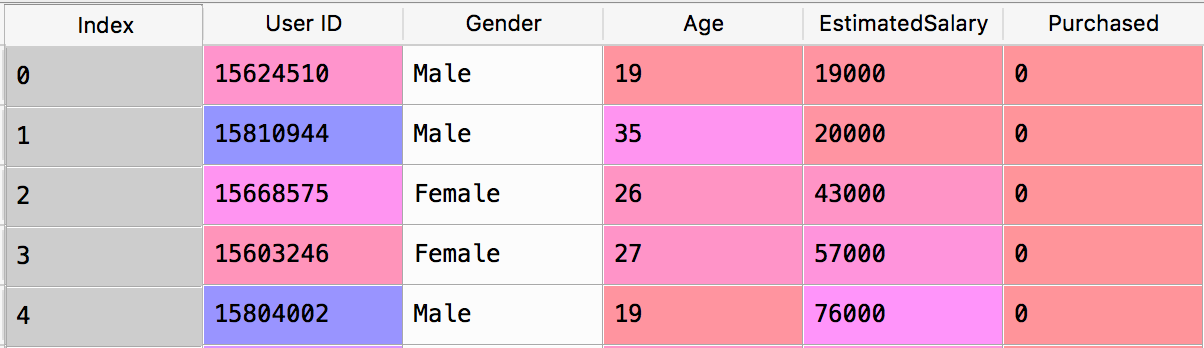
Logistic Regression Findings

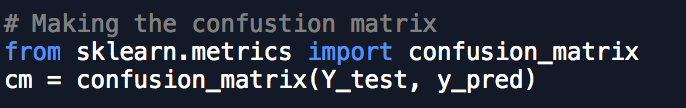
We received a dataset with 400 users showing their gender, age, estimated salary, and whether they purchased a specific vehicle that was on sale (where 1 = yes and 2 = no).

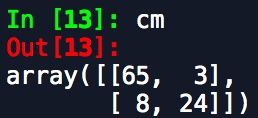
Our model is using *age* and *estimated salary* is going to predict whether the individual purchased the vehicle. Therefore *User ID* and *Gender* are discarded.



We split the data into a training (75%) and test set (25%). Thus 300 rows will be used to train the model, and 100 will be used to test the model.

After running the logistic regression model we used a confusion matrix to see how accurate the predictions were. From the picture below we observe there was 89 correct predictions and 11 incorrect predictions giving us 89% accuracy.

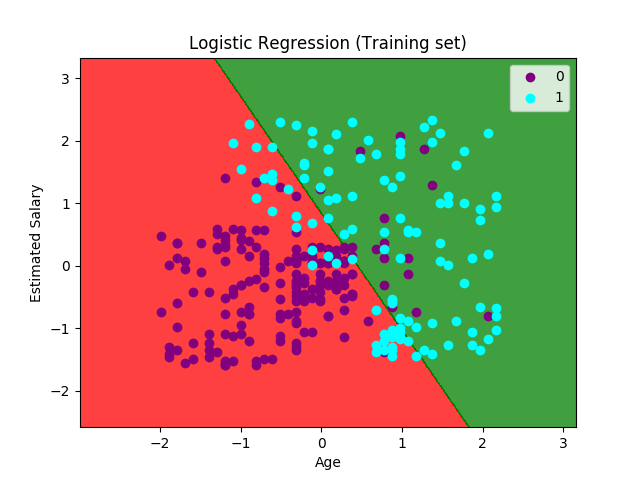




Visualizing the data

The diagram below shows the results of the **training set**.

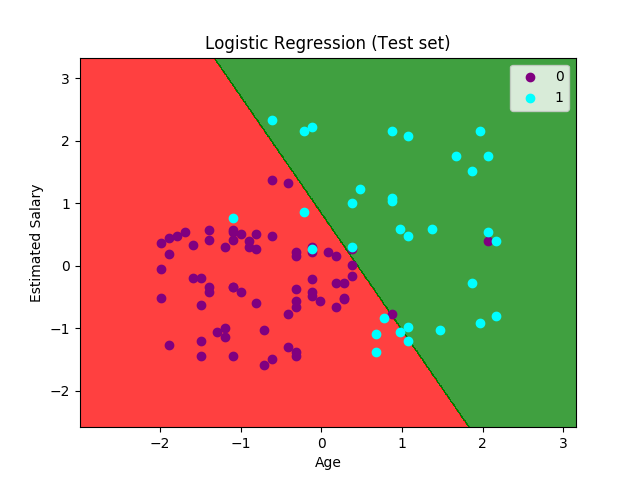
* The points are the individuals in the dataset.
* The *purple* *points* are the training set observations where the dependent variable *purchased* is 0 (didn’t buy the vehicle).
* The *cyan* *points* are the training set observations where the dependent variable *purchased* is 1 (bough the vehicle).
* The points within *red* *region* are the members our classifier will predict who won’t buy the vehicle.
* The points within *green* *region* are the members our classifier will predict who will buy the vehicle.



As you can see from the line dividing the two regions that we are using a linear classifier. Our classifier picked the best possible line in order to learn how to properly classify the data.

The diagram below shows the results of the **test set**.

* The points are the individuals in the dataset.
* The *purple* *points* are the training set observations where the dependent variable *purchased* is 0 (didn’t buy the vehicle).
* The *cyan* *points* are the training set observations where the dependent variable *purchased* is 1 (bough the vehicle).
* The points within *red* *region* are the members our classifier will predict who won’t buy the vehicle.
* The points within *green* *region* are the members our classifier will predict who will buy the vehicle.



From this we can observe

* Youngest members with a low estimated salary didn’t buy the vehicle.
* A few young members with a higher estimated salary bought the vehicle.
* Most older people bought the vehicle.

Conclusion

While the logistic regression model is a good one it is a linear model. A linear model doesn’t seem to be best suited for this type of program. A polynomial regression model might be better suited. We shall look at other classifiers for this in the future.