Week 03 Report

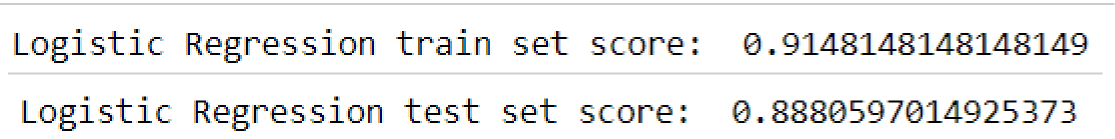
Jeffrey Li

Previous Procedure

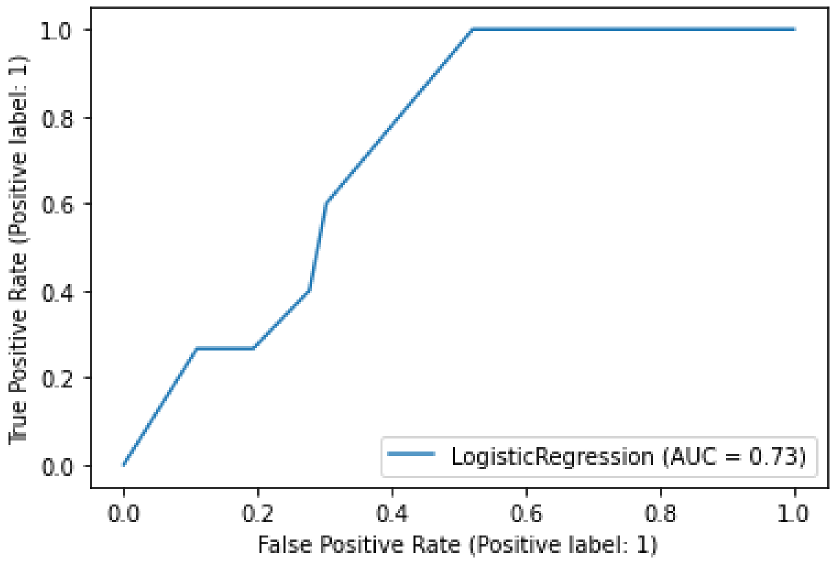
- 404 total data instances
  - 366 small chi, 38 large chi
- Preprocessing:
  - Binned features into nominal, categorical labels.
  - Created binary labels for chi (0 = small chi, 1 = large chi).
- Classification Algorithms Used: Logistic Regression, Naïve Bayes, Random Forest, AdaBoost.
- Split Training and Testing using 67:33 ratio.

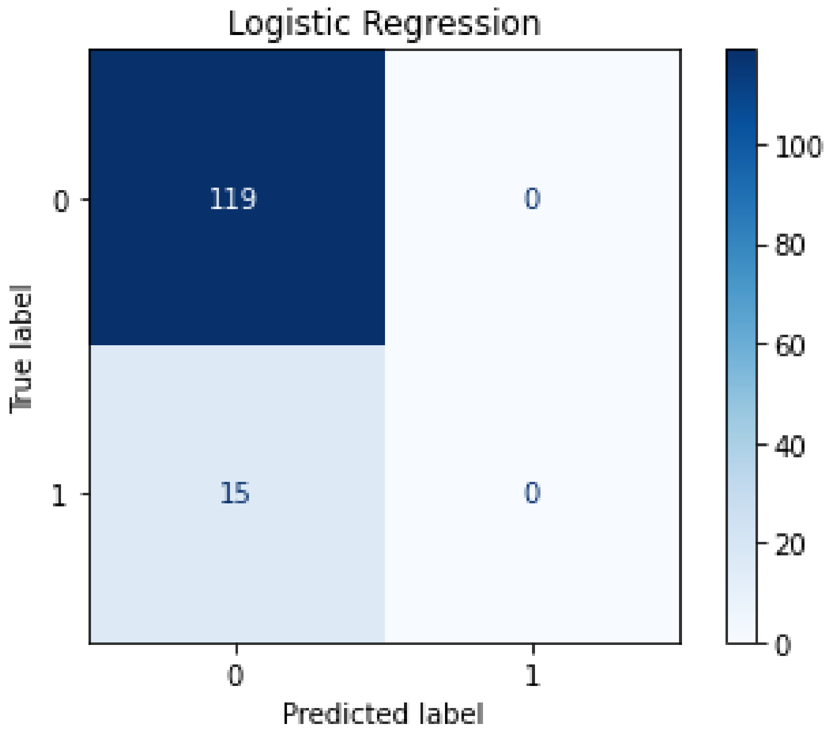
Previous Results

- High training and high testing score due to high True Negative score.
  - Same reason for Area Under the Curve (AUC) score.
- Only True Negative and False Negative scores reported. The models are only predicting small chi labels and not large chi labels.
- Results most likely caused by imbalance in target labels.



```
Logistic Regression train set score:  0.9148148148148149
Logistic Regression test set score:  0.8880597014925373
```



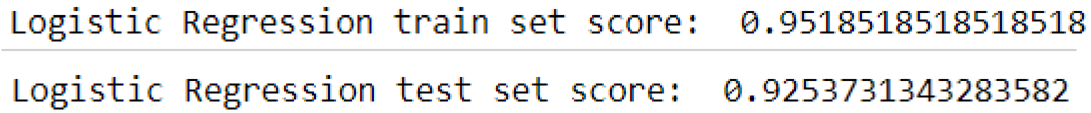


Updates in Preprocessing

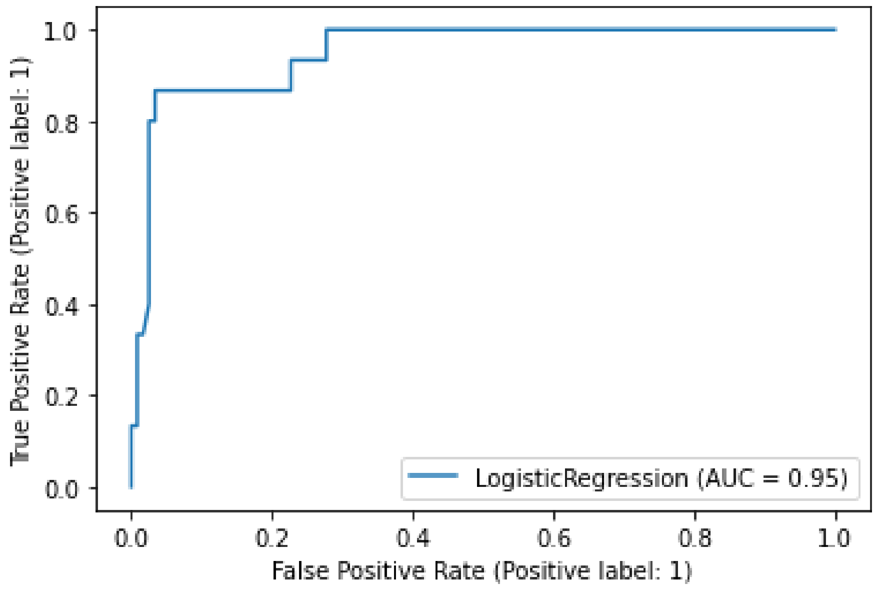
1. Removed Binning
2. Added New Feature
3. 50/50 Split of Small Chi Data
4. Feature Selection

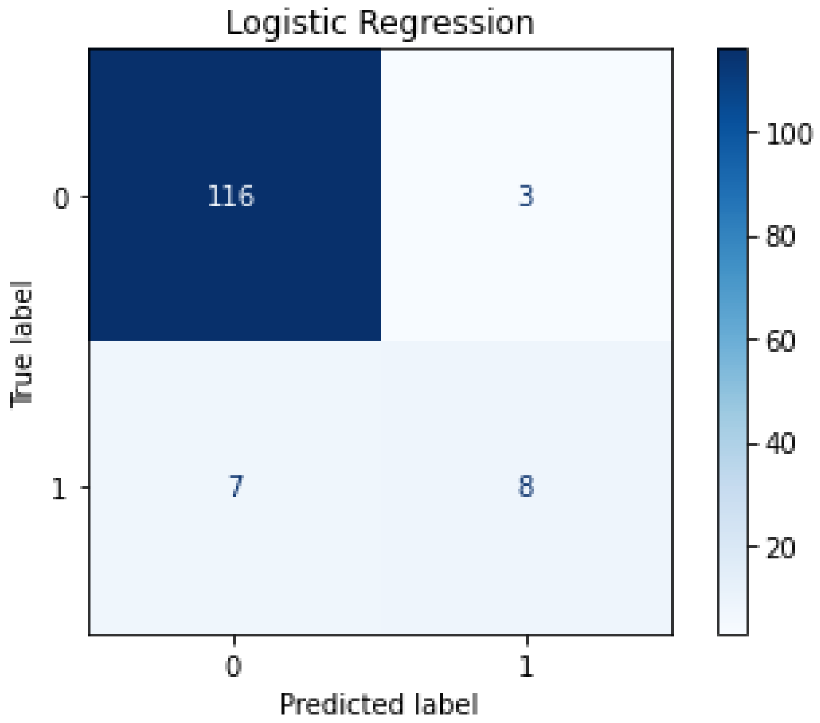
Removed Binning

- Converting continuous data to categorical data removed a lot of information.
- Kept features as continuous but maintained binary labels for target.
- High AUC score.
  - Good scores for True Negative and True Positive means the model attempts to predict small and large chi (w/ some error).
- Potential Issue:
  - Model may benefit from being trained/tested with more data or noise.



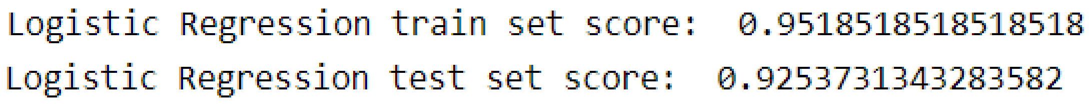
```
Logistic Regression train set score:  0.9518518518518518

Logistic Regression test set score:  0.9253731343283582
```



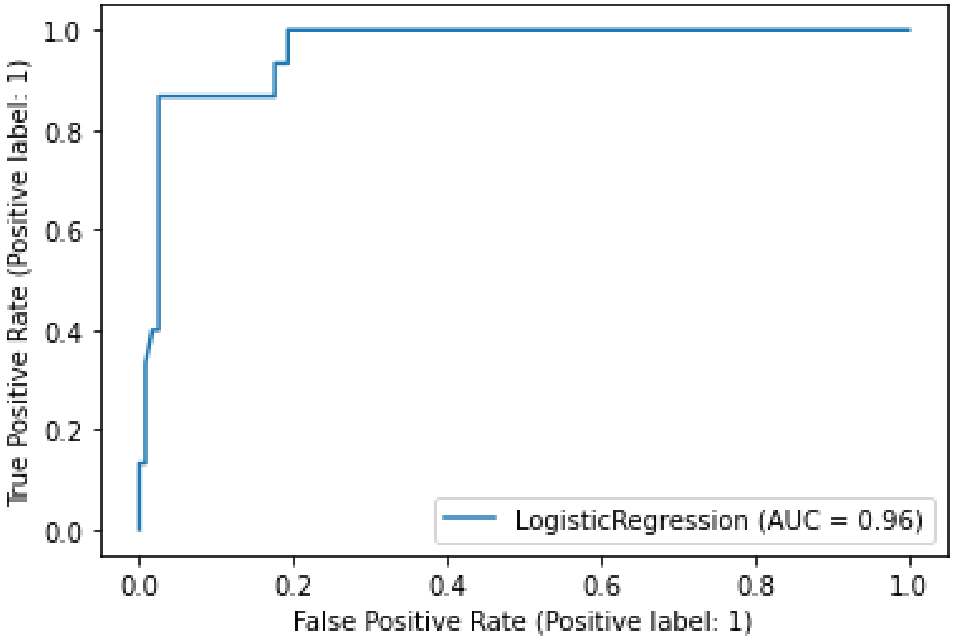


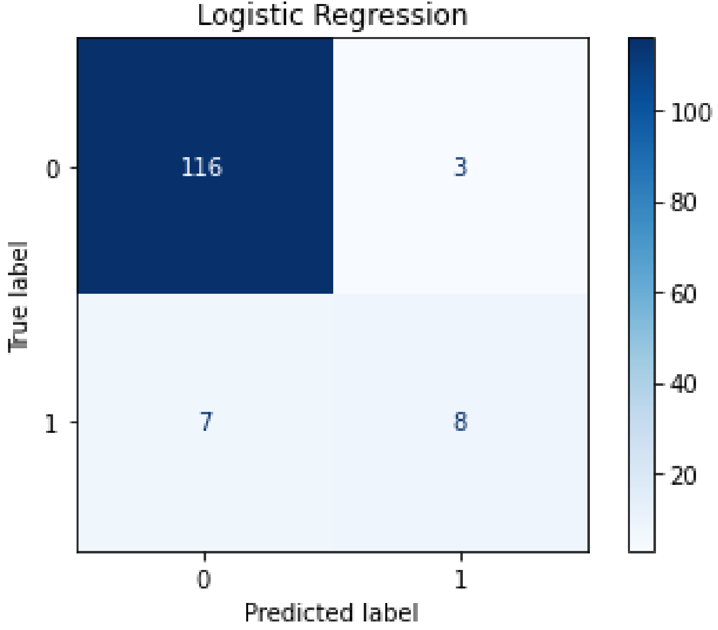
Added New Feature

- Adding new features may speed up data transformation and ultimately model accuracy.
- Kept features as continuous but maintained binary labels for target.
- Created new feature ‘diff’ from determining the **absolute difference between Alph1 and Alph2**.
  - Kept Alph1 and Alph2 as features.
  - Total of 5 features.
- High AUC score.
  - Good scores for True Negative and True Positive means the model attempts to predict small and large chi (w/ some error).
- Potential Issue:
  - Model may benefit from being trained/tested with more data or noise.



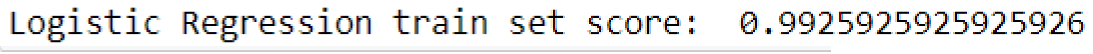
```
Logistic Regression train set score:  0.9518518518518518
Logistic Regression test set score:  0.9253731343283582
```



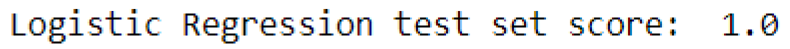


Added New Feature (cont.)

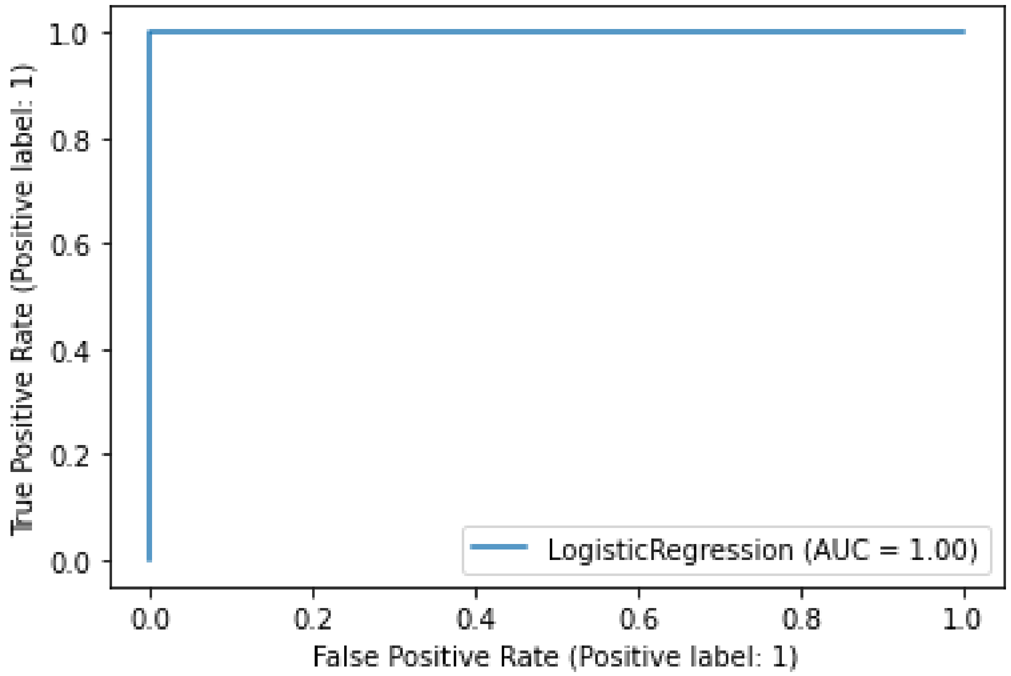
- Kept the new feature ‘diff’ but **removed Alph1 and Alph2 as features**, since they may be redundant.
  - Total of 3 features.
- Kept features as continuous but maintained binary labels for target.
- Perfect AUC score.
  - The model perfectly predicts small chi and large chi with no errors.
- Potential Issue:
  - Model may benefit from being further trained/tested with more data or noise.

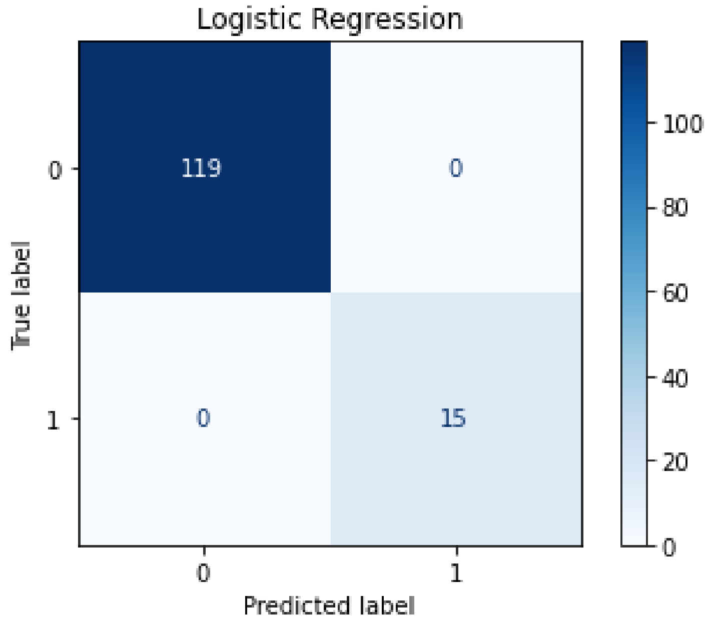


```
Logistic Regression train set score:  0.9925925925925926
```



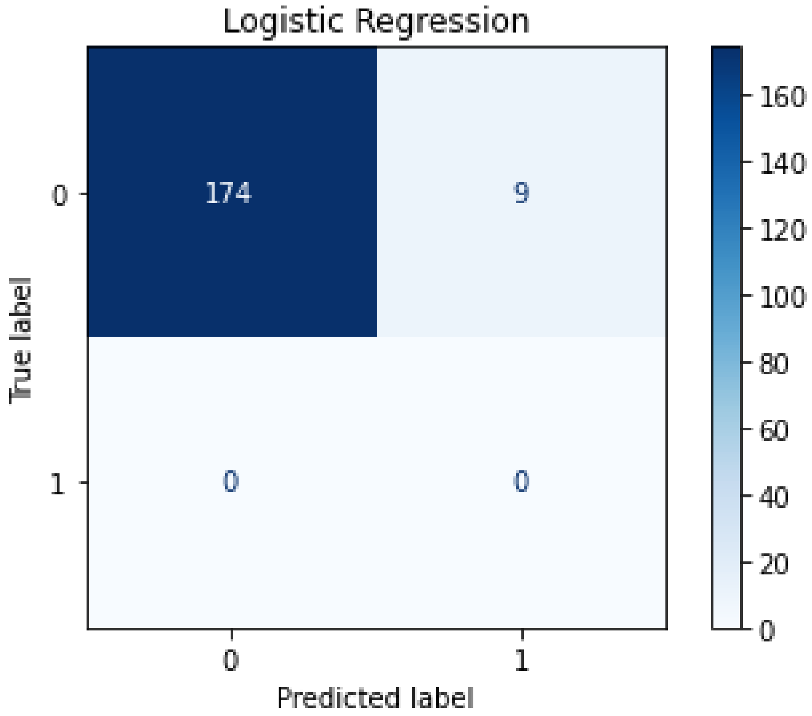
```
Logistic Regression test set score:  1.0
```

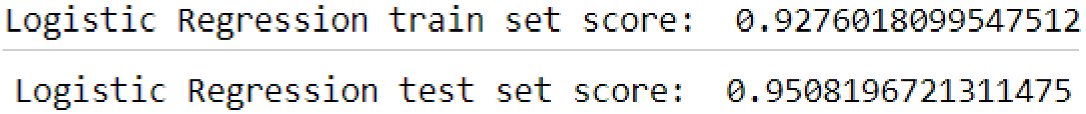




Random Split (50:50) of Small Chi Data

- We want to see if the test data gets assigned to small chi with no errors.
  - Training data consists of 183 random instances of small chi, all (38) instances of large chi.
  - Testing data consists of the remaining 183 random instances of small chi.
- Kept features as continuous but maintained binary labels for target.
  - Kept diff, alph1, and alph2 as features.
- No AUC curve because there is no False Positive.
- Confusion matrix shows 174 instances were correctly identified as small chi (True Negative) and 9 were incorrectly identified as 'large chi (False Positive).
- Potential Issue:
  - Model may benefit from a different random split ratio.

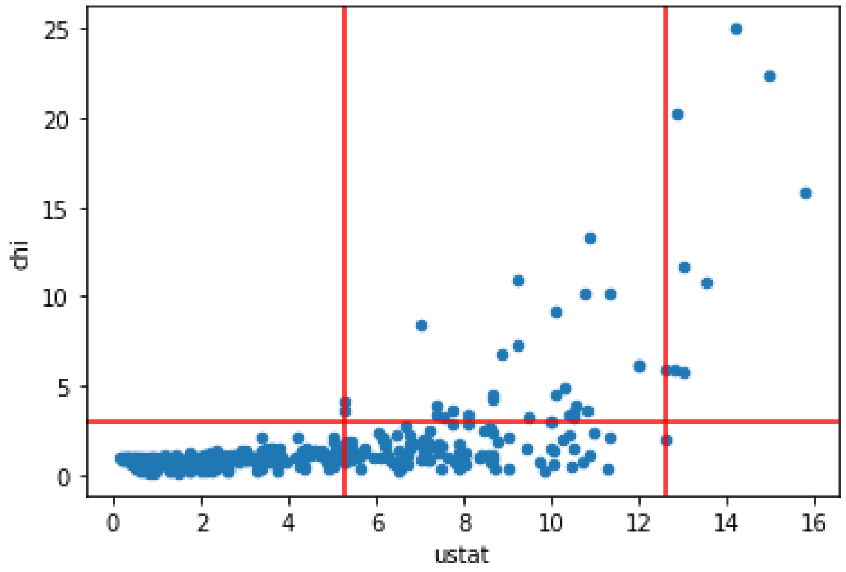


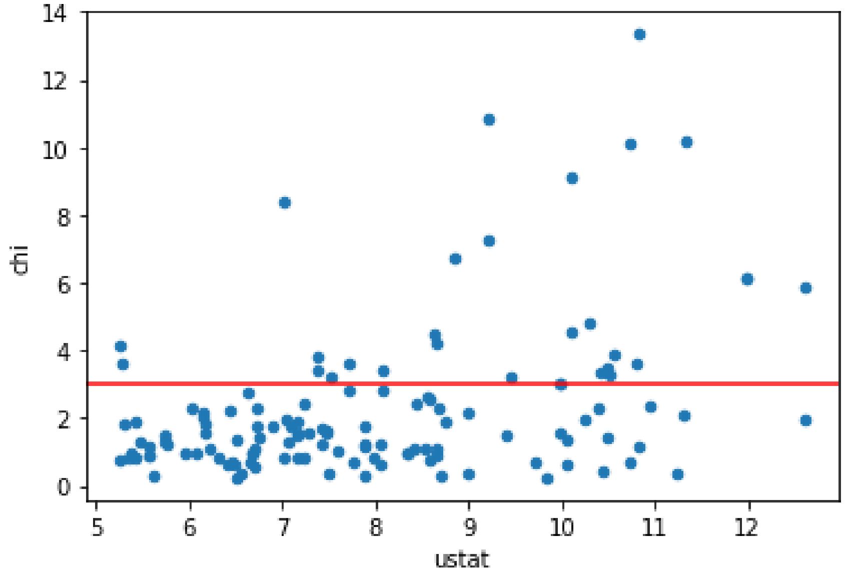


```
Logistic Regression train set score:  0.9276018099547512
Logistic Regression test set score:  0.9508196721311475
```

Feature Selection

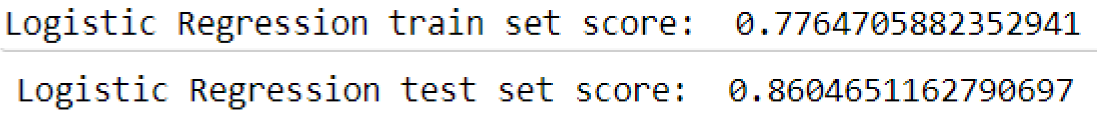
- Process only data instances that lie between the overlapping scatter area seen on the top right.
    - Min ustat value for large chi: 5.26.
    - Max ustat value for small chi: 12.6.
- Only data instances containing ustat value **between 5.26 and 12.6** are considered.
    - See Bottom Right.
    - 128 data instances: 98 small chi and 30 large chi.



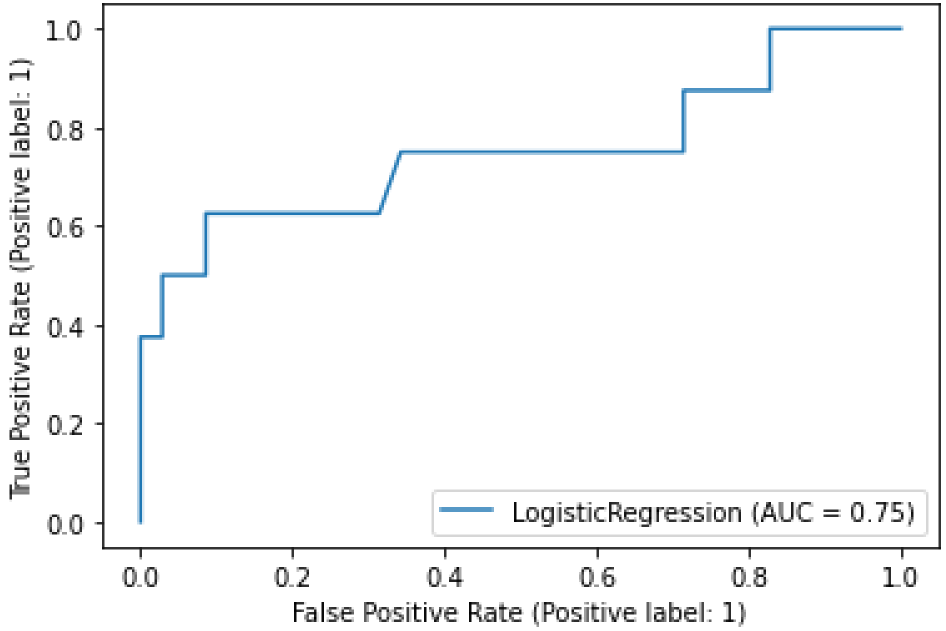


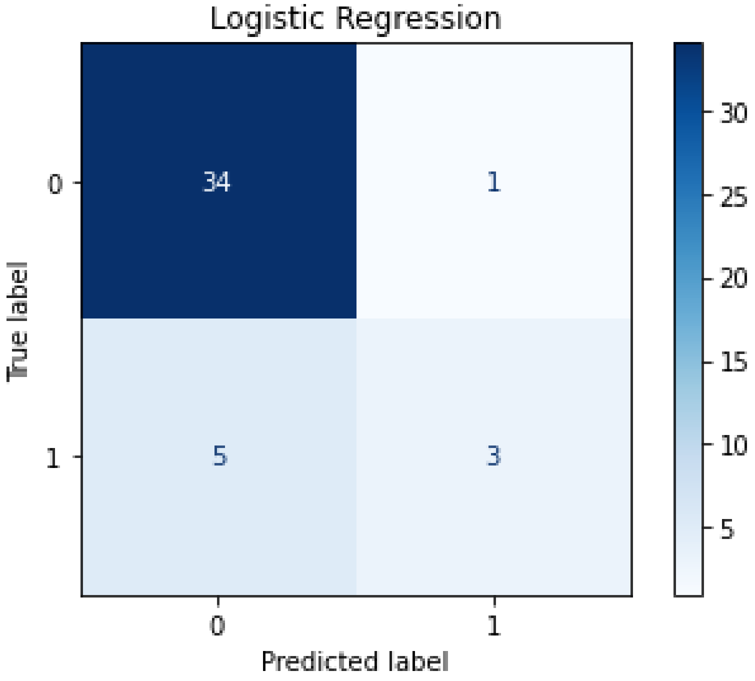
Feature Selection (cont.)

- Kept features as continuous but maintained binary labels for target.
  - Kept diff, alph1, and alph2 as features.
- High testing accuracy and good AUC score.
  - Model predicts both small chi and large chi (w/ errors).



```
Logistic Regression train set score:  0.7764705882352941
Logistic Regression test set score:  0.8604651162790697
```


True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
LogisticRegression (AUC = 0.75)


Logistic Regression
True label
Predicted label