Ericsson Group Data Scientist Hiring Test - Santa Clara

34m
to test end

☆ Machine Learning 1

You are working in a Data Science team that has been assigned the task to work on customer churning for a telecom operator. You are provided a dataset of 100,000 records. Each of these records has the following fields:

- user-id: a unique identifier for the mobile subscriber (string)
- voice-consumption: the number of minutes spent calling (double)
- data-consumption: the number of consumed kilobytes of the data (double)
- sms-consumption: the amount of sent or received SMS (int)
- monthly-bill: the last amount of money billed to the subscriber (double)
- churned: indicates whether the subscriber has churned or not (boolean, true is the subscriber churned)

Out of the 100,000 records, roughly 80,000 are with churned equal to false.

You are required to create a model that should predict whether a subscriber churned knowing the remaining fields. This problem can be described as:

Pick the correct choices

- [ ] **an unsupervised learning problem because you are asked to cluster subscribers in two clusters (did churn, did not churn), as task for which the k-means algorithm can be used and k-means is an unsupervised learning algorithm**
- [x] **a supervised learning problem because you are asked to predict a label that is provided in the input dataset**
- [ ] **a supervised learning problem because churning is an anomaly from the operator's perspective and anomaly detection is a supervised learning kind of problem**
- [ ] **an unsupervised learning because the operator does not know why subscribers are churning (or not churning)**

Clear selection

☆ Machine Learning 2

You are taking over the work another team member that was too busy to work on the problem. He did manage to spend a few hours on it and came up with a first model that could accurately predict if a customer is churning with an accuracy of 80%. He thinks this is a good start. Do you agree?

Pick the correct choices

- [x] Yes
- [ ] No NO

Clear selection

☆ Machine Learning 3

Moving forward, you decide to first split the dataset into a training dataset and a testing dataset. Which training/testing ratio should you be using?

Pick the correct choices

- [ ] 99/1
- [ ] 99.9/0.1
- [x] 80/20
- [ ] 50/50

Clear selection

☆ Machine Learning 4

You have trained a first model and the results are out. Your model performs very well on the training dataset (accuracy is 98%) but not so well on the testing dataset (accuracy drops to 88%). Meanwhile, you have also shown the testing dataset to a domain expert in order to get a better baseline and he was able to predict is a given subscriber was going to churn with an accuracy of 98.5%. How would you characterize model that you trained?

Pick the correct choices

- [x] **The model has high variance**
- [ ] **The model has high bias**

Clear selection

☆ Machine Learning 5

What can you do to try to improve this model?

Pick the correct choices

- [ ] **Acquire or engineer more features in the dataset**
- [x] **Collect more data samples**
- [x] **Remove one or more feature from the dataset**
- [x] **Use regularization in your cost function**

Clear selection

☆ Machine Learning 6

You have thought of a completely different approach and have now trained another model. The results are different this time: The accuracy on the training dataset is 92% and the accuracy on the testing dataset is 90.5%. How would you characterize This new model?

Pick the correct choices

- [x] **The model has high bias**
- [ ] **The model has high variance**

Clear selection

☆ Machine Learning 7

What can you try to do to improve this new model?

Pick the correct choices

- [ ] **Use regularization in your cost function**
- [ ] **Remove one or more feature from the dataset**
- [x] **Acquire or engineer more features in the dataset**
- [ ] **Collect more data**

Clear selection

☆ Machine Learning 8

You now have a good model with good performance but it turns out the training is very slow. The model is trained with gradient descent. What can you do to reduce the training time?

Pick the correct choices

- [ ] **Decrease the step size**
- [x] **Use stochastic gradient descent instead of "pure" gradient descent**
- [ ] **Increase the step size**
- [ ] **Try different training / testing splits of the dataset**

Clear selection

☆ Machine Learning 9

You are now trying a different promising model that you are training with stochastic gradient descent but you are struggling with irregular results as the algorithms converges to relatively different values every time you run it. What can you do to stabilize the results?

Pick the correct choices

- [x] **Increase the batch size**
- [ ] **Try different training / testing splits of the dataset**
- [ ] **Decrease the batch size**

X ☑ **Increase the step size**

☑ **Decrease the step size**

Clear selection

☆ Machine Learning 10

When applied to different validation dataset composed of 100,000 records, your final model has the following confusion matrix when predicting the value of the churned field:

- True positive: 15,000
- True negative: 79,000
- False positive: 1,000
- False negative: 5,000

What is the accuracy and the recall of this model?

Pick the correct choices

- ☐ **Accuracy: 84% Recall: 75%**
- ☐ **Accuracy: 94% Recall: 98.75%**
- ☑ **Accuracy: 94% Recall: 75%**
- ☐ **Accuracy: 84% Recall: 98.75%**

Clear selection

Continue