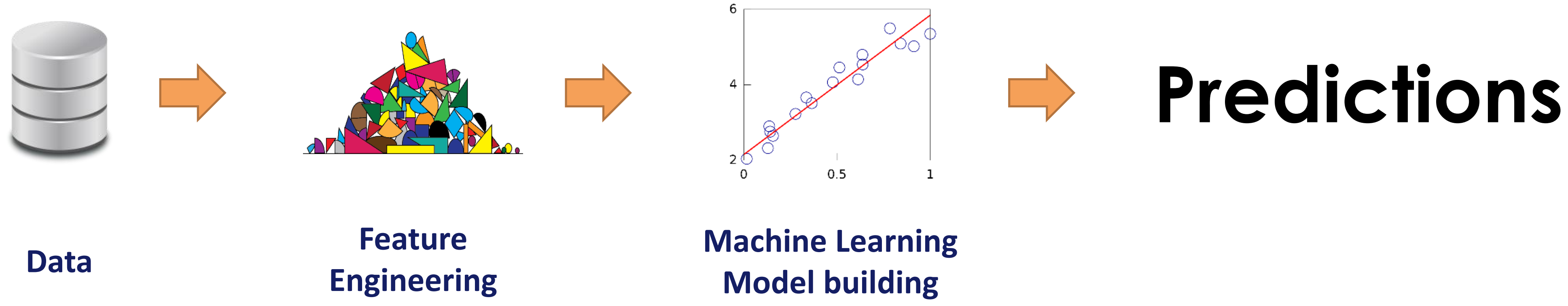
Train In Data

# Assembling a Pipeline

# Assembling a pipeline: steps



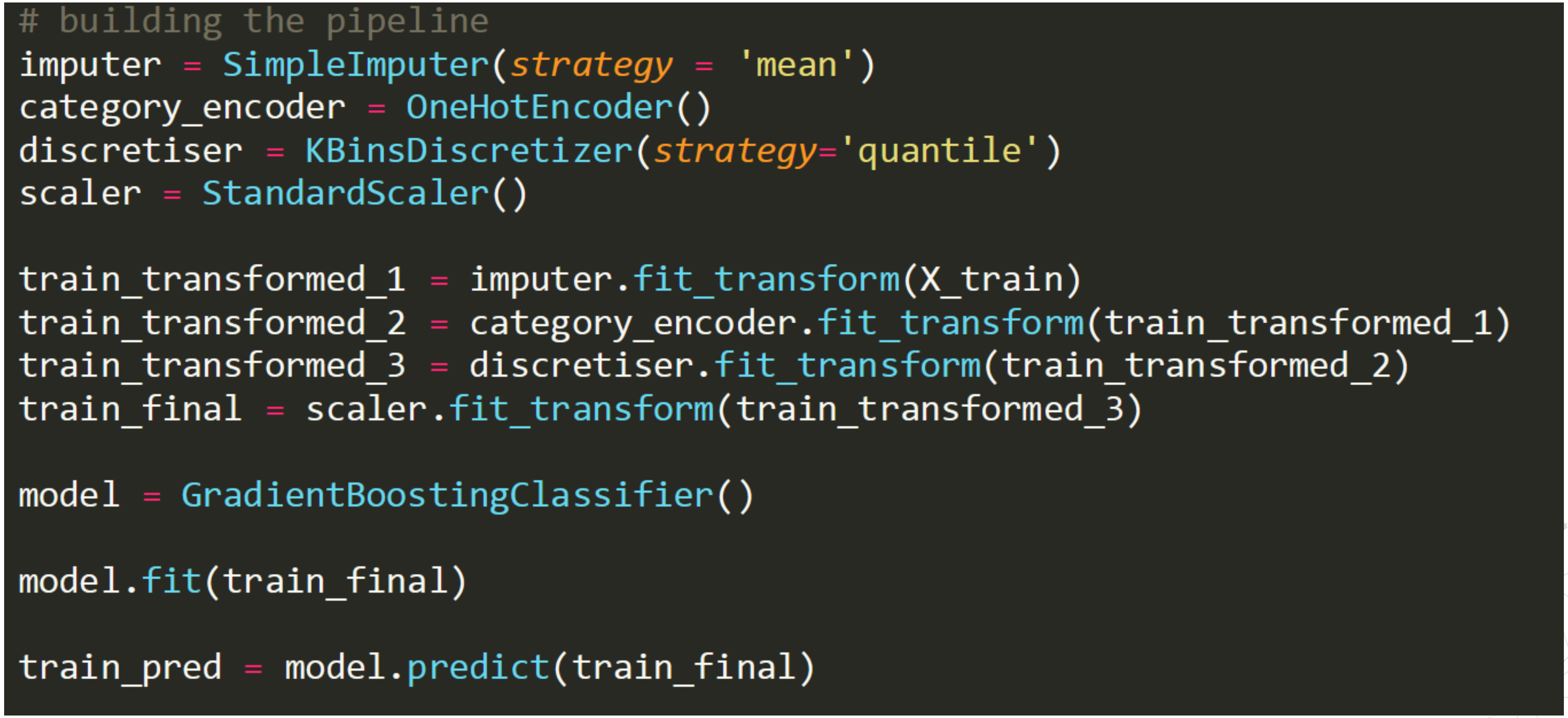
Data

Feature Engineering

Machine Learning Model building

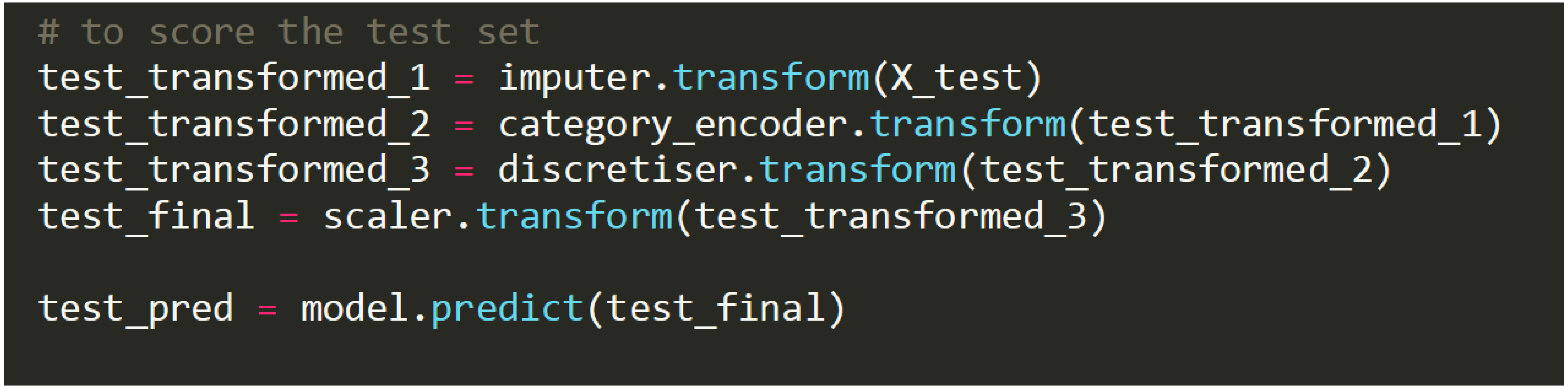
Predictions

# Pipeline steps: training



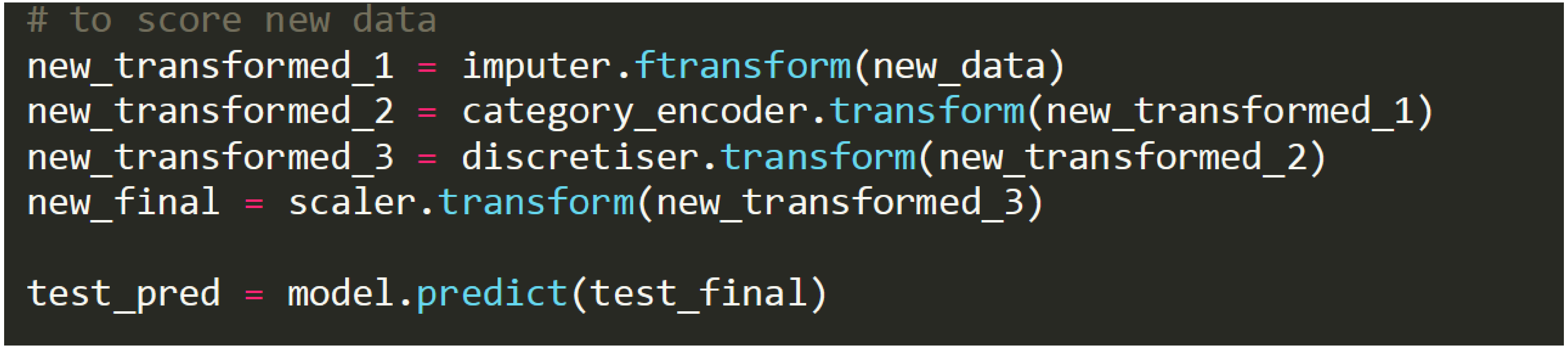
```
# building the pipeline
imputer = SimpleImputer(strategy = 'mean')
category_encoder = OneHotEncoder()
discretiser = KBinsDiscretizer(strategy='quantile')
scaler = StandardScaler()

train_transformed_1 = imputer.fit_transform(X_train)
train_transformed_2 = category_encoder.fit_transform(train_transformed_1)
train_transformed_3 = discretiser.fit_transform(train_transformed_2)
train_final = scaler.fit_transform(train_transformed_3)

model = GradientBoostingClassifier()

model.fit(train_final)

train_pred = model.predict(train_final)
```

# Pipeline steps: testing



```
# to score the test set
test_transformed_1 = imputer.transform(X_test)
test_transformed_2 = category_encoder.transform(test_transformed_1)
test_transformed_3 = discretiser.transform(test_transformed_2)
test_final = scaler.transform(test_transformed_3)

test_pred = model.predict(test_final)
```

# Pipeline steps: new data

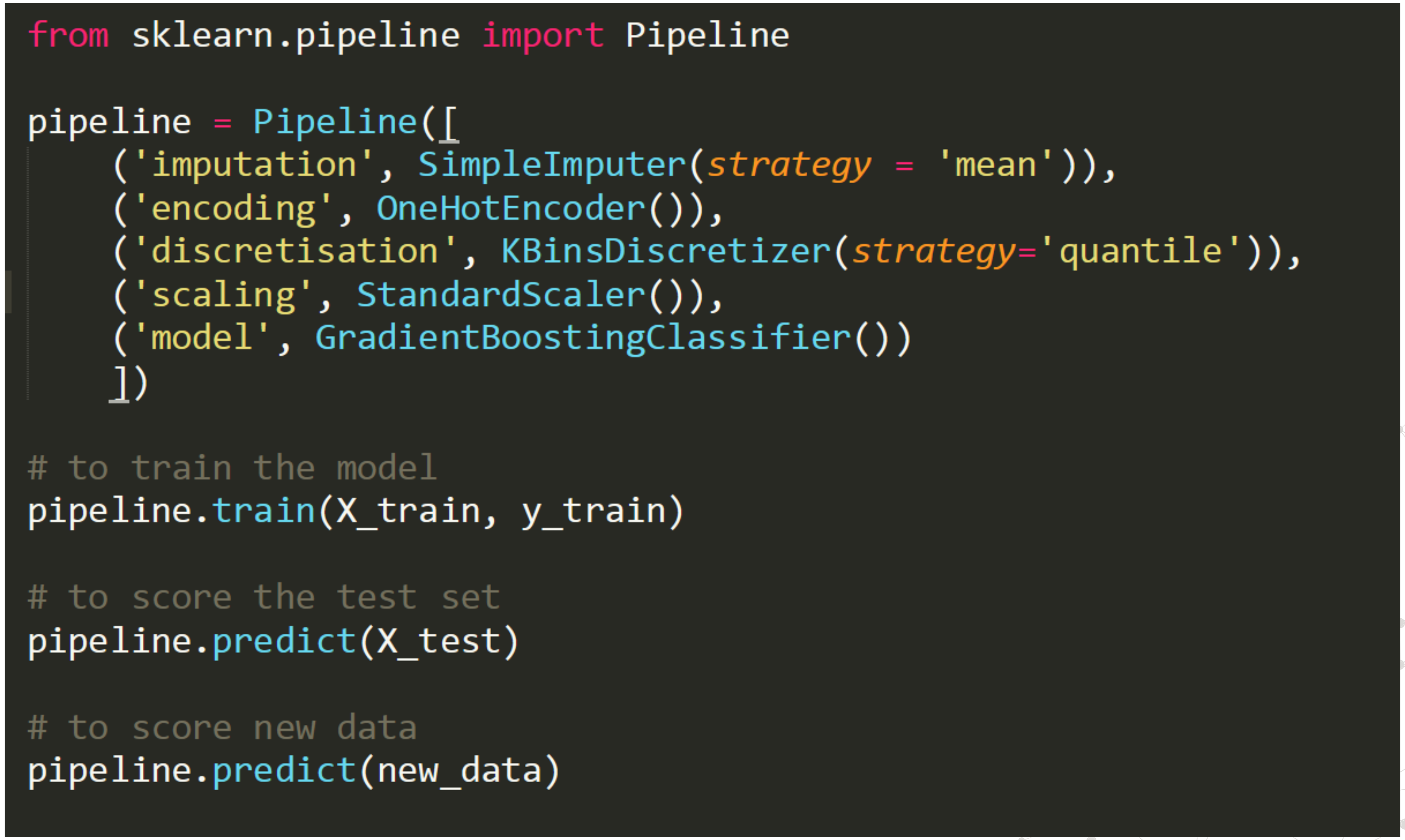


```
# to score new data
new_transformed_1 = imputer.ftransform(new_data)
new_transformed_2 = category_encoder.transform(new_transformed_1)
new_transformed_3 = discretiser.transform(new_transformed_2)
new_final = scaler.transform(new_transformed_3)

test_pred = model.predict(test_final)
```

# Assembling a Pipeline

**Pipeline** - class that allows to run transformers and a machine learning model in sequence.

- Most steps are Transformers
- Last step can be an Estimator



```
from sklearn.pipeline import Pipeline

pipeline = Pipeline([
    ('imputation', SimpleImputer(strategy = 'mean')),
    ('encoding', OneHotEncoder()),
    ('discretisation', KBinsDiscretizer(strategy='quantile')),
    ('scaling', StandardScaler()),
    ('model', GradientBoostingClassifier())
    ])

# to train the model
pipeline.train(X_train, y_train)

# to score the test set
pipeline.predict(X_test)

# to score new data
pipeline.predict(new_data)
```

# Scikit-learn transformers

- Missing Data Imputation
  - SimpleImputer
- Categorical Variable Encoding
  - OneHotEncoder
  - LabelEncoder
- Discretisation
  - KBinsDiscretizer

- Variable Transformation
  - PowerTransformer
  - FunctionTransfomer
- Scaling
  - StandardScaler
  - MinMaxScaler
  - RobustScaler
  - Normalizer

# Feature-engine transformers

- Missing Data Imputation
  - MeanMedianImputer
  - RandomSampleImputer
  - EndTailImputer
  - AddNaNBinaryImputer
  - CategoricalVariableImputer
  - FrequentCategoryImputer

- Categorical Variable Encoding
  - CountFrequencyCategoricalEncoder
  - OrdinalCategoricalEncoder
  - MeanCategoricalEncoder
  - WoERatioCategoricalEncoder
  - OneHotCategoricalEncoder
  - RareLabelCategoricalEncoder

- Outlier Removal
  - Windsorizer
  - ArbitraryOutlierCapper

- Discretisation
  - EqualFrequencyDiscretiser
  - EqualWidthDiscretiser
  - DecisionTreeDiscretiser

- Variable Transformation
  - LogTransformer
  - ReciprocalTransformer
  - PowerTransformer
  - BoxCoxTransformer

Train In Data

# THANK YOU

www.trainindata.com