**Cross-Validation Results on Last 50 Training Data**

SVM:

0.8109393579072532 100.0

[0.76470588 0.70588235 0.6875]

K-nearest neighbor:

[0.70588235 0.70588235 0.625]

Random forest:

[0.76470588 0.76470588 0.6875]

Logistic regression:

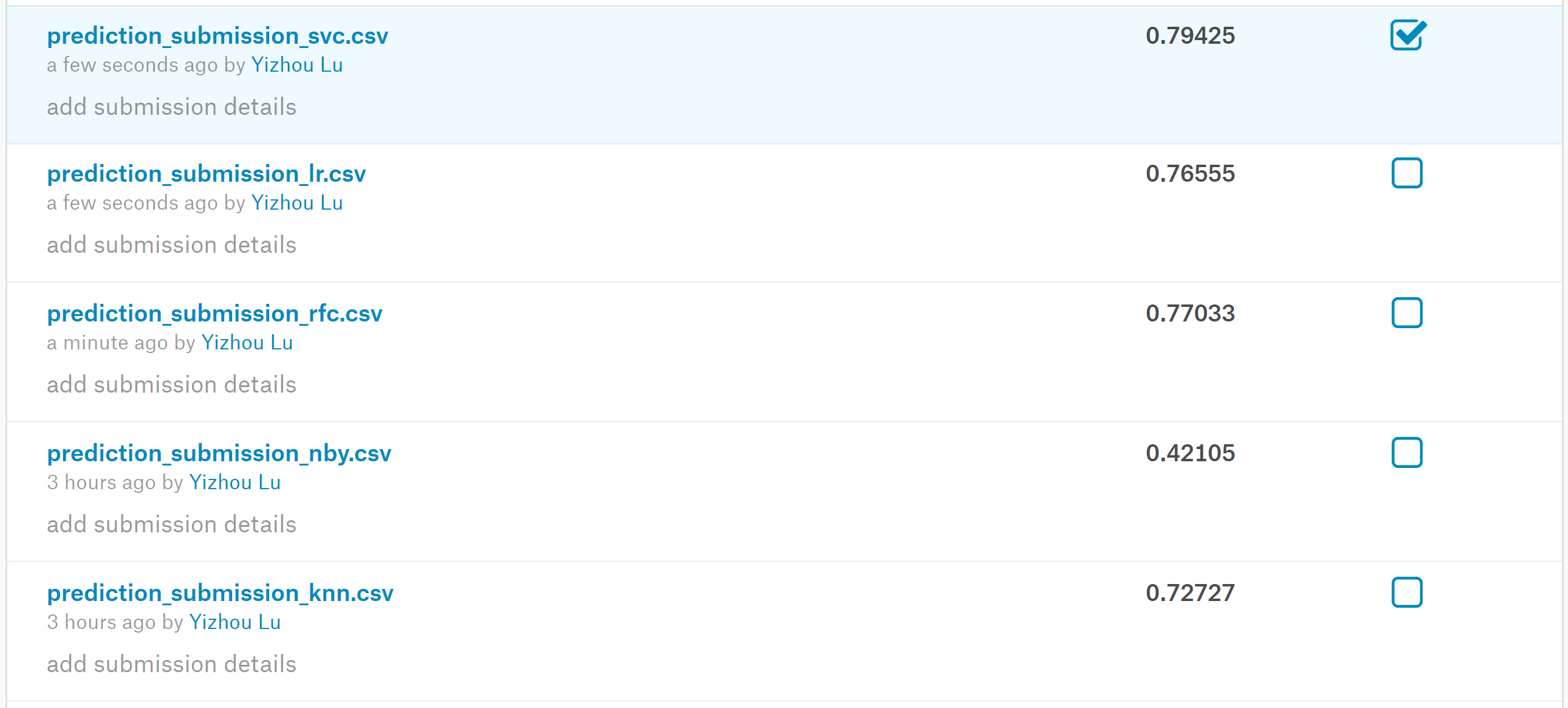
0.8335315101070154 4.281332398719393

[0.82352941 0.70588235 0.75]

Naïve bayes

[0.58823529 0.35294118 0.5]

**Test Data Accuracy Scores:**



SVM is the best performer, but logistic regression and random forest are not far behind. Naïve Bayes is eliminated from further consideration.

**Adding PCA to Remove Redundancy:**

SVM:

0.7883790882901204 100.0

[0.76470588 0.70588235 0.6875]

K-nearest neighbor:

[0.70588235 0.64705882 0.625]

Random forest:

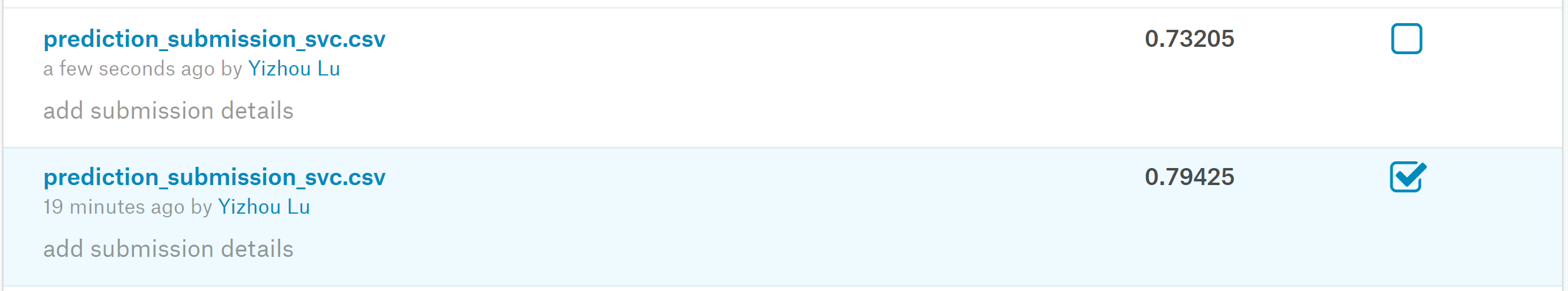
[0.82352941 0.64705882 0.6875]

Logistic regression:

0.7942975766819184 0.12742749857031338

[0.70588235 0.64705882 0.8125]

**Test Data Accuracy Scores with PCA:**



As expected, the PCA makes prediction accuracy considerably worse. With such a small number of features, PCA does not help and does not really make sense.

**Applying Grid Search to Random Forest:**

SVM:

0.8109642433485851 100.0

[0.76470588 0.70588235 0.6875]

K-nearest neighbor:

[0.70588235 0.70588235 0.625]

Random forest:

0.8204880528723946 50

[0.88235294 0.76470588 0.6875]

Logistic regression:

0.8335451618369768 4.281332398719393

[0.82352941 0.70588235 0.75]

**Test Data Accuracy Scores:**

No improvement over the case without grid search! It didn’t help for random forest!

**After Improvements on Establishment of Some Features (Ticket, Cabin):**

SVM:

0.762222504660227 4.281332398719393

[0.58823529 0.70588235 0.75]

K-nearest neighbor:

[0.64705882 0.52941176 0.5625]

Random forest:

0.8097737671581088 55

[0.82352941 0.70588235 0.6875]

Logistic regression:

0.8228605321132011 2.0691380811147897

[0.76470588 0.70588235 0.75]

**Test Data Accuracy Scores:**

SVM:

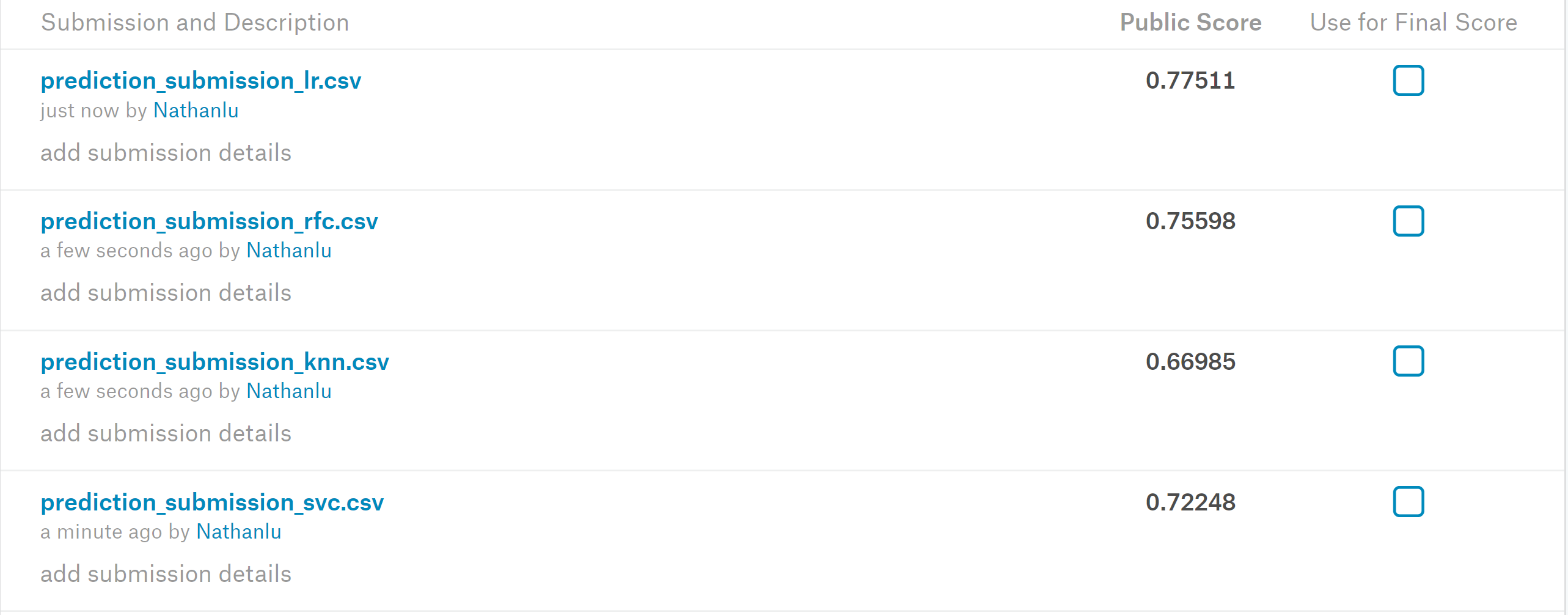
0.7654320987654321 8.858667904100825

Random forest:

0.8193041526374859 35

Logistic regression:

0.8260381593714926 4.832930238571752



Logistic regression overtook SVM as the best performer, while random forest is only slightly behind. KNN is eliminated from further consideration because of significantly poorer performance. **However, the overall performance of algorithms become even slightly poorer. No improvement!**

**Including Neural Network and Sensitivity Analysis Without Cabin:**

SVM:

0.7705388917132688 29.76351441631318

[0.58823529 0.64705882 0.8125]

Random forest:

0.8121462463989154 35

[0.82352941 0.70588235 0.8125]

Logistic regression:

0.8299652601254025 1.6237767391887217

[0.76470588 0.76470588 0.75]

Neural Network:

0.8327721661054994 180 identity

[0.70588235 0.82352941 0.5625]

**Test Data Accuracy Scores:**

0.7800224466891134 37.926901907322495

0.8103254769921437 35

0.8282828282828283 2.3357214690901213

0.830527497194164 160 identity

**Back to Original Feature Sets (with the highest score) and Sensitivity Analysis: Without Fare and Cabin**

SVM:

0.8276054905948144 61.584821106602604

[0.64705882 0.76470588 0.875]

Random forest:

0.8097695305880359 10

[0.76470588 0.76470588 0.8125]

Logistic regression:

0.8323631587866464 3.3598182862837818

[0.76470588 0.76470588 0.75]

Neural Network:

0.8395061728395062 70 tanh

[0.76470588 0.82352941 0.875]

**Test Data Accuracy Scores:**

0.8338945005611672 100.0

0.8215488215488217 80

0.8428731762065095 2.9763514416313175

0.8395061728395062 40 identity

Almost the same as with Cabin and Fare and with Fare only! Neural Network is not performing better.

**After Mean Normalization of Age and Fare:**

SVM:

0.8323589222165735 329.03445623126674

[0.82352941 0.70588235 0.8125]

Random forest:

0.8133409591594646 45

[0.82352941 0.70588235 0.8125]

Logistic regression:

0.8299652601254025 2.636650898730358

[0.82352941 0.76470588 0.75]

Neural Network:

0.8058361391694725 70 identity

[0.82352941 0.76470588 0.8125]

**Test Data Accuracy Scores:**

SVM:

0.8395061728395062 204.33597178569417

Random forest:

0.809203142536476 35

Logistic regression:

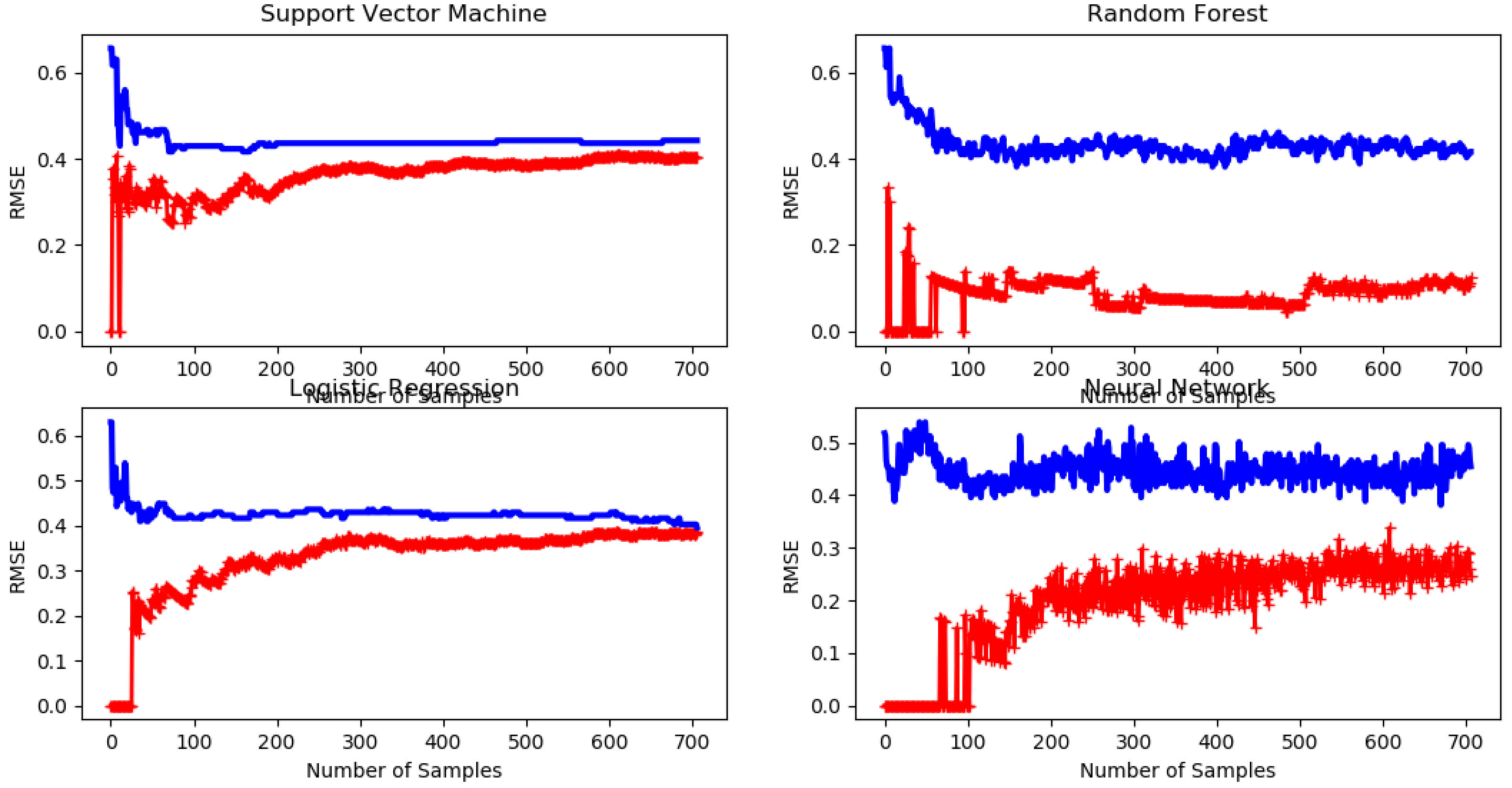
0.8316498316498316 2.3357214690901213

Neural network:

0.8069584736251403 60 identity

No improvement!

**Implementing Learning Curves:**



Random forest (performing ok in training, but worse in cv data) seems overfitting, while all the other algorithms seem underfitting.

SVM:

0.8342138543180986 14.38449888287663

[0.76666667 0.78333333 0.81355932]

Random forest:

0.8356498717630512 20

[0.73333333 0.71666667 0.81355932]

Logistic regression:

0.8314245529435403 1.2742749857031337

[0.8 0.78333333 0.84745763]

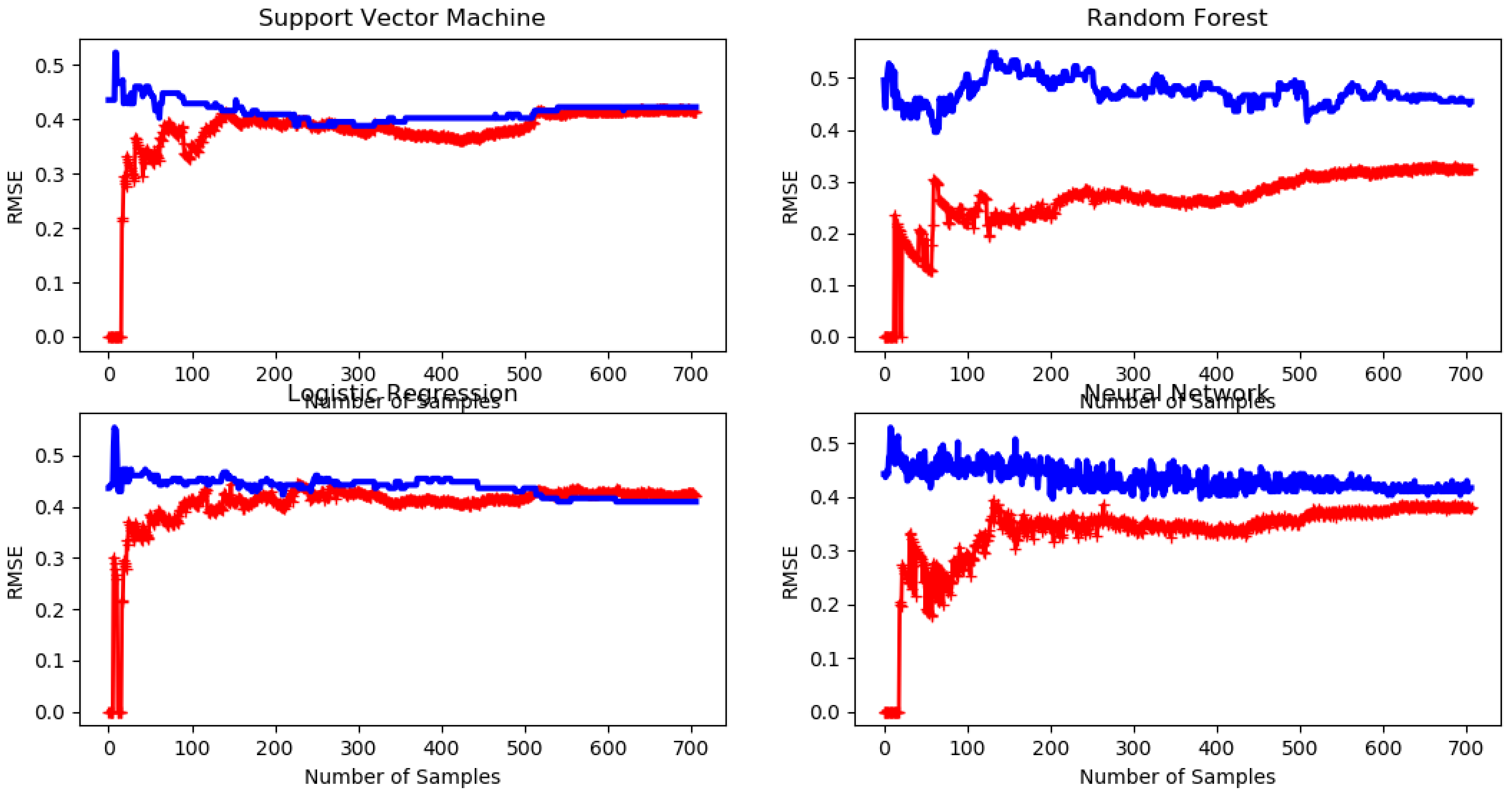
Neural Network:

0.8089564939900011 10 logistic

[0.76666667 0.68333333 0.81355932]

**Modifying and Simplifying Features:**

Learning Curves:



Random forest (performing ok in training, but worse in cv data) seems overfitting, while all the other algorithms seem underfitting.

SVM:

0.8048198182226477 100.0

[0.85 0.83333333 0.72881356]

Random forest:

0.7949568012859153 10

[0.68333333 0.8 0.74576271]

Logistic regression:

0.8019891500904159 3.3598182862837818

[0.78333333 0.83333333 0.71186441]

Neural Network:

0.8300653594771242 120 relu

[0.81666667 0.8 0.71186441]