Football Manager Dataset

Player Ability Analysis

BFOR 516 Project:
Nicholas Lopez, Sajjad Khan, Jackson Nahom, Omer Keskin

Dataset

Football Manager 2020 Dataset (Kaggle)

- This dataset provides a collection of football players and their stats, such as age, position, club, nationality, value, wage, all player attribute components.

- Size: 35MB, 64 Columns, 144750 rows/players

- Key Columns: Wage, Value, Age, Long Shots, Long Throws, Passing, Dribbling, Aggression, Stamina, etc. (56 of the columns were used)

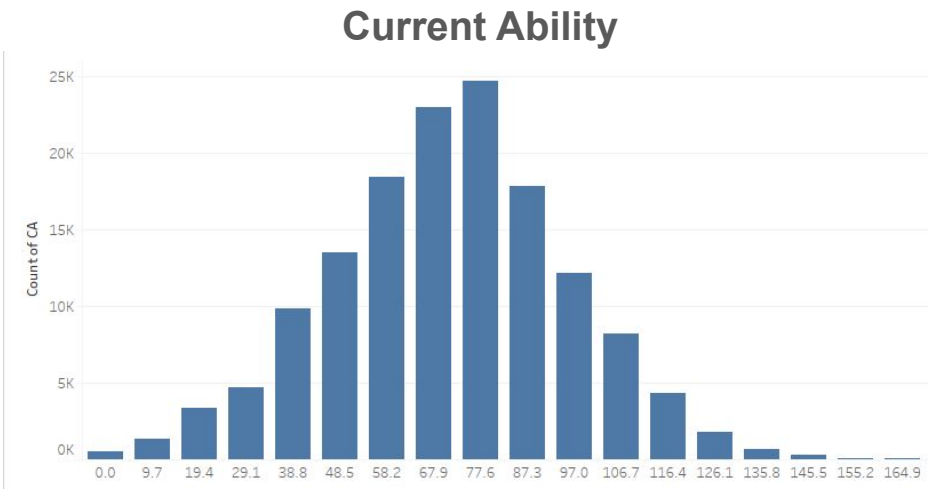
What is this Data

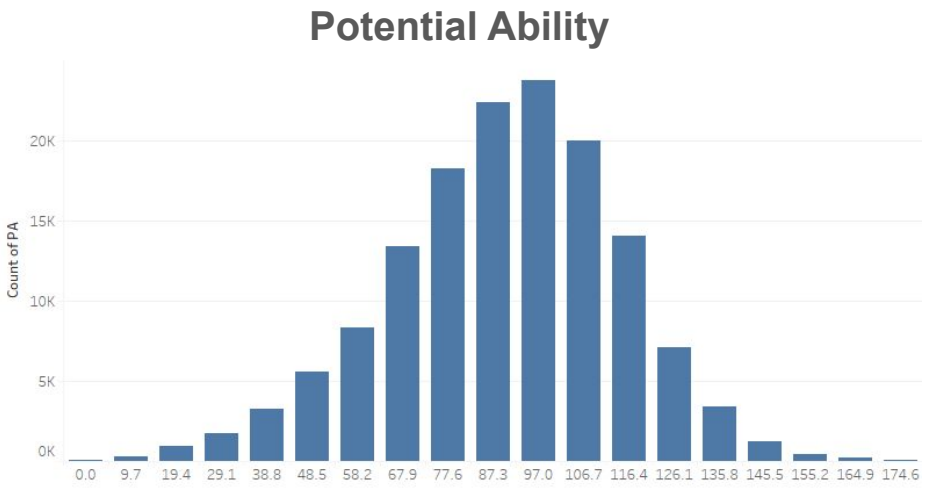
- Football Manager is a simulation game that could best be described as a spreadsheet simulator
- The data used comes from a network of 13,000 scouts of at varying levels that send reports to the developer Sports Interactive
- The data is so expansive and detailed that some high level clubs pay Sports Interactive for access to their database
- Everton publicly signed a database deal with Sports Interactive back in 2008

Business Questions

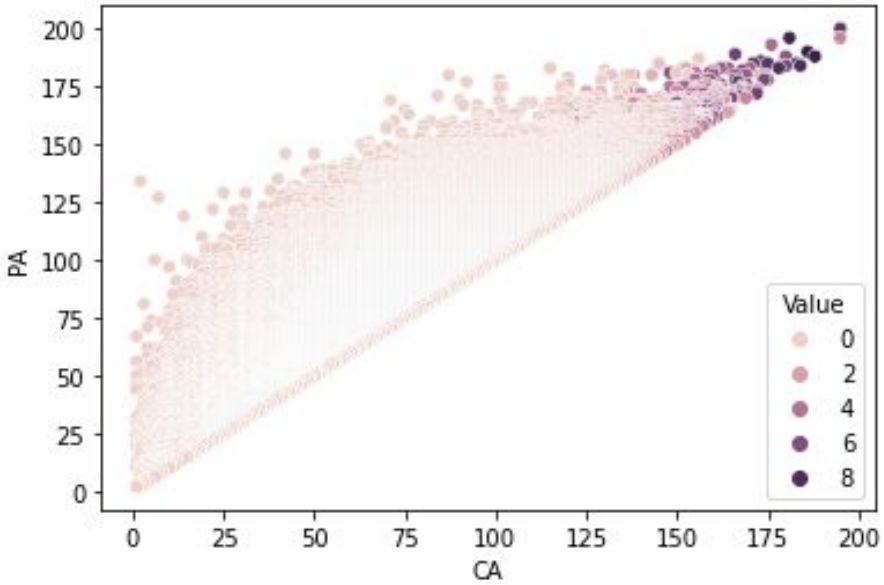
1. Can we predict a player's Current Ability?
2. Can we predict a player's Potential Ability?
3. Can we predict which players are in a Top 10 league?

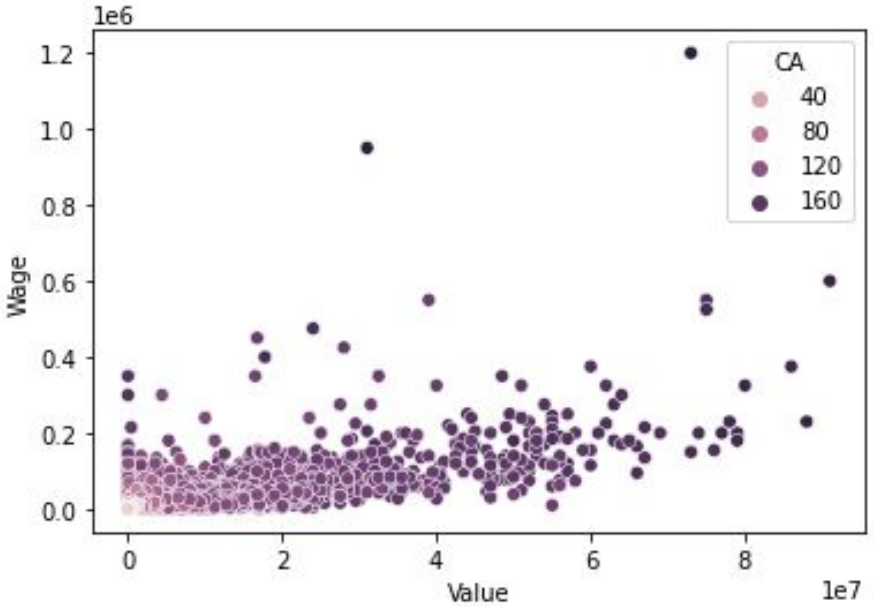
Summary Statistics



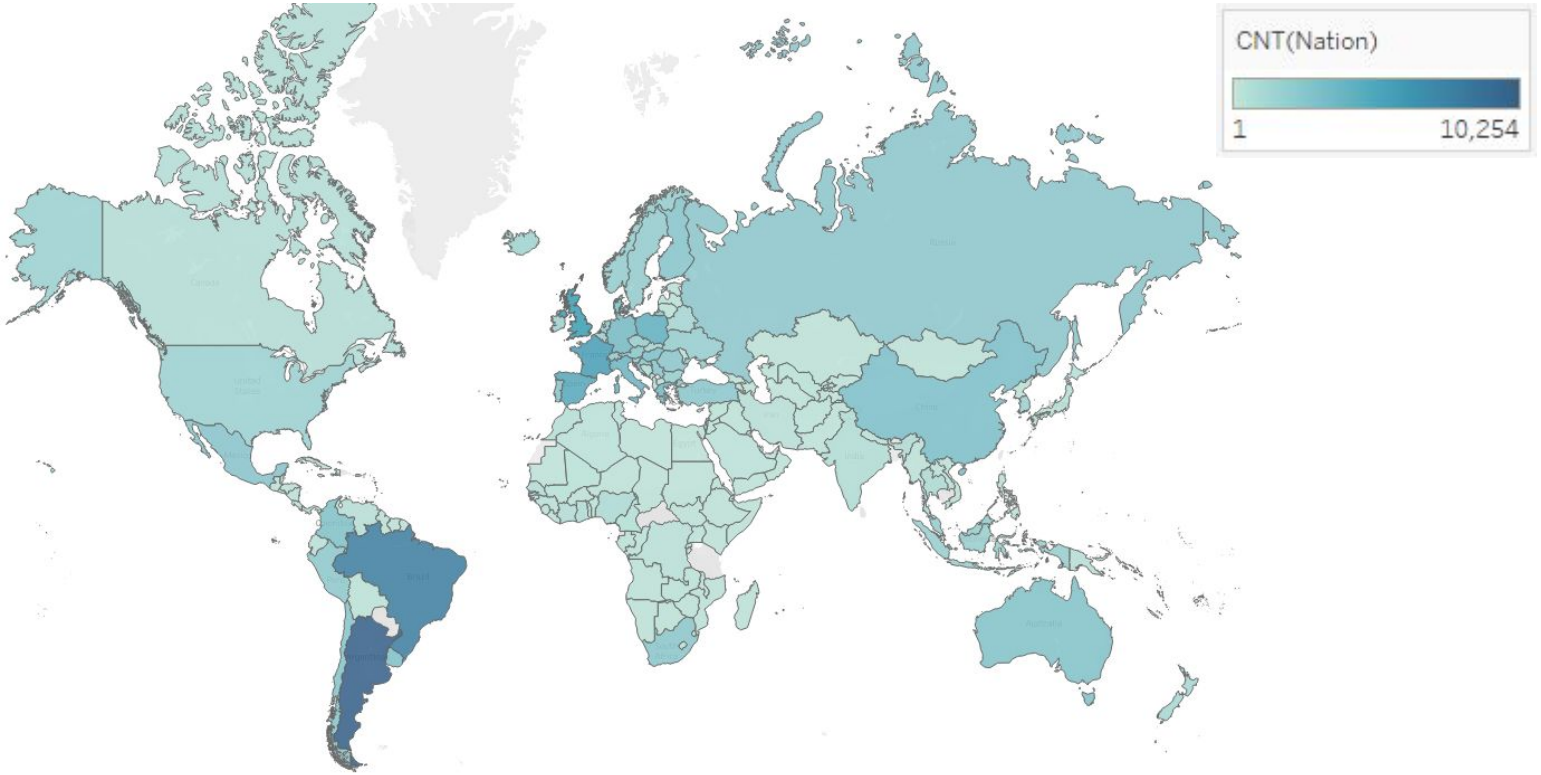


Current Ability, Potential Ability, and Value





Number of Players per Country



Question 1

Can we predict a player's Current Ability (CA)?

- Used 6 regression Models: MLP, Random Forest, K Nearest Neighbors, Decision Tree, and SVR
- The two best models came out to be the Random Forest and MLP models

Evaluations	MLP Regressor	Random Forest Regressor
RMSE	5.0661	6.6949
MSE	25.6650	44.8221
R-squared	0.9628	0.9899
Avg Predicted Difference from actual	4.0952	4.9672

Question 2

Can we predict a player's Potential Ability (PA)?

- Used 6 regression Models: MLP, Random Forest, K Nearest Neighbors, Decision Tree, and SVR
- The two best models came out to be the Random Forest and MLP models

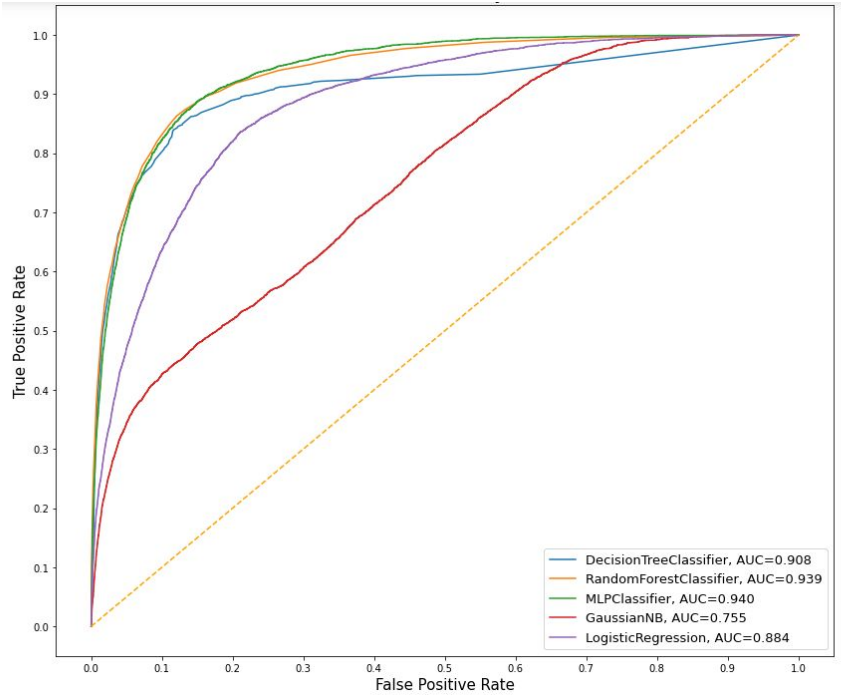
Evaluations	MLP Regressor	Random Forest Regressor
RMSE	11.1589	11.6928
MSE	124.5214	136.7206
R-squared	0.8098	0.9690
Avg Predicted Difference from actual	8.5016	8.9136

Question 3

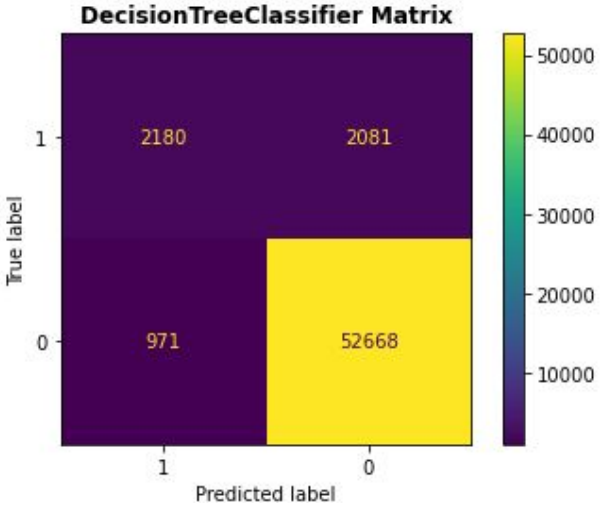
Can we predict which players are in a Top 10 league?

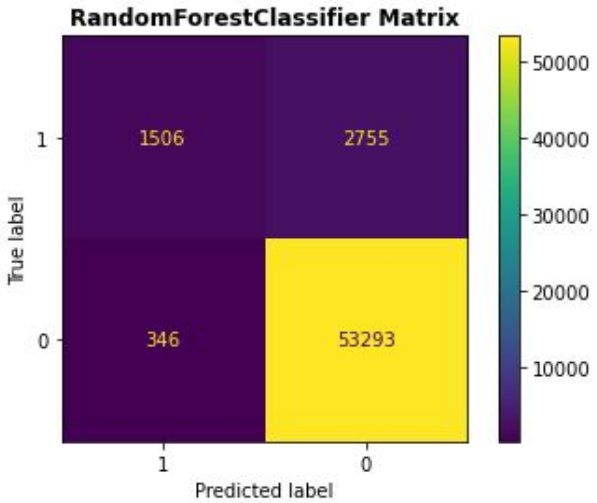
- Solution: We predicted which players are in a Top 10 League by using five classification models.

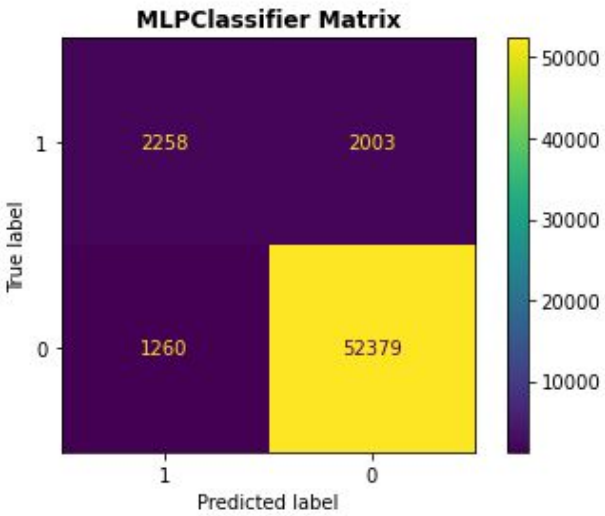
- The MLP Classifier had the best AUC, Accuracy, Log Loss
- The Random Forest Classifier had the fewest false predictions
- The Decision Tree had average results



Classification Matrices







Statistic Evaluations

---- statistics for DecisionTreeClassifier ----

	precision	recall	f1-score	support
0	0.96	0.98	0.97	53639
1	0.69	0.51	0.59	4261
accuracy			0.95	57900
macro avg	0.83	0.75	0.78	57900
weighted avg	0.94	0.95	0.94	57900

Model log loss: 0.3762755894684658

---- statistics for RandomForestClassifier ----

	precision	recall	f1-score	support
0	0.95	0.99	0.97	53639
1	0.81	0.35	0.49	4261
accuracy			0.95	57900
macro avg	0.88	0.67	0.73	57900
weighted avg	0.94	0.95	0.94	57900

Model log loss: 0.15232815295472713

---- statistics for MLPClassifier ----

	precision	recall	f1-score	support
0	0.96	0.98	0.97	53639
1	0.64	0.53	0.58	4261
accuracy			0.94	57900
macro avg	0.80	0.75	0.78	57900
weighted avg	0.94	0.94	0.94	57900

Model log loss: 0.14463179575678917

Dataset Limitations

- Like other sports data, this dataset is likely to be incomplete to some degree. There will be errors/omissions and it will not encompass each goal or kick every player has made

- This data set also does not include statistics about players' likelihood of injuries. This likely would have had a tangible effect on players' long-term Potential Ability

Future Work

- Overall we are content with the outcomes of predicting Potential and current ability from our Regression models

- We believe that we can adjust the classification models to better predict if a player is in a top 10 league

Conclusions

- We can fairly accurately predict a player’s current ability
- Predicting a player’s potential we are not extremely accurate
- Our Classification model to predict if a player is in a top 10 league needs some work, most the accuracy comes from predicting a player is not in a top 10 league

Additional Problem - Digit Recognizer

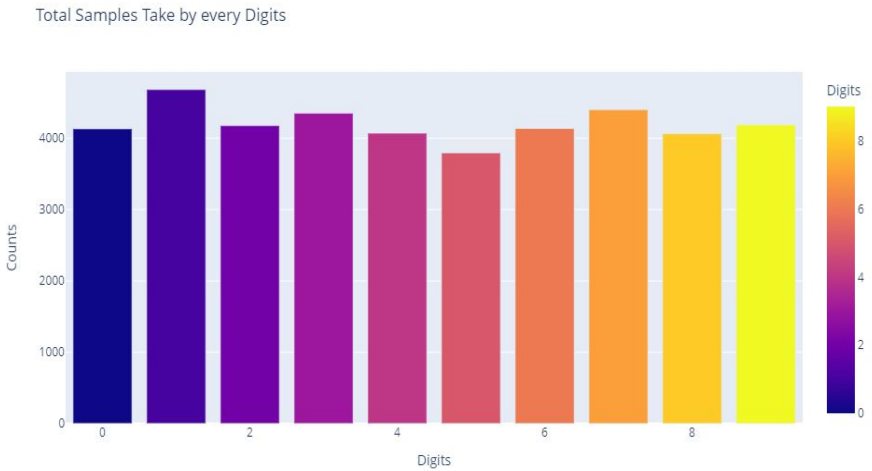
DataSet

digit-recognizer (Kaggle)

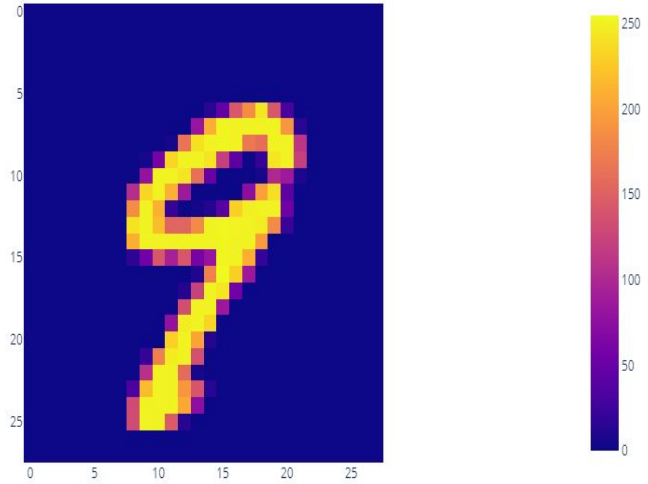
- This dataset contain pixel values of gray-scale images (28x28) of hand-drawn digits, from zero through nine.
- Size: 128MB, 785 Columns, 42,000 Rows
- Key Columns: All

Business Question: Can we recognize hand drawn digits?

Basic Stats



Frequency of digits



Visuals of digit 9

Solution

We used three classification models for predictions. SVM, Neural Network, and Random Forest.

- Random Forest Classifier had the best AUC, Accuracy, F1 score.

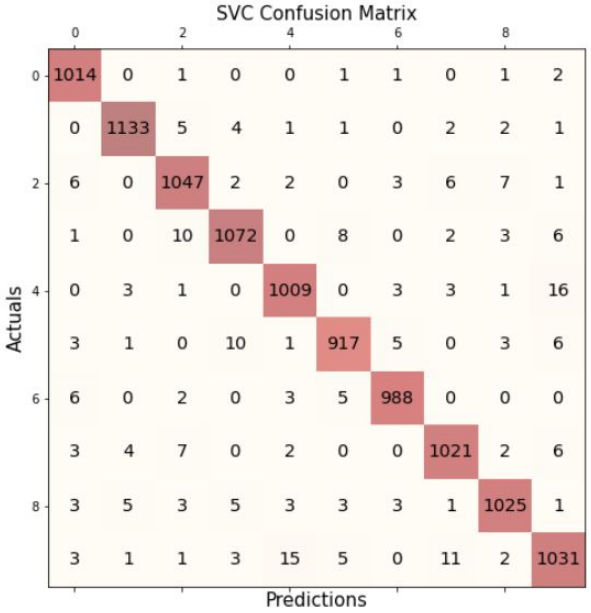
SVC_AUC: 0.99961
RF_AUC: 0.99982
MLP_AUC: 0.99875

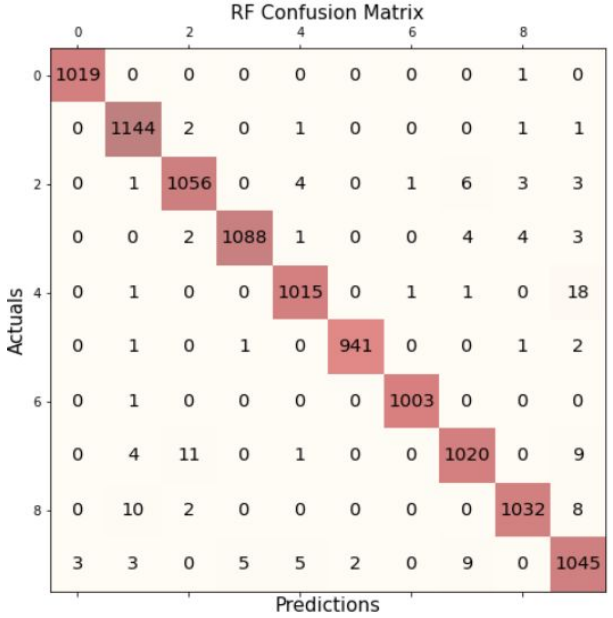
Statistic Evaluations

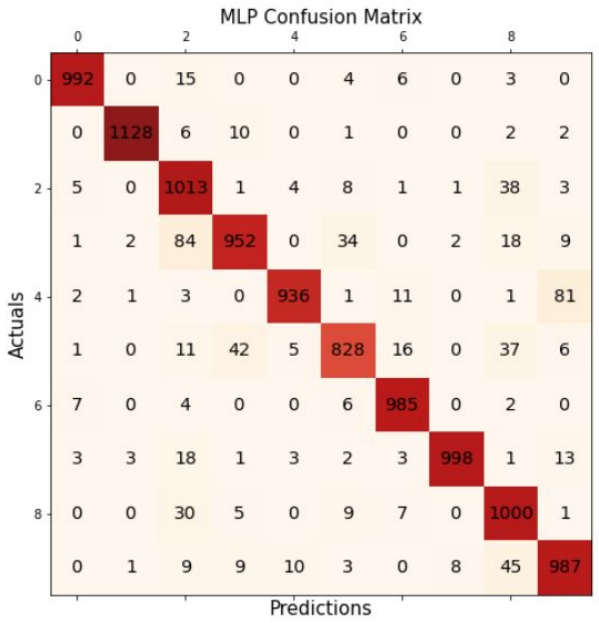
```
        ---- Stats for SVM classifier ----
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1043
                ... values omited ...
           9       0.96      0.97      0.97      1038

    accuracy                           0.98     10500
   macro avg       0.98      0.98      0.98     10500
weighted avg       0.98      0.98      0.98     10500
--------------------------------------------------------

        ---- Stats for RandomForest classifier ----
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1043
                ... values omited ...
           9       0.96      0.98      0.97      1038

    accuracy                           0.99     10500
   macro avg       0.99      0.99      0.99     10500
weighted avg       0.99      0.99      0.99     10500
```

```
        ---- Stats for MLP classifier ----
              precision    recall  f1-score   support

           0       1.00      0.96      0.98      1043
           1       0.98      0.99      0.98      1159
           2       0.96      0.97      0.96      1049
           3       0.94      0.97      0.95      1099
           4       0.96      0.97      0.97       990
           5       0.99      0.96      0.97       955
           6       0.98      0.98      0.98      1058
           7       1.00      0.95      0.97      1112
           8       0.92      0.99      0.95       997
           9       0.96      0.92      0.94      1038

    accuracy                           0.97     10500
   macro avg       0.97      0.97      0.97     10500
weighted avg       0.97      0.97      0.97     10500
```

Confusion Matrices







Limitations

- Provided test data does not contain actual values.
- Predictions using images not supported.

Future Work

- Implement methods to use images rather than csv data for predictions.

Conclusion

- Used models are fairly good at recognizing digits from grey-scale images.

Thank You!