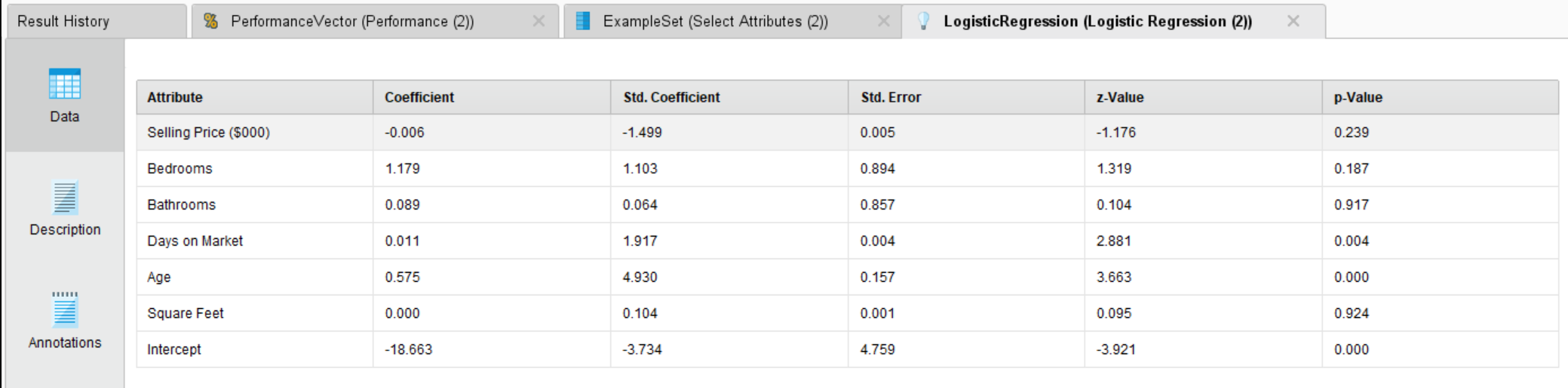
**ISM6136-Assignment 3 – Classification**

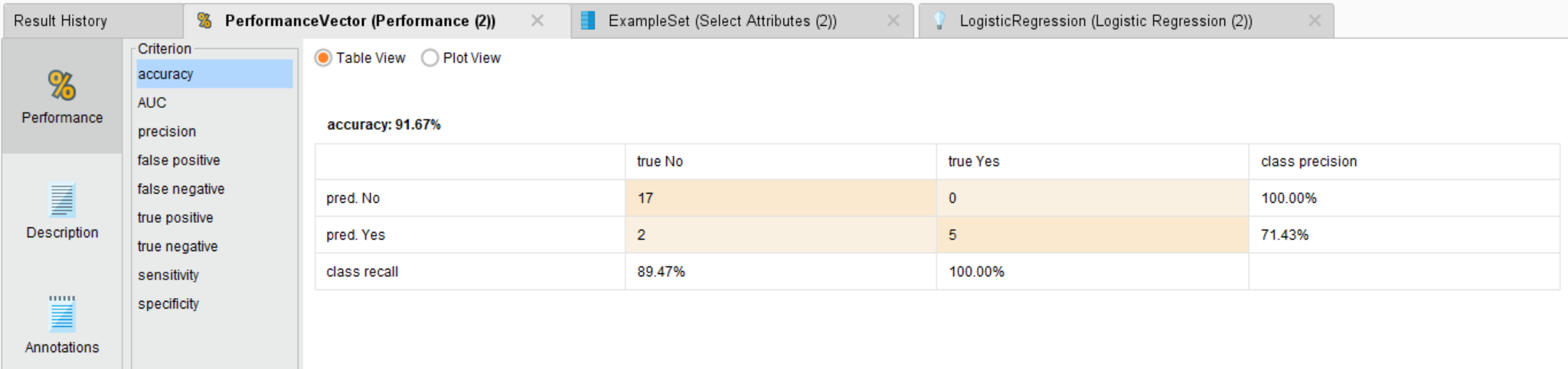
In order to predict whether a house will be in foreclosure or not, I selected Selling Price ($000), Bedrooms, Bathrooms, Days on Market, Age and Square Feet as my attributes. My target variable is foreclosure.

Because outcome variable is categorical (yes and no), I have to change that to binary classification (i.e. y=0 or y=1).

My first predictive model is using the logistic regression to find the association between my attributes and the probability of foreclosure.



The table above shows that there are associations between days on market as well as age and the probability of foreclosure based on the p-value.

The Evaluating Classification Performance Confusion Matrix table is showed below:

The Performance Vector are listed below:

Accuracy: 91.67%

AUC: 0.916

Precision: 71.43%

False\_positive: 2

False\_negative: 0

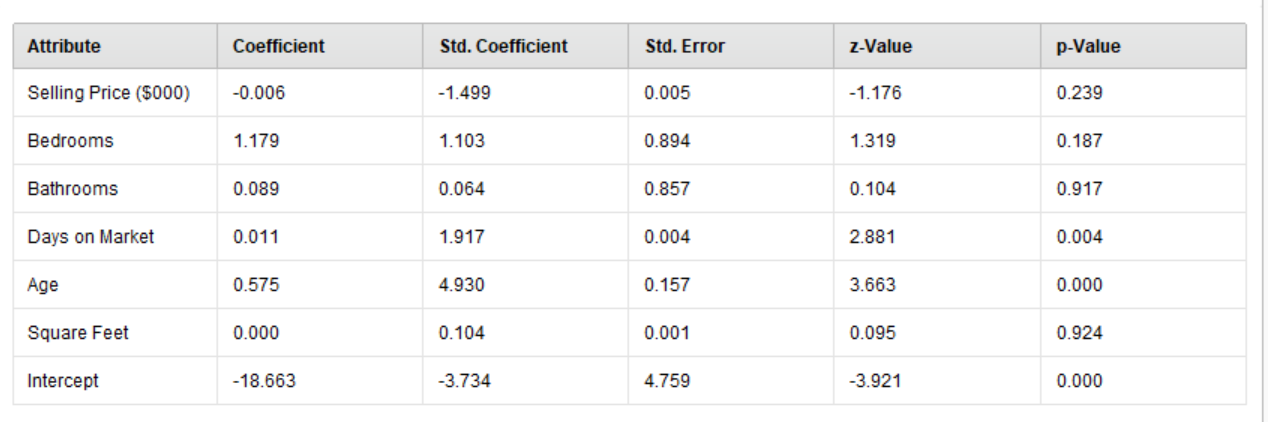
True\_positive: 5

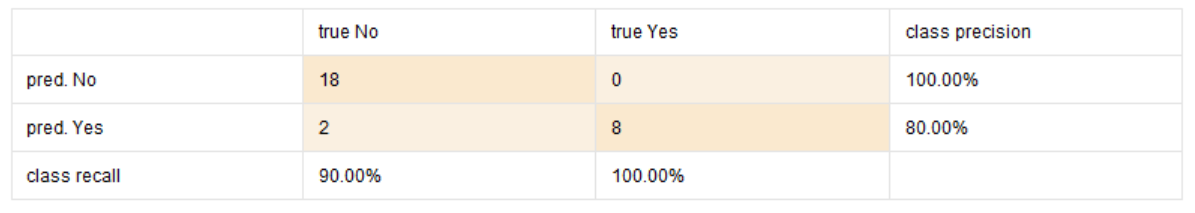
True\_negative: 17

Sensitivity: 100.00%

Specificity: 89.47%

Then I optimized my model by adjusted the split ratio from 0.75 to 0.7 and got results below:





Accuracy: 92.86%

AUC: 0.925

Precision: 80.00%

False\_positive: 2

False\_negative: 0

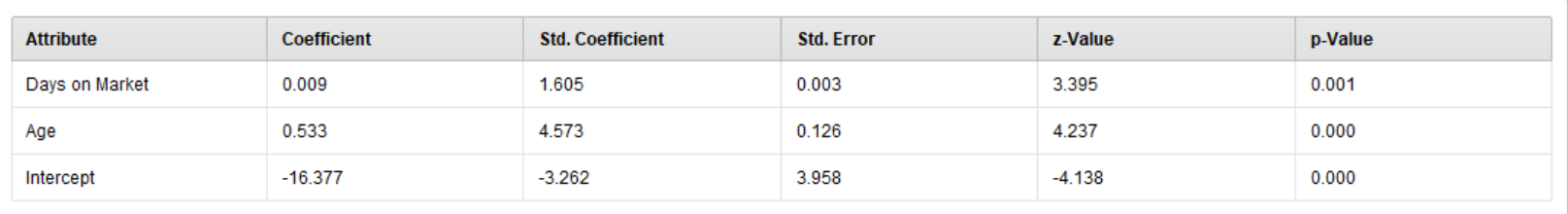
True\_positive: 8

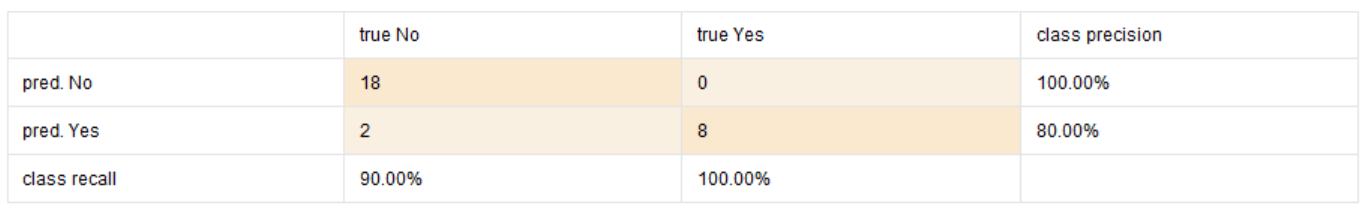
True\_negative: 18

Sensitivity: 100.00%

Specificity: 90.00%

Next, I reperform the logistic regression using attribute Days on Market and Age because based my previous analysis, other attributes are irrelevant. This model improves the AUC to 0.937 which is better than the previous model. The accuracy, sensitivity and specificity remain the same.





Accuracy: 92.86%

AUC: 0.950

Precision: 80.00%

False\_positive: 2

False\_negative: 0

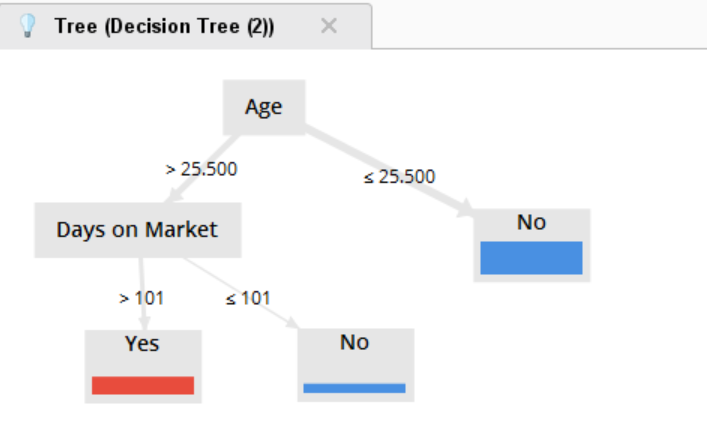
True\_positive: 8

True\_negative: 18

Sensitivity: 100.00%

Specificity: 90.00%

My second predictive model is using the decision tree to find the association between my attributes and the probability of foreclosure. My target variable is still foreclosure. My selected attributes are Selling Price ($000), Bedrooms, Bathrooms, Days on Market, Age and Square Feet.



My decision tree rule model is listed below:

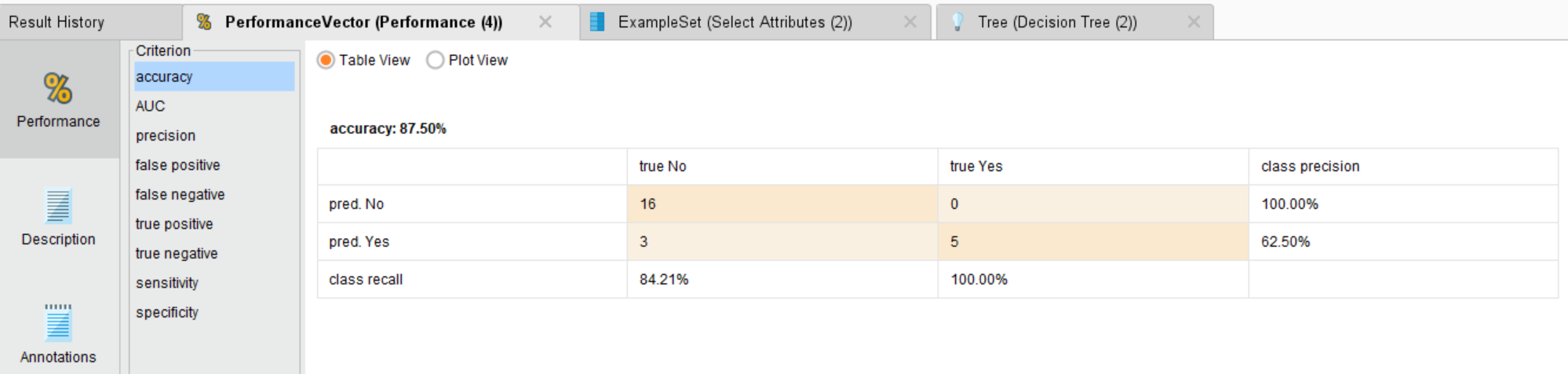
Age > 25.500

| Days on Market > 101: Yes {No=0, Yes=27}

| Days on Market ≤ 101: No {No=13, Yes=0}

Age ≤ 25.500: No {No=55, Yes=0}

And below is its Evaluating Classification Performance Confusion Matrix table:



The Performance Vector are listed below:

Accuracy: 87.50%

AUC: 0.500

Precision: 62.50%

False\_positive: 3

False\_negative: 0

True\_positive: 5

True\_negative: 16

Sensitivity: 100.00%

Specificity: 84.21%

By comparing the Logistic Regression method and the Decision Tree method in the chart below, the Logistic Regression method has a better performance on predicting the possibility of foreclosure.

|  |  |  |
| --- | --- | --- |
|  | Logistic Regression | Decision Tree |
| Accuracy | 91.67% | 87.50% |
| AUC | 0.937 | 0.500 |
| Precision | 71.43% | 62.50% |
| Sensitivity | 100.00% | 100.00% |
| Specificity | 89.47% | 84.21% |