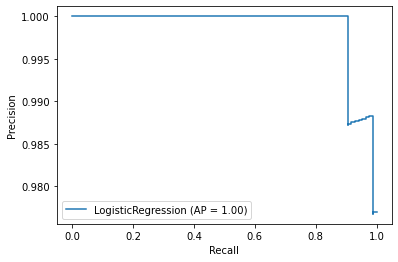
4sl3 lab 2

David Wang

400073796

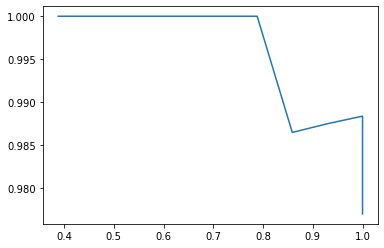
Part 1

**Scikit Logistic Regression**



Above is the PR graph for our scikit logistic regression model

**Manual logistic regression**



Above is the PR graph for out Manual logistic regression model, vector of parameters in range of -20 to 100 was found through trial and error to produce said graph

**KNN neighbors manual**

K selected was 5.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Manual KNN | | | | |  |
| K Fold | | | | |  |
| Nearest Neighbors | 1 | 2 | 3 | 4 | 5 | average |
| 1 | 0.087719 | 0.070175 | 0.061404 | 0.131579 | 0.106195 | 0.091414 |
| 2 | 0.087719 | 0.061404 | 0.035088 | 0.131579 | 0.097345 | 0.082627 |
| 3 | 0.087719 | 0.105263 | 0.087719 | 0.061404 | 0.115044 | 0.09143 |
| 4 | 0.096491 | 0.078947 | 0.078947 | 0.087719 | 0.061947 | 0.08081 |
| 5 | 0.087719 | 0.149123 | 0.052632 | 0.157895 | 0.088496 | 0.107173 |

The selected classifier in this case was 4 nearest neighbors.

**KNN neighbors scikit**

K selected was 5

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Scikit KNN | | | | |  |
| K Fold | | | | |  |
| Nearest Neighbors | 1 | 2 | 3 | 4 | 5 | average |
| 1 | 0.087719 | 0.070175 | 0.061404 | 0.131579 | 0.106195 | 0.091414 |
| 2 | 0.087719 | 0.070175 | 0.052632 | 0.140351 | 0.097345 | 0.089644 |
| 3 | 0.070175 | 0.070175 | 0.04386 | 0.078947 | 0.079646 | 0.068561 |
| 4 | 0.087719 | 0.078947 | 0.061404 | 0.070175 | 0.070796 | 0.073808 |
| 5 | 0.070175 | 0.061404 | 0.035088 | 0.078947 | 0.088496 | 0.066822 |

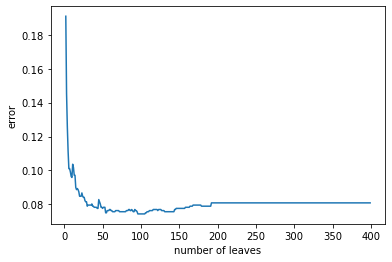
The selected classifier in this case was 5 nearest neighbors.

**Analysis of all 4 models**

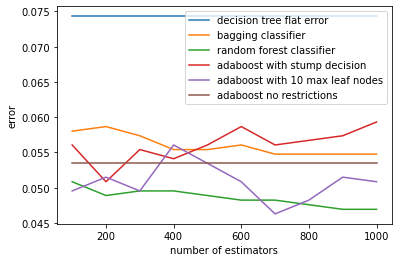
|  |  |  |
| --- | --- | --- |
|  | Error | F1 Score |
| Manual Logistic | 0.006993 | 0.994152 |
| sci-kit logistic | 0.013986 | 0.988235 |
| Manual KNN | 0.06993 | 0.939759 |
| sci-kit KNN | 0.013986 | 0.988235 |

The manual logistic regression model has the best accuracy and F1 score based on the chart above. The manual logistic regression model seems to have perform better than the scikit given one. However the KNN neighbor manual model seem to lag behind in performance when compared to its scikit version.

Part 2



The graph above displays the accuracy of the decision tree classifier based on leaves. The lowest region is between 50 to 110 leaves and then the error starts to grow a little bit again.



The best classifier based on the data presented is the random forest classifier. It has the lowest error at almost all point of the graph on all number of estimators. A close second is the adaboost with 10 max leaf nodes. Its error actually is the lowest at the 700 estimator interval but becomes less efficient at 400 estimators. Interestingly enough,, the adaboost with no restrictions is the third best option in this case, however it has the same error at all estimators and is the only classifier that has this property. The bagging classifier and the adaboost with stump decision are tied for 4th and 5th place. The adaboost seems to perform better with less estimators and its error goes up after 200 estimators, while the bagging classifier seems to do better with additional classifiers. The two intersect at around the 500 estimator point and that’s when the bagging classifier overtakes the adaboost with stump decision classifier. The decision tree classifier is the worst classifier out of all the rest with the highest average of error by a relative margin.