Kaggle Master-Week 2 Quiz

Toplam puan 100/100

E-posta adresi *

sevdanurgenc@gmail.com

Name-Last Name: *

Sevdanur GENC

✓ Q1- Which of the following statements are true about the intended use of cross-validation? * 10/10

I - To reduce randomness while measuring model performance.

II - To get a better measure of model performance.

III - To increase model's training performance.

IV - To increase MAE (mean absolute error) or MSE (mean squared error).

- ○ I, II, IV
- ○ II, III
- ◉ I, II ✓

All of them.

✓ Q2- Which of the following statements are true about LabelEncoder and OneHotEncoder? * 10/10

I-They help us to deal with categorical values.

II-Label Encoding assigns each value to a different integer whether it is unique or not.

III-One Hot Encoding creates new column for every possible value in the original data.

IV-For large number of categorical variable count value (such as 15 different values) it is not good to use One Hot Encoder generally.

I, II, IV

I, III, IV ✓

I, II, III

All of them.

✓ Q3- Which of the following statement is inconsistent with pipelines? * 10/10

With pipelines, there is less probability to forget a preprocessing step.

It's hard to productionize a model with pipelines. ✓

You won't need to manually keep track of your training and validation data at each step with a pipeline.

With a pipeline, we can use the cross-validation technique easily.

✓ Q4- print(df(head).method()) Assume that you want to print locations of the missing values in the top 10 rows. Which method is suitable for this? * 10/10

- ○ dropna(how='any')
- ○ isnan
- ○ notnull
- ◉ isnull ✓

✓ Q5- Which of the following is not a Booster parameter of XGBoost? * 10/10

- ○ min_child_weight
- ◉ objective ✓
- ○ max_leaf_nodes
- ○ colsample_bylevel

✓ Q6- What do the highlighted code pieces mean? * 10/10

```
X_train_plus = X_train.copy()
X_valid_plus = X_valid.copy()
for col in cols_with_missing:
    X_train_plus[col + '_was_missing'] = X_train_plus[col].isnull()
    X_valid_plus[col + '_was_missing'] = X_valid_plus[col].isnull()
my_imputer = SimpleImputer()
imputed_X_train_plus = pd.DataFrame(my_imputer.fit_transform(X_train_plus))
imputed_X_valid_plus = pd.DataFrame(my_imputer.transform(X_valid_plus))
imputed_X_train_plus.columns = X_train_plus.columns
imputed_X_valid_plus.columns = X_valid_plus.columns
```

- ○ To make new columns indicating what will be imputed
- ○ For imputation
- ○ To make copy to avoid changing original data
- ◉ To put removed column names back ✓

✓ Q7- Which of the below is/are nominal variable(s)? * 10/10

I - Gender

II - Genotype

III - Religious preference

IV- IQ

V - Income earned in a week.

- ○ I, II
- ◉ I, II, III ✓
- ○ II, III, IV
- ○ All of them.

✓ Q8- Which of the following statements are true about “max_depth” hyperparameter in Random Forest? * 10/10

```
I-      Lower is better parameter in case of same validation accuracy
II-     Higher is better parameter in case of same validation accuracy
III-    Increase the value of max_depth may overfit the data
IV-     Increase the value of max_depth may underfit the data
```

- ○ I, IV
- ○ II, IV

- (●) I, III ✓
- ( ) II, III

✓ Q9- You will build a model to predict housing prices. The model will be deployed on an ongoing basis, to predict the price of a new house when a description is added to a website. Here are four features that could be used as predictors. Which of the features is most likely to be a source of leakage? * 10/10

- ( ) Size of the house (in square meters)
- (●) Average sales price of homes in the same neighborhood ✓
- ( ) Latitude and longitude of the house
- ( ) Whether the house has a basement

✓ Q10- How is the Gradient Boosting cycle proceed? Please choose the correct order from the mixed statements below. * 10/10

I- We add the new model to ensemble.
II- We use the current ensemble to generate predictions for each observation in the dataset.
III- We use the loss function to fit a new model that will be added to the ensemble.

- ( ) I-II-III
- ( ) I-III-II
- ( ) II-I-III
- (●) II-III-I ✓