# Feature preprocessing and generation with respect to models

Practice Quiz, 4 questions

## Congratulations! You passed!

Next Item

1 / 1 point

1.
What type does a feature with values: ['low', 'middle', 'high'] most likely have?

- Datetime
- Text
- Numeric
- Ordinal (ordered categorical)

**Correct**
Correct!

- Categorical
- Coordinates

---

2 / 2 points

2.
Suppose you have a dataset X, and a version of X where each feature has been standard scaled.

For which model types training or testing quality can be much different depending on the choice of the dataset?

- [ ] Random Forest

**Un-selected is correct**

Linear models

**Correct**
Correct! There are two reasons for this: first, amount of regularization applied to a feature depends on the feature's scale. Second, optimization methods can perform differently depending on relative scale of features.

Neural network

**Correct**
Correct! There are two reasons for this: first, amount of regularization applied to a feature depends on the feature's scale. Second, optimization methods can perform differently depending on relative scale of features.

Nearest neighbours

**Correct**
Correct! The reason for it is that the scale of features impacts the distance between samples. Thus, with different scaling of the features nearest neighbors for a selected object can be very different.

GBDT

**Un-selected is correct**

---

1 / 1 point

3.
Suppose we want to fit a GBDT model to a data with a categorical feature. We need to somehow encode the feature. Which of the following statements are true?

Depending on the dataset either of label encoder or one-hot encoder could be better

**Correct**
Correct! It's good idea to try both, if you don't have any better ideas to try.

Label encoding is always better to use than one-hot encoding

One-hot encoding is always better than label encoding

---

# Feature preprocessing and generation with respect to models

2/2 points

Practice Quiz, 4 questions

4.

What can be useful to do about missing values?

- [x] Replace them with a constant (-1/-999/etc.)

**Correct**

This is one of the most frequent ways to deal with missing values.

- [x] Nothing, but use a model that can deal with them out of the box

**Correct**

Some models like XGBoost and CatBoost can deal with missing values out-of-box. These models have special methods to treat them and a model's quality can benefit from it.

- [x] Reconstruct them (for example train a model to predict the missing values)

**Correct**

This one is tricky, but sometimes it can prove useful.

- [ ] Impute with feature variance

**Un-selected is correct**

- [ ] Apply standard scaler

**Un-selected is correct**

- [x] Remove rows with missing values

**Correct**

This one is possible, but it can lead to loss of important samples and a quality decrease.

- [x] Impute with a feature mean

**Correct**

This is one of the most frequent ways to deal with missing values.

## Feature preprocessing and generation with respect to models

Practice Quiz, 4 questions