T1. If the starting points are (3,3), (2,2), and (-3,-3). Describe each assign and update step. What are the points assigned? What are the updated centroids? You may do this calculation by hand or write a program to do it.

Text

Description automatically generated with medium confidenceChart, scatter chart

Description automatically generated

From the pictures, we can see that there are two steps in order to fit the model. Points are grouped together likes [(-7, -7), (-2, -4), (-3, -3)] [(1, 2), (2, 2), (3, 3)] [(6, 6), (7, 7), (8, 8)] as we can see in the second plot. The centroids will be shown in the image as well.

T2. If the starting points are (-3,-3), (2,2), and (-7,-7), what happens?

The cluster is different from the first trial as you can see in the following image.

Chart, scatter chart

Description automatically generated

T3. Between the two starting set of points in the previous two questions, which one do you think is better? How would you measure the ‘goodness’ quality of a set of starting points?

In my opinion, the first one, by using the explainable variance, the first one accounts for 0.929 while the second one’s is 0.814. With this metric, I believe that we should consider using the first case as the starting point for centroids.

Graphical user interface, text, application, email

Description automatically generated

T4. What is the median age of the training set? You can easily modify the age in the dataframe by train["Age"] = train["Age"].fillna(train["Age"].median())

28

Text

Description automatically generated

Note that you need to modify the code above a bit to fill with mode() because mode() returns a series rather than a single value.

T5. Some fields like ‘Embarked’ are categorical. They need to be converted to numbers first. We will represent S with 0, C with 1, and Q with 2. What is the mode of Embarked? Fill the missing values with the mode. You can set the value of Embarked easily with the following command. train.loc[train["Embarked"] == "S", "Embarked"] = 0 Do the same for Sex.

0 from the above image.

T6. Write a logistic regression classifier using gradient descent as learned in class. Use PClass, Sex, Age, and Embarked as input features.

Text

Description automatically generated with medium confidence

T7. Submit a screenshot of your submission (with the scores). Upload your code to courseville.

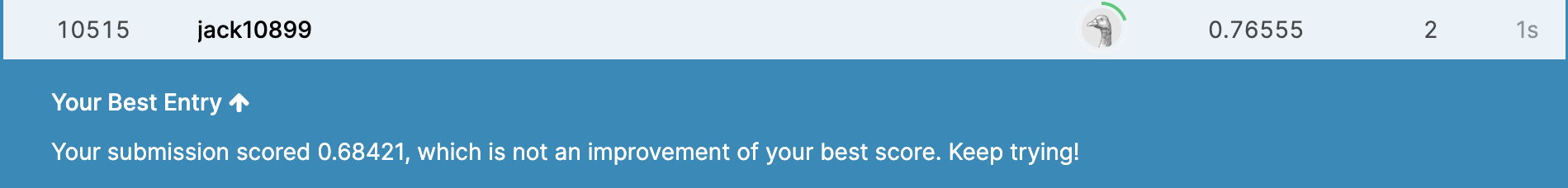
Graphical user interface, text

Description automatically generated

T8. Try adding some higher order features to your training (x 2 1 , x1x2,...). Does this model has better accuracy on the training set? How does it perform on the test set?

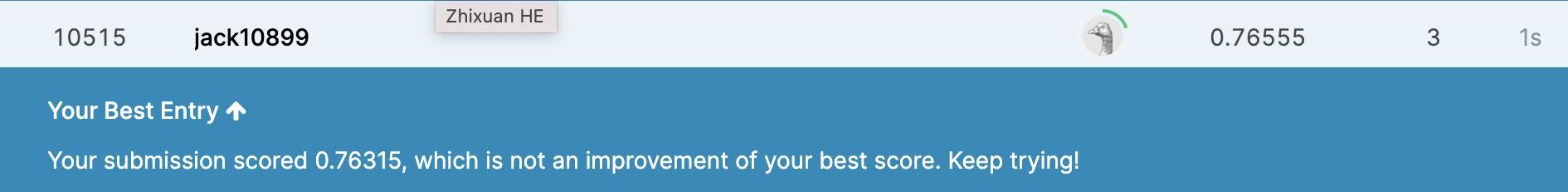
A picture containing chart

Description automatically generated



I think it worse than not doing anything.

T9. What happens if you reduce the amount of features to just Sex and Age?



Still lost to the first attempt, but after dropping two features, the score is very close to the first one. This is interesting because the other two features might not be that necessary when predict if the one is survived from Titanic.