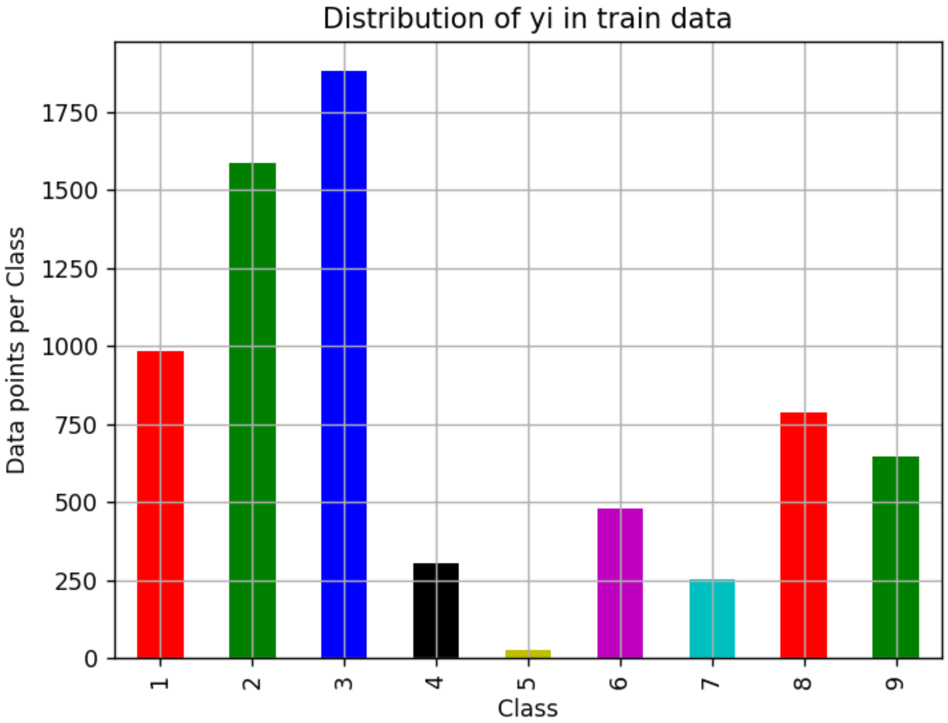
Observation:- Class 2 and 3 are clearly separated whereas other classes have poor distinctions

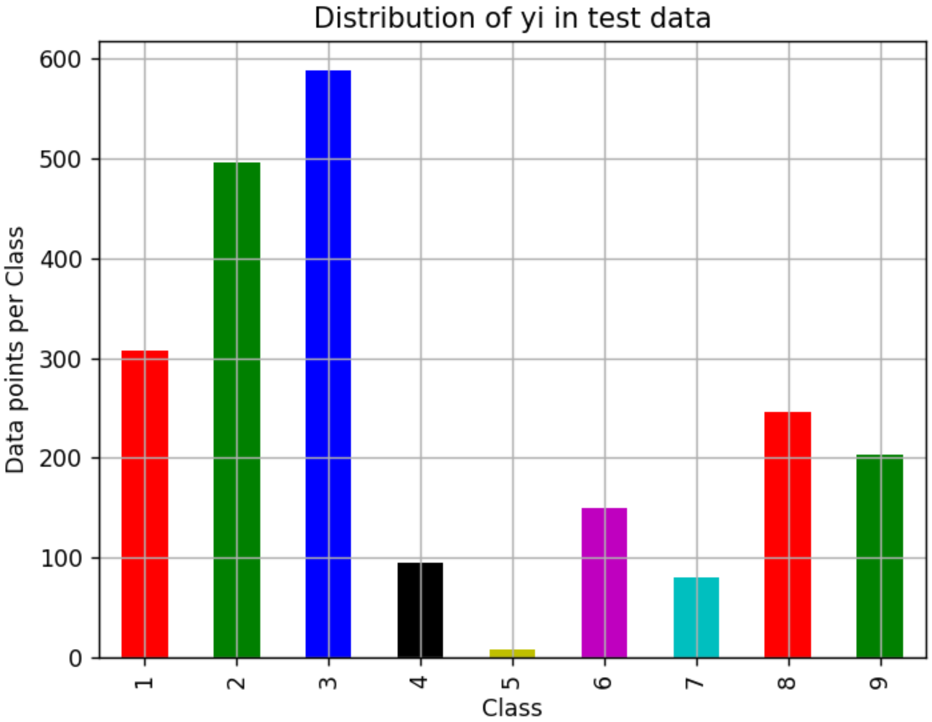
Test Train Split

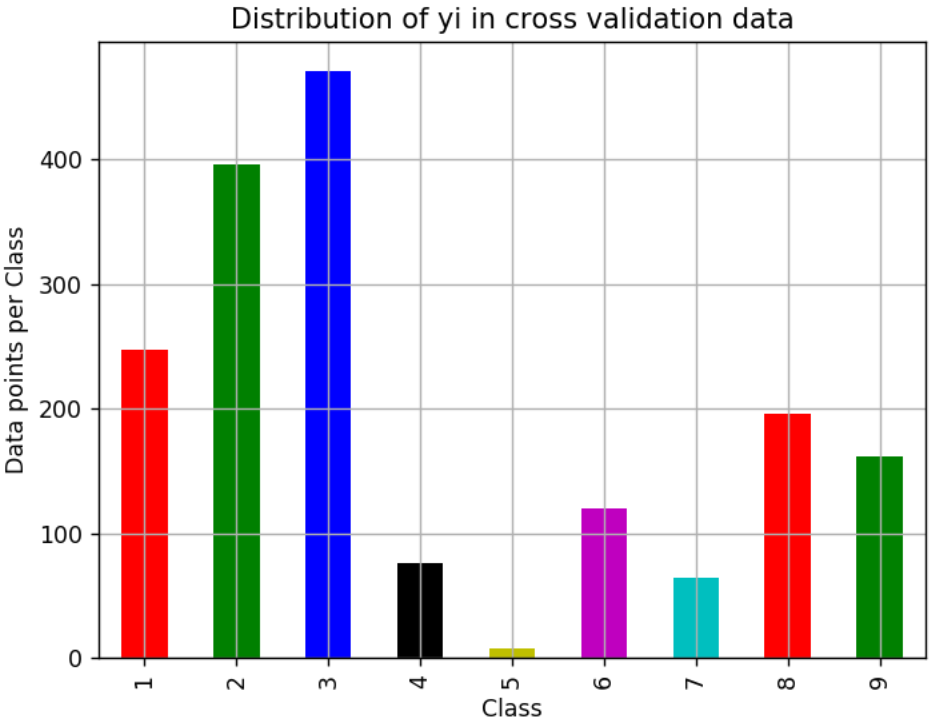
Number of data points in train data: 6955
Number of data points in test data: 2174
Number of data points in cross validation data: 1739

Check for distribution of data

We check for the distribution of classes in each split by plotting a histogram.


Distribution of yi in train data
Data points per Class
0
250
500
750
1000
1250
1500
1750
1
2
3
4
5
6
7
8
9
Class


Distribution of yi in test data
Data points per Class
0
100
200
300
400
500
600
1
2
3
4
5
6
7
8
9
Class

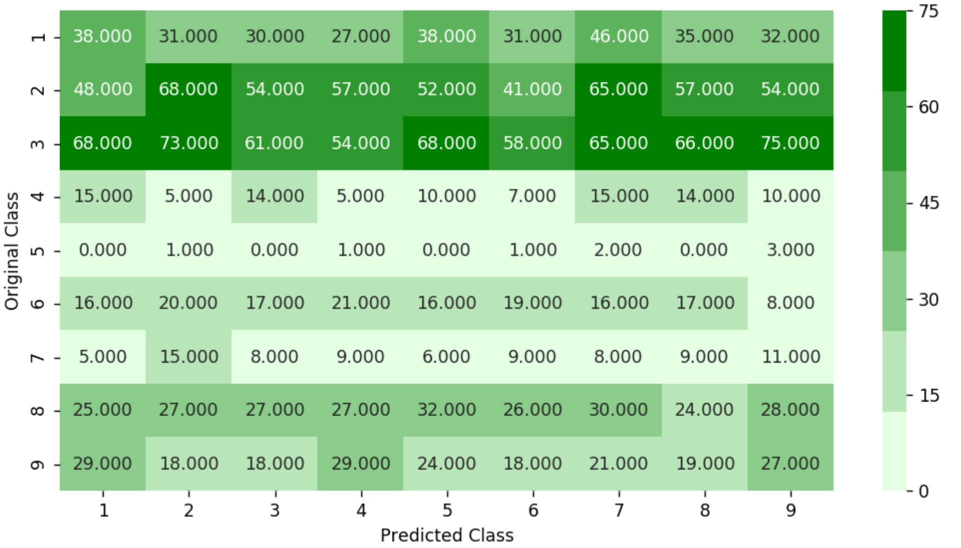


Machine Learning Model

Random Model

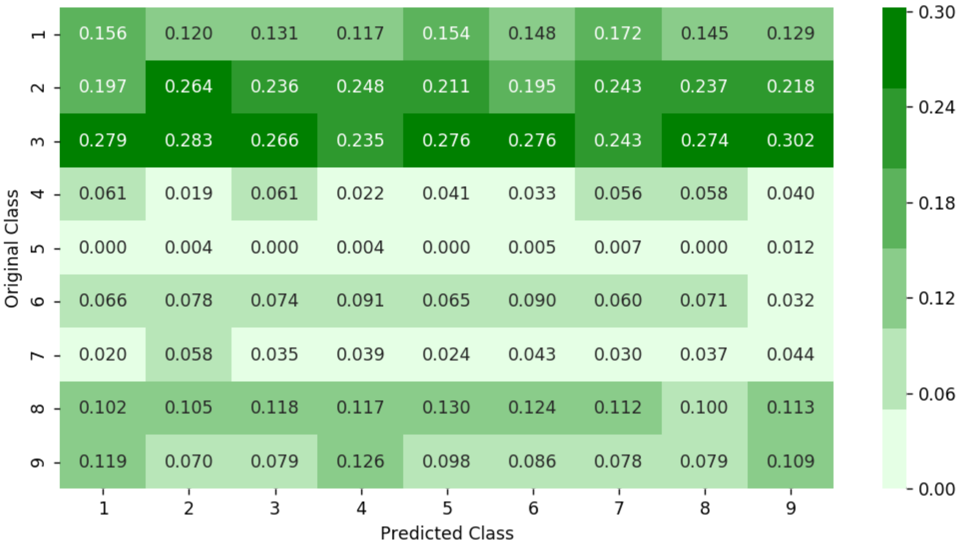
Log loss on Cross Validation Data using Random Model 2.46
Log loss on Test Data using Random Model 2.48
Accuracy 11.49

Confusion Matrix



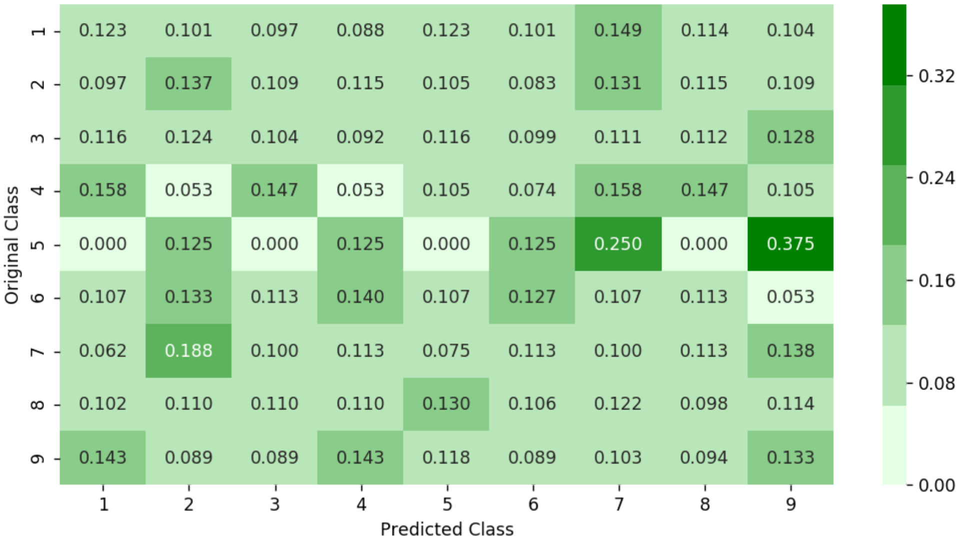
	1	2	3	4	5	6	7	8	9
1	38.000	31.000	30.000	27.000	38.000	31.000	46.000	35.000	32.000
2	48.000	68.000	54.000	57.000	52.000	41.000	65.000	57.000	54.000
3	68.000	73.000	61.000	54.000	68.000	58.000	65.000	66.000	75.000
4	15.000	5.000	14.000	5.000	10.000	7.000	15.000	14.000	10.000
5	0.000	1.000	0.000	1.000	0.000	1.000	2.000	0.000	3.000
6	16.000	20.000	17.000	21.000	16.000	19.000	16.000	17.000	8.000
7	5.000	15.000	8.000	9.000	6.000	9.000	8.000	9.000	11.000
8	25.000	27.000	27.000	27.000	32.000	26.000	30.000	24.000	28.000
9	29.000	18.000	18.000	29.000	24.000	18.000	21.000	19.000	27.000

Precision Matrix



	1	2	3	4	5	6	7	8	9
1	0.156	0.120	0.131	0.117	0.154	0.148	0.172	0.145	0.129
2	0.197	0.264	0.236	0.248	0.211	0.195	0.243	0.237	0.218
3	0.279	0.283	0.266	0.235	0.276	0.276	0.243	0.274	0.302
4	0.061	0.019	0.061	0.022	0.041	0.033	0.056	0.058	0.040
5	0.000	0.004	0.000	0.004	0.000	0.005	0.007	0.000	0.012
6	0.066	0.078	0.074	0.091	0.065	0.090	0.060	0.071	0.032
7	0.020	0.058	0.035	0.039	0.024	0.043	0.030	0.037	0.044
8	0.102	0.105	0.118	0.117	0.130	0.124	0.112	0.100	0.113
9	0.119	0.070	0.079	0.126	0.098	0.086	0.078	0.079	0.109

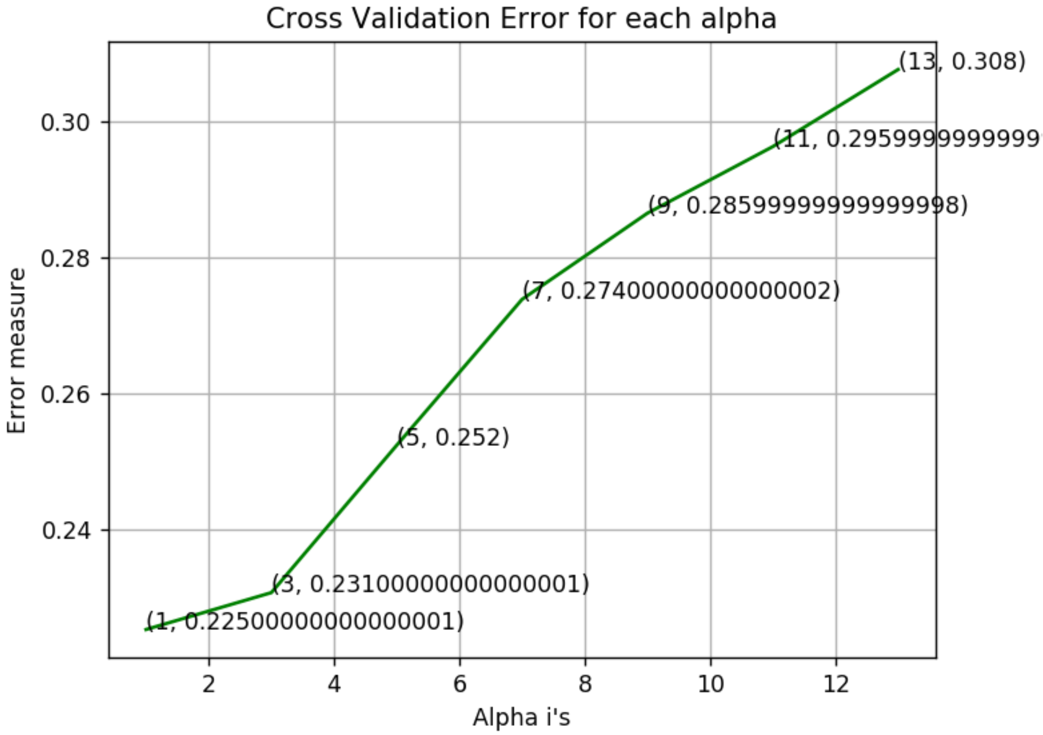
Recall Matrix



K Nearest Neighbor Classification

Hyperparameter Search

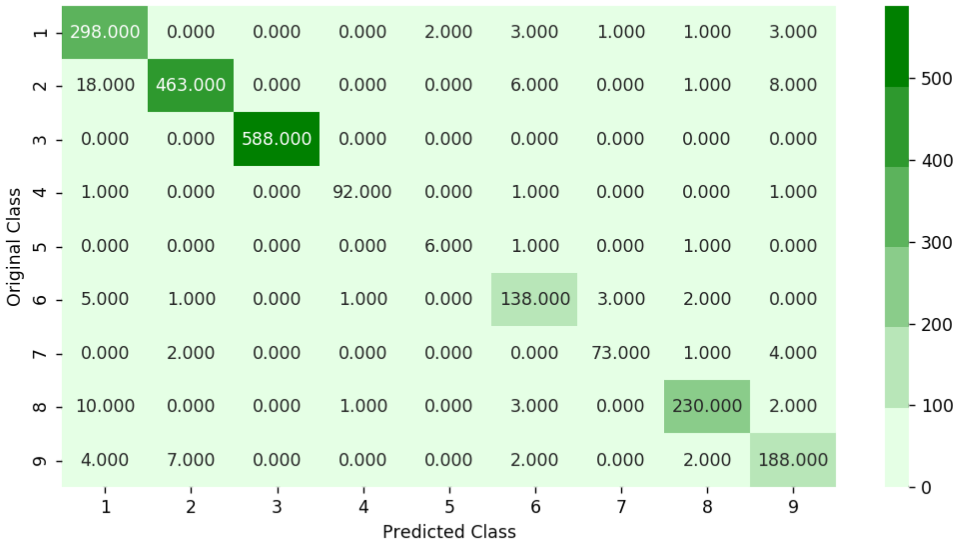
```
log_loss for k =  1 is 0.225386237304
log_loss for k =  3 is 0.230795229168
log_loss for k =  5 is 0.252421408646
log_loss for k =  7 is 0.273827486888
log_loss for k =  9 is 0.286469181555
log_loss for k =  11 is 0.29623391147
log_loss for k =  13 is 0.307551203154
```



Results from the Best Model

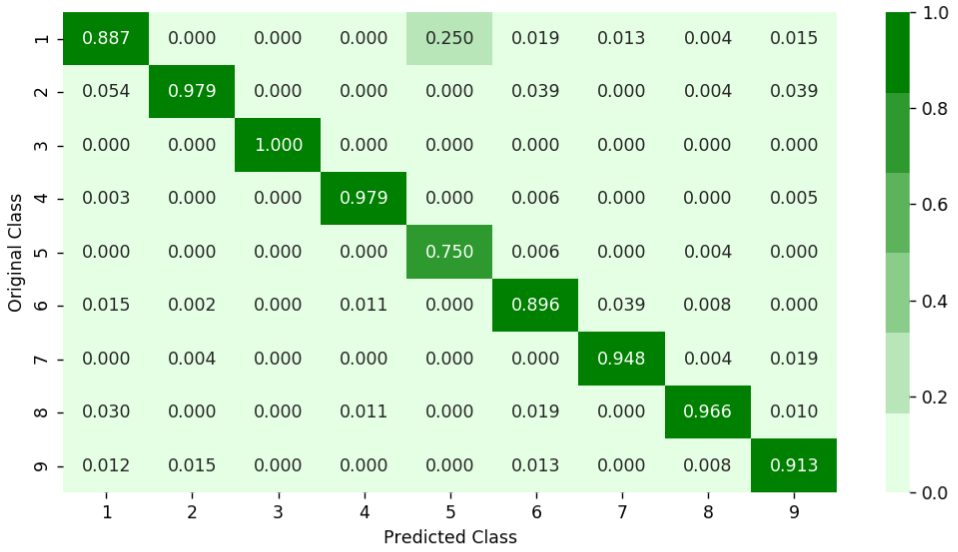
For values of best alpha = 1 The train log loss is: 0.08
For values of best alpha = 1 The cross validation log loss is: 0.23
For values of best alpha = 1 The test log loss is: 0.24
Accuracy 95.49

Confusion Matrix

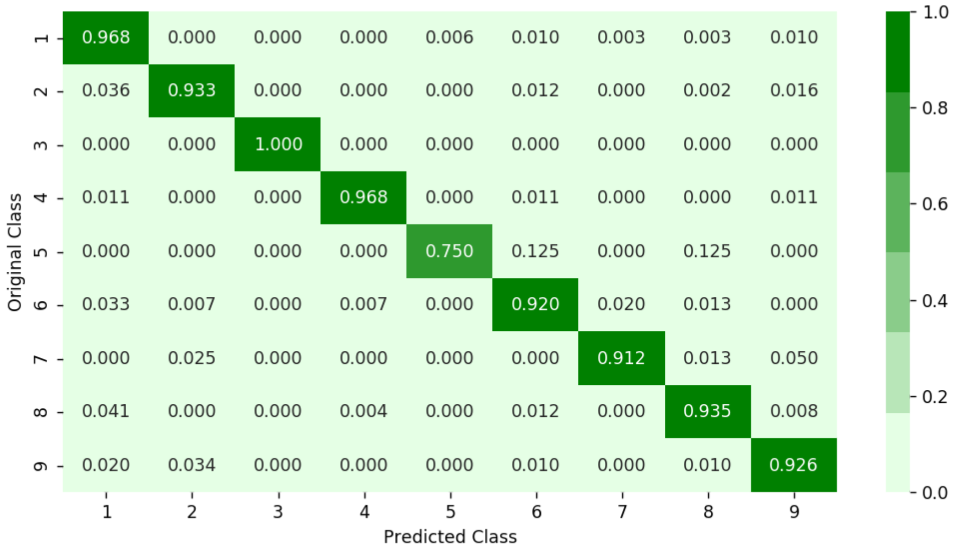


	1	2	3	4	5	6	7	8	9
1	298.000	0.000	0.000	0.000	2.000	3.000	1.000	1.000	3.000
2	18.000	463.000	0.000	0.000	0.000	6.000	0.000	1.000	8.000
3	0.000	0.000	588.000	0.000	0.000	0.000	0.000	0.000	0.000
4	1.000	0.000	0.000	92.000	0.000	1.000	0.000	0.000	1.000
5	0.000	0.000	0.000	0.000	6.000	1.000	0.000	1.000	0.000
6	5.000	1.000	0.000	1.000	0.000	138.000	3.000	2.000	0.000
7	0.000	2.000	0.000	0.000	0.000	0.000	73.000	1.000	4.000
8	10.000	0.000	0.000	1.000	0.000	3.000	0.000	230.000	2.000
9	4.000	7.000	0.000	0.000	0.000	2.000	0.000	2.000	188.000

Precision Matrix



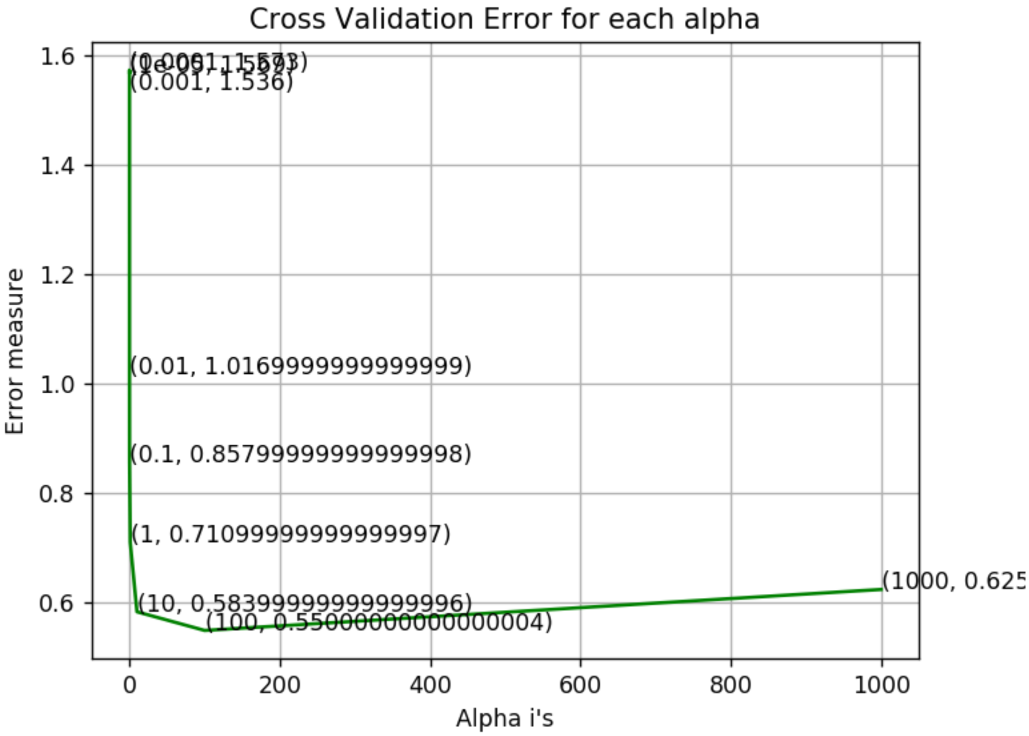
Recall Matrix



Logistic Regression

Hyperparameter Search

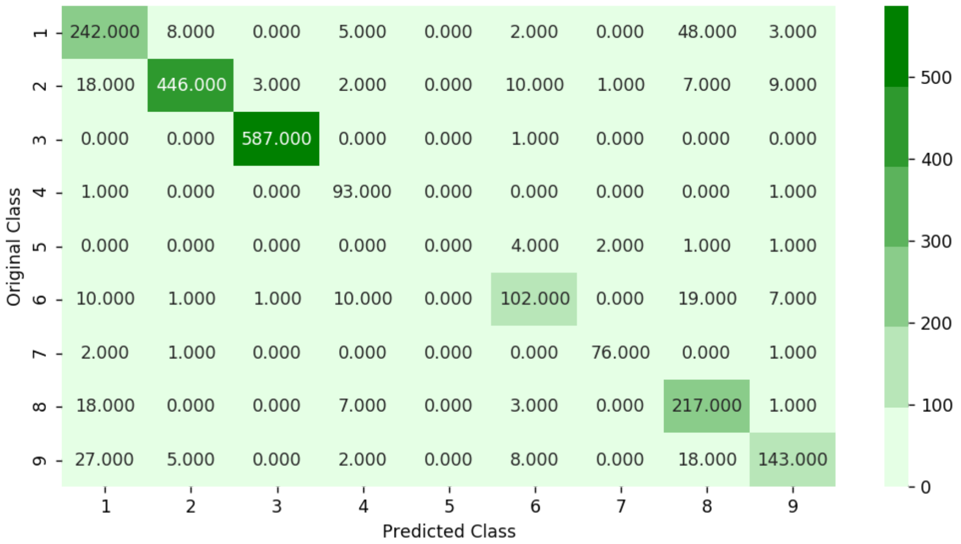
```
log_loss for c =  1e-05 is 1.56916911178
log_loss for c =  0.0001 is 1.57336384417
log_loss for c =  0.001 is 1.53598598273
log_loss for c =  0.01 is 1.01720972418
log_loss for c =  0.1 is 0.857766083873
log_loss for c =  1 is 0.711154393309
log_loss for c =  10 is 0.583929522635
log_loss for c =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



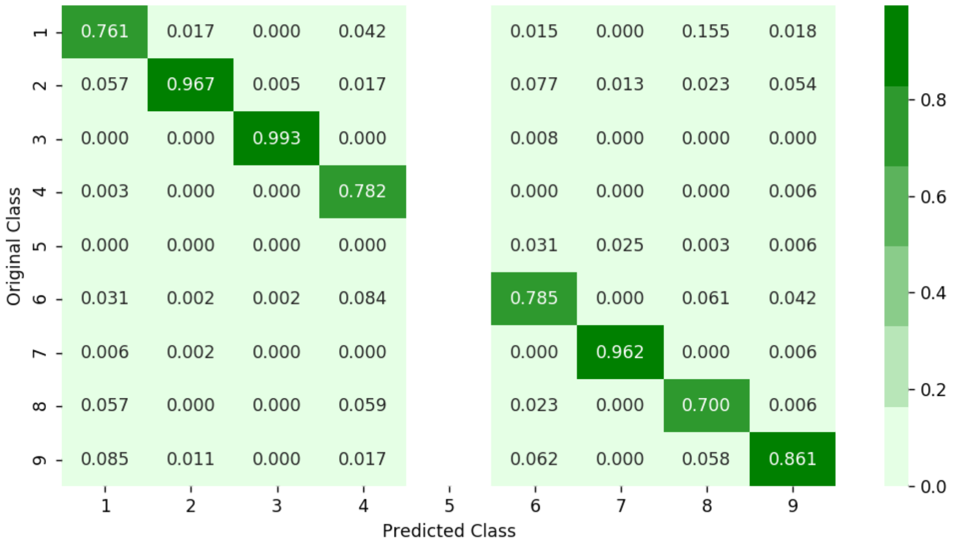
Results from the Best Model

```
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
```

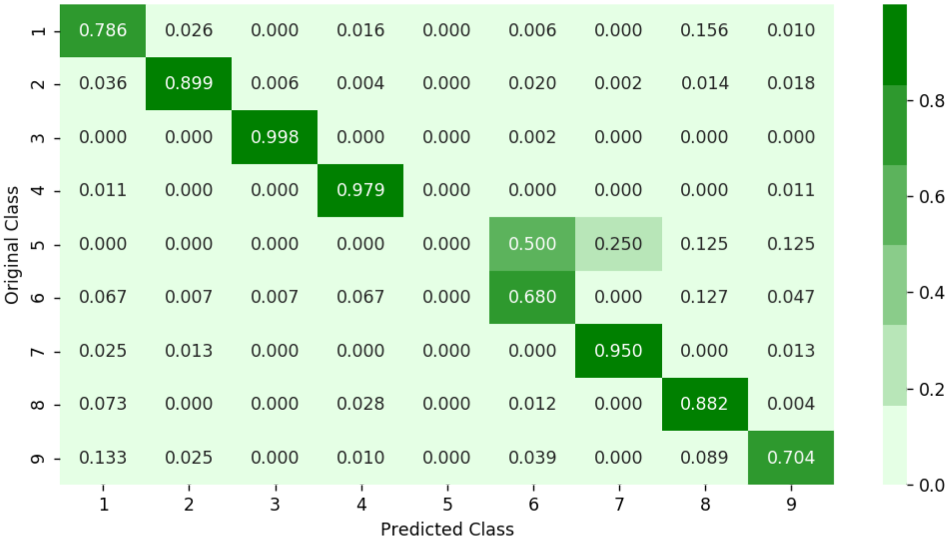
Confusion Matrix



Precision Matrix



Recall Matrix

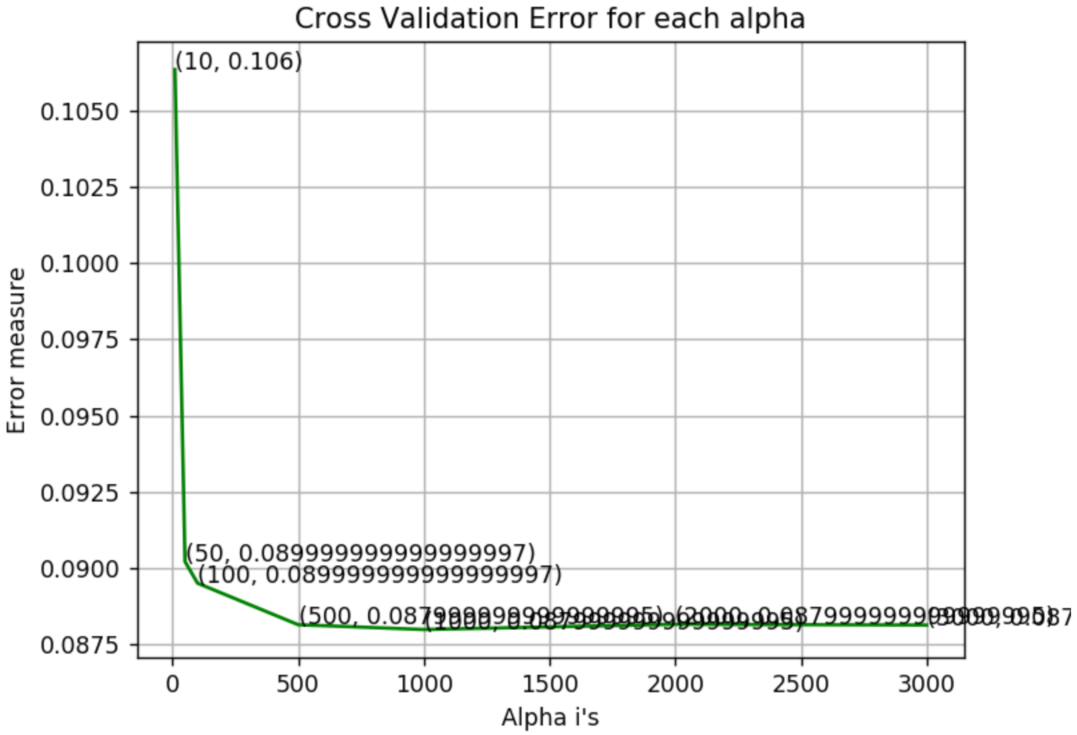


	1	2	3	4	5	6	7	8	9
1	0.786	0.026	0.000	0.016	0.000	0.006	0.000	0.156	0.010
2	0.036	0.899	0.006	0.004	0.000	0.020	0.002	0.014	0.018
3	0.000	0.000	0.998	0.000	0.000	0.002	0.000	0.000	0.000
4	0.011	0.000	0.000	0.979	0.000	0.000	0.000	0.000	0.011
5	0.000	0.000	0.000	0.000	0.000	0.500	0.250	0.125	0.125
6	0.067	0.007	0.007	0.067	0.000	0.680	0.000	0.127	0.047
7	0.025	0.013	0.000	0.000	0.000	0.000	0.950	0.000	0.013
8	0.073	0.000	0.000	0.028	0.000	0.012	0.000	0.882	0.004
9	0.133	0.025	0.000	0.010	0.000	0.039	0.000	0.089	0.704

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

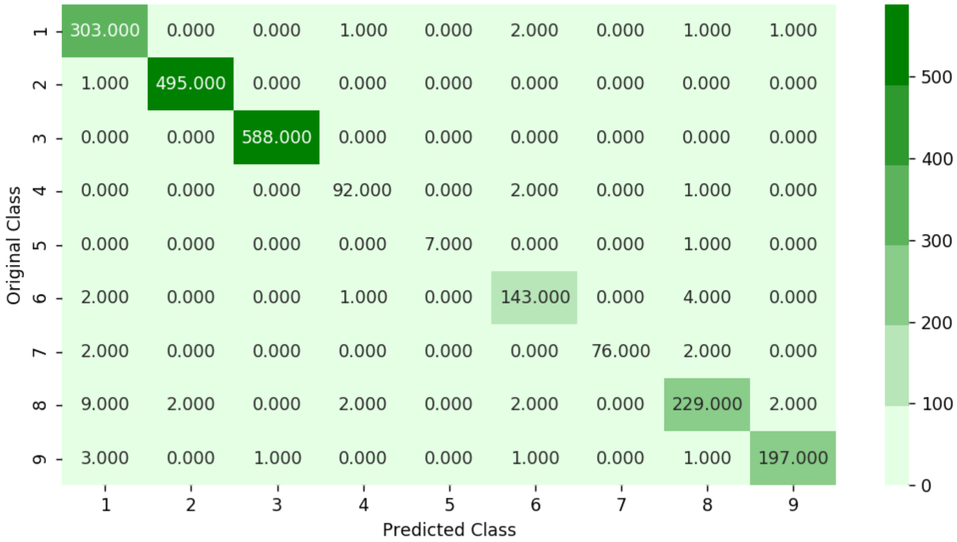
For values of best alpha = 1000 The train log loss is: 0.031

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

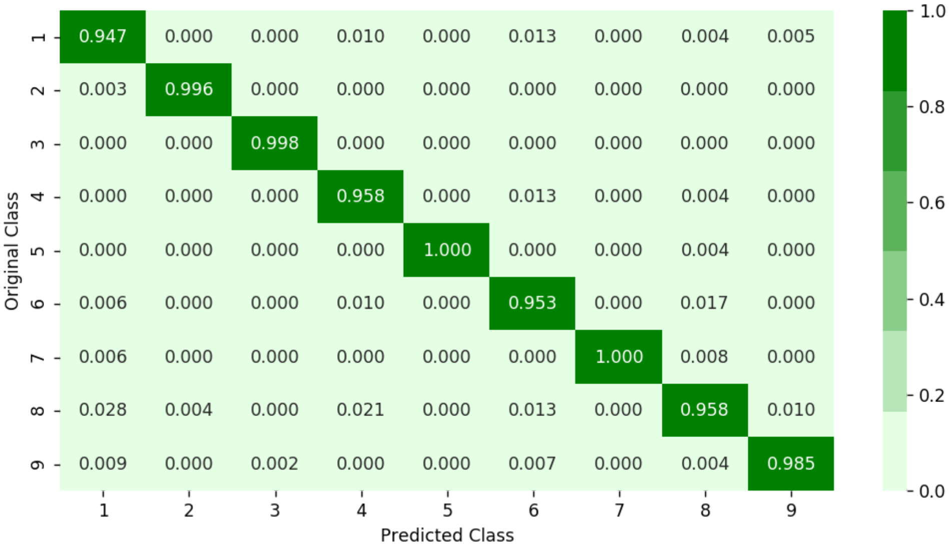
Accuracy 96.76

Confusion Matrix



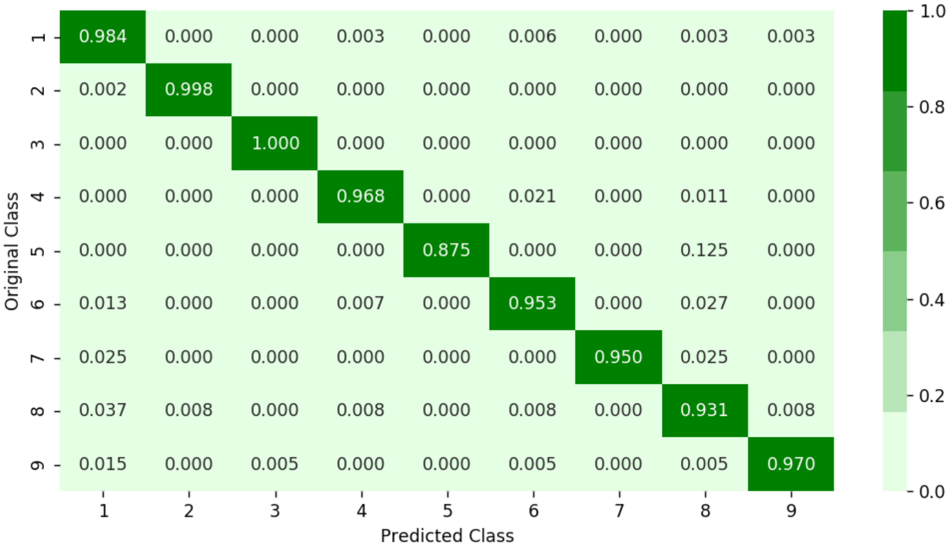
	1	2	3	4	5	6	7	8	9
1	303.000	0.000	0.000	1.000	0.000	2.000	0.000	1.000	1.000
2	1.000	495.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	588.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	92.000	0.000	2.000	0.000	1.000	0.000
5	0.000	0.000	0.000	0.000	7.000	0.000	0.000	1.000	0.000
6	2.000	0.000	0.000	1.000	0.000	143.000	0.000	4.000	0.000
7	2.000	0.000	0.000	0.000	0.000	0.000	76.000	2.000	0.000
8	9.000	2.000	0.000	2.000	0.000	2.000	0.000	229.000	2.000
9	3.000	0.000	1.000	0.000	0.000	1.000	0.000	1.000	197.000

Precision Matrix



	1	2	3	4	5	6	7	8	9
1	0.947	0.000	0.000	0.010	0.000	0.013	0.000	0.004	0.005
2	0.003	0.996	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.998	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.958	0.000	0.013	0.000	0.004	0.000
5	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.004	0.000
6	0.006	0.000	0.000	0.010	0.000	0.953	0.000	0.017	0.000
7	0.006	0.000	0.000	0.000	0.000	0.000	1.000	0.008	0.000
8	0.028	0.004	0.000	0.021	0.000	0.013	0.000	0.958	0.010
9	0.009	0.000	0.002	0.000	0.000	0.007	0.000	0.004	0.985

Recall Matrix

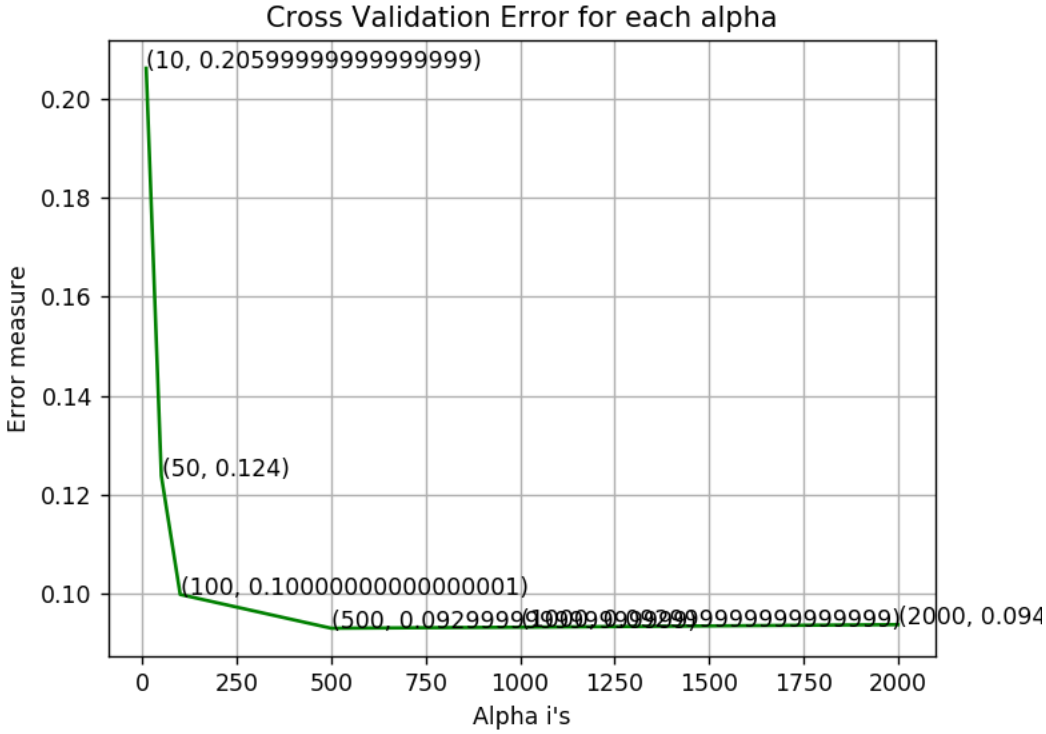


	1	2	3	4	5	6	7	8	9
1	0.984	0.000	0.000	0.003	0.000	0.006	0.000	0.003	0.003
2	0.002	0.998	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.968	0.000	0.021	0.000	0.011	0.000
5	0.000	0.000	0.000	0.000	0.875	0.000	0.000	0.125	0.000
6	0.013	0.000	0.000	0.007	0.000	0.953	0.000	0.027	0.000
7	0.025	0.000	0.000	0.000	0.000	0.000	0.950	0.025	0.000
8	0.037	0.008	0.000	0.008	0.000	0.008	0.000	0.931	0.008
9	0.015	0.000	0.005	0.000	0.000	0.005	0.000	0.005	0.970

XgBoost Classification

Hyperparameter Search

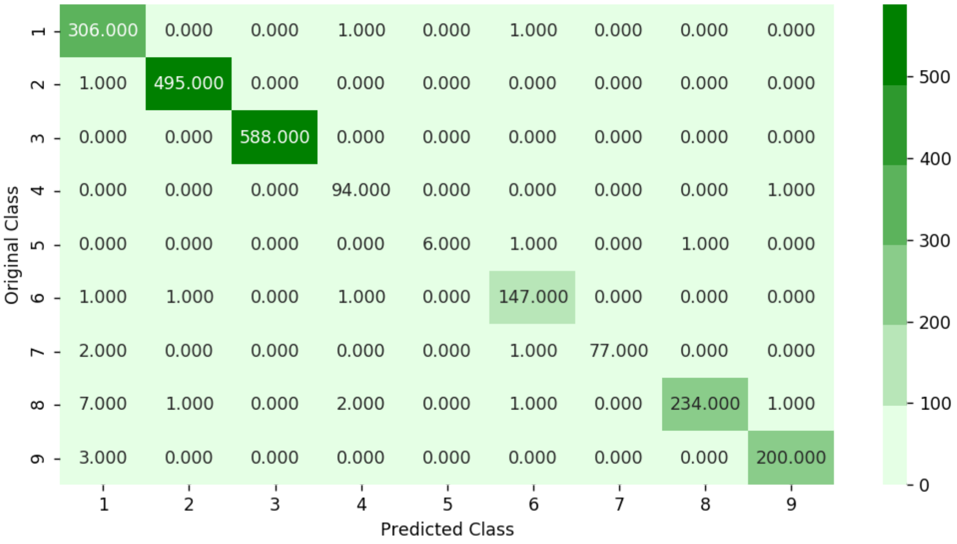
```
log_loss for c =  10 is 0.20615980494
log_loss for c =  50 is 0.123888382365
log_loss for c =  100 is 0.099919437112
log_loss for c =  500 is 0.0931035681289
log_loss for c =  1000 is 0.0933084876012
log_loss for c =  2000 is 0.0938395690309
```



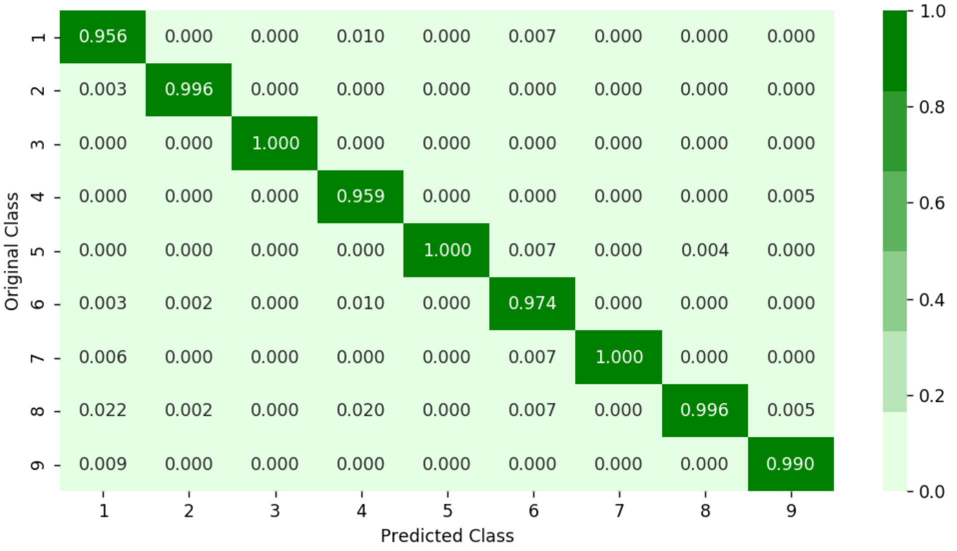
Results from the Best Model

```
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
```

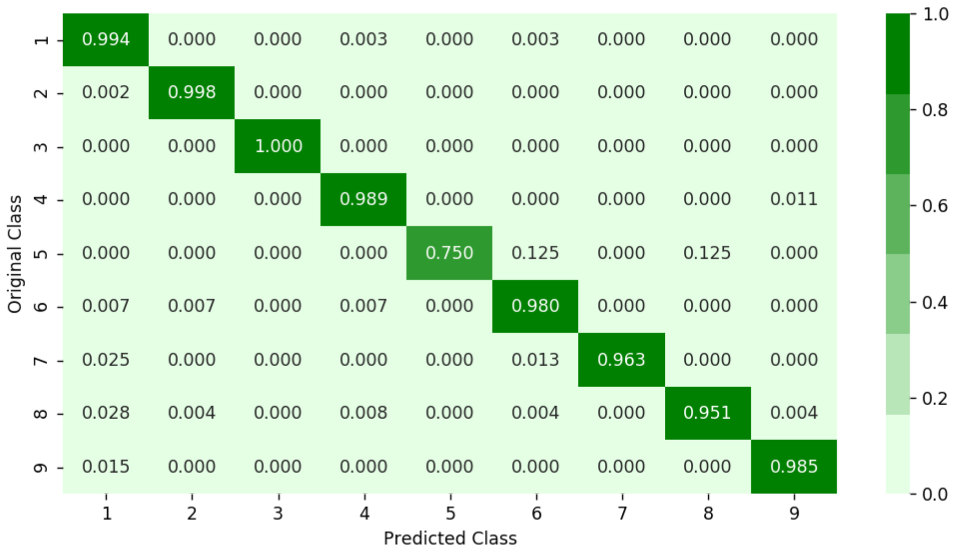
Confusion Matrix



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_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': 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