타이타닉 #4 데이터 전처리 심화

2018-08-24 오전.

[지난 수업 피드백](#_gulw6k1nof1b)

[수업개요](#_py0l4j804t95)

[복습](#_438499xv4qgk)

[데이터 전처리 심화](#_jn87qsqzfc65)

[Missing Value](#_32n9d3fve4et)

[Binning](#_7kblbxnvrxhs)

[One-Hot-Encoding](#_h38hnjkv0mkr)

[Feature Scaling](#_cojafgfiha3a)

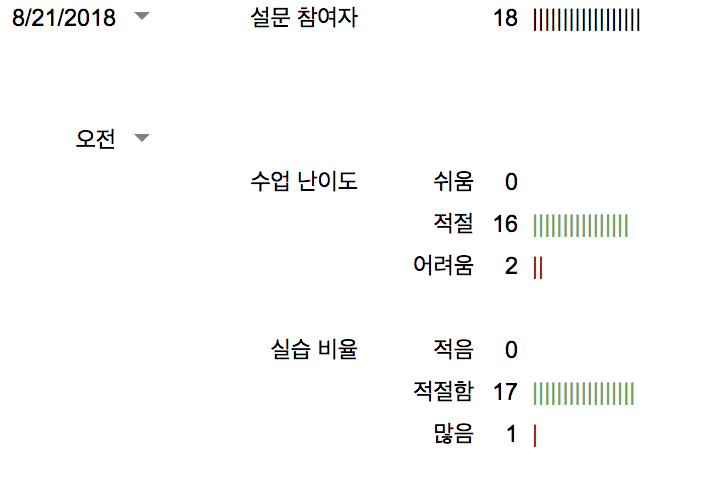
[GitHub에 올리기](#_cz9qmiqeqavm)

[수업 피드백 5m](#_sudhqi62fmhk)

# 출석 3m

# 아침회의 10m

# 지난 수업 피드백 5m



* 실습이 길고 실습시간이 조정되어 좋았음
* 눈코딩보다 손코딩하니 이해가 전 보다 잘 되었음
* 이 정도 속도가 적당 VS 여전히 빠르고 중간에 놓치면 따라가기 어려움

- 왜 pclass도 인코딩해줘야 하는지 모르겠어요ㅠ.ㅠ

- 코드에 대한 설명이 더 많았으면 좋겠습니다

- 지금 배우고 있는 것이 무엇인지 정확히 감이 잡히지가 않아요. ㅠ

- one-hot-encoding을 통해 범주형 데이터를 수치형으로 바꾸어준 다음 다시 범주형으로 바꾸는 것을 들었는데, 그렇게 해 주는 이유는 원상태에서 정제된 데이터를 보기 위함인지요? 원래는 범주형으로 보았기 때문에 이를 해석하기 좋게 다시금 바꿔주는 것이라고 이해하면 될까요?

# 수업개요

지난 수업 내용에 대해 얘기해 보고 사이킷런과 판다스를 활용하여 했던 작업들을 개선해 봅니다.

오늘 수업에서는 샘플 코드를 미리 보며 데이터 전처리를 설명하며, 해당 코드를 활용해 캐글 스코어를 올려보는 작업을 해봅니다.

# 복습

# 데이터 전처리 심화

동일 폴더에 아래의 이름으로 파일을 추가해 놓았습니다. 각각의 데이터 전처리 설명을 듣고 직접 적용해 보는 실습을 해봅니다.

## Missing Value

## Binning

## One-Hot-Encoding

## Feature Scaling

|  |
| --- |
| from sklearn.preprocessing import MinMaxScaler  def normalize\_age(data):  scaler = MinMaxScaler()  data["Age"] = scaler.fit\_transform(data["Age"].values.reshape(-1,1))  return data   def normalize\_fare(data):  scaler = MinMaxScaler()  data["Fare"] = scaler.fit\_transform(data["Fare"].values.reshape(-1,1))  return data  train\_data = normalize\_age(train\_data) test\_data = normalize\_age(test\_data)  train\_data = normalize\_fare(train\_data) test\_data = normalize\_fare(test\_data)  train\_data.head() |

# GitHub에 올리기

12:30~12:45

# 수업 피드백 5m

12:45~12:50

<https://goo.gl/forms/791qKEIT8SZRK1aC2>