Logistic Regression Findings

We received a dataset (*Social_Network_Ads.csv*) 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.

Index	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

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.

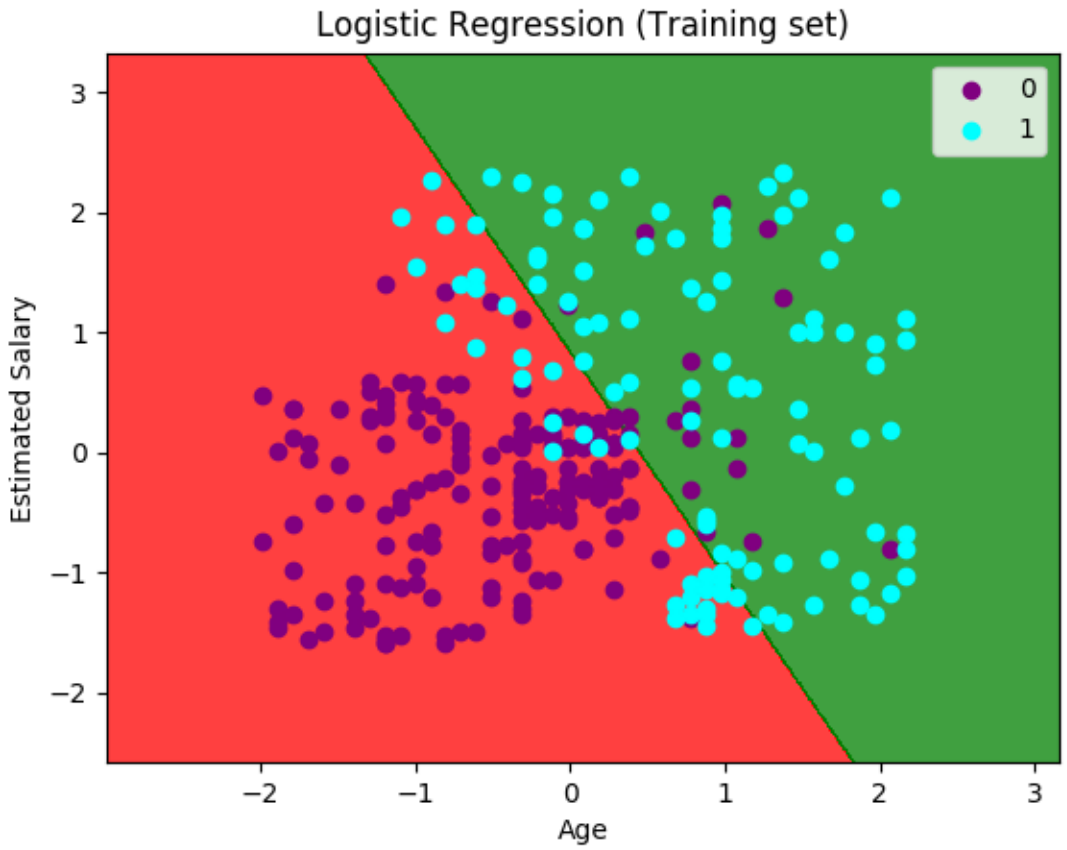
```
# Making the confustion matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test, y_pred)
```

```
In [13]: cm
Out[13]:
array([[65,  3],
       [ 8, 24]])
```

Visualizing the data

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

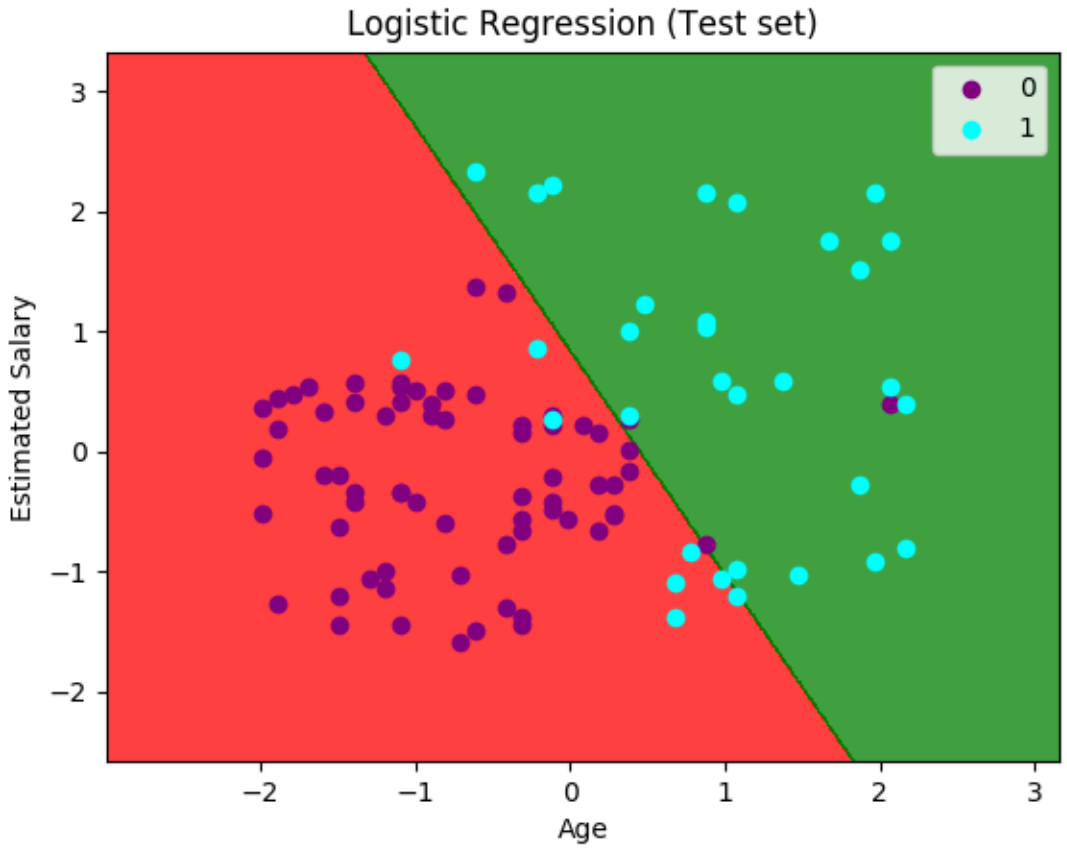
- The points are the individuals in the dataset.
- The *purple points* are the training set observations where the dependent variable *purchased* is 0 (didn't buy the vehicle).
- The *cyan points* are the training set observations where the dependent variable *purchased* is 1 (bough the vehicle).
- The 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 (*see "Classification Findings.pdf")*.