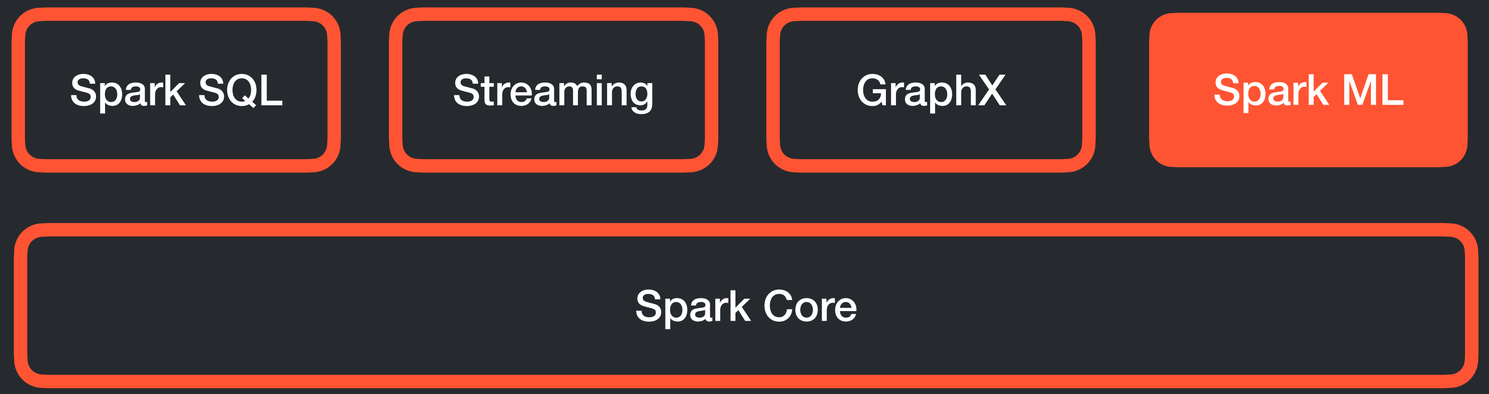
SPARK ML

KARPOV.COURSES

SPARK ML



SPARK < 2.0

Spark MLlib - RDD

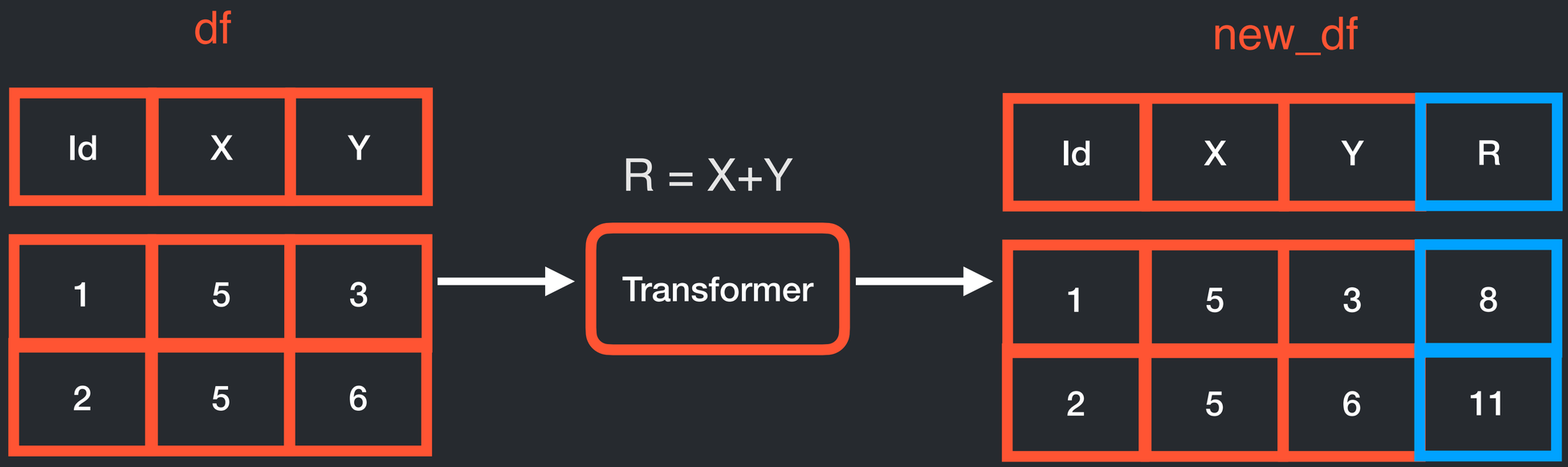
2.0 <= SPARK

Spark ML - DataFrame

КОМПОНЕНТ TRANSFORMER

Transformer - Алгоритм преобразования одного набора данных в другой.

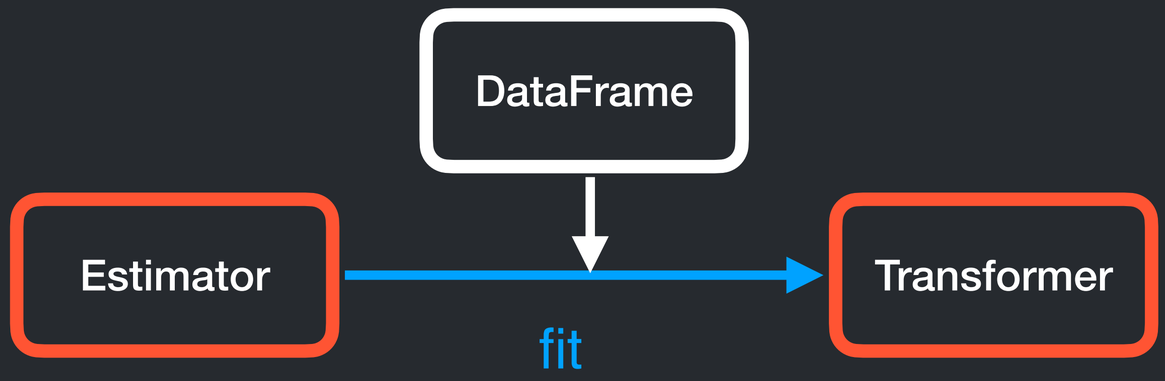
new_df = Transformer.transform(df)



КОМПОНЕНТ ESTIMATOR

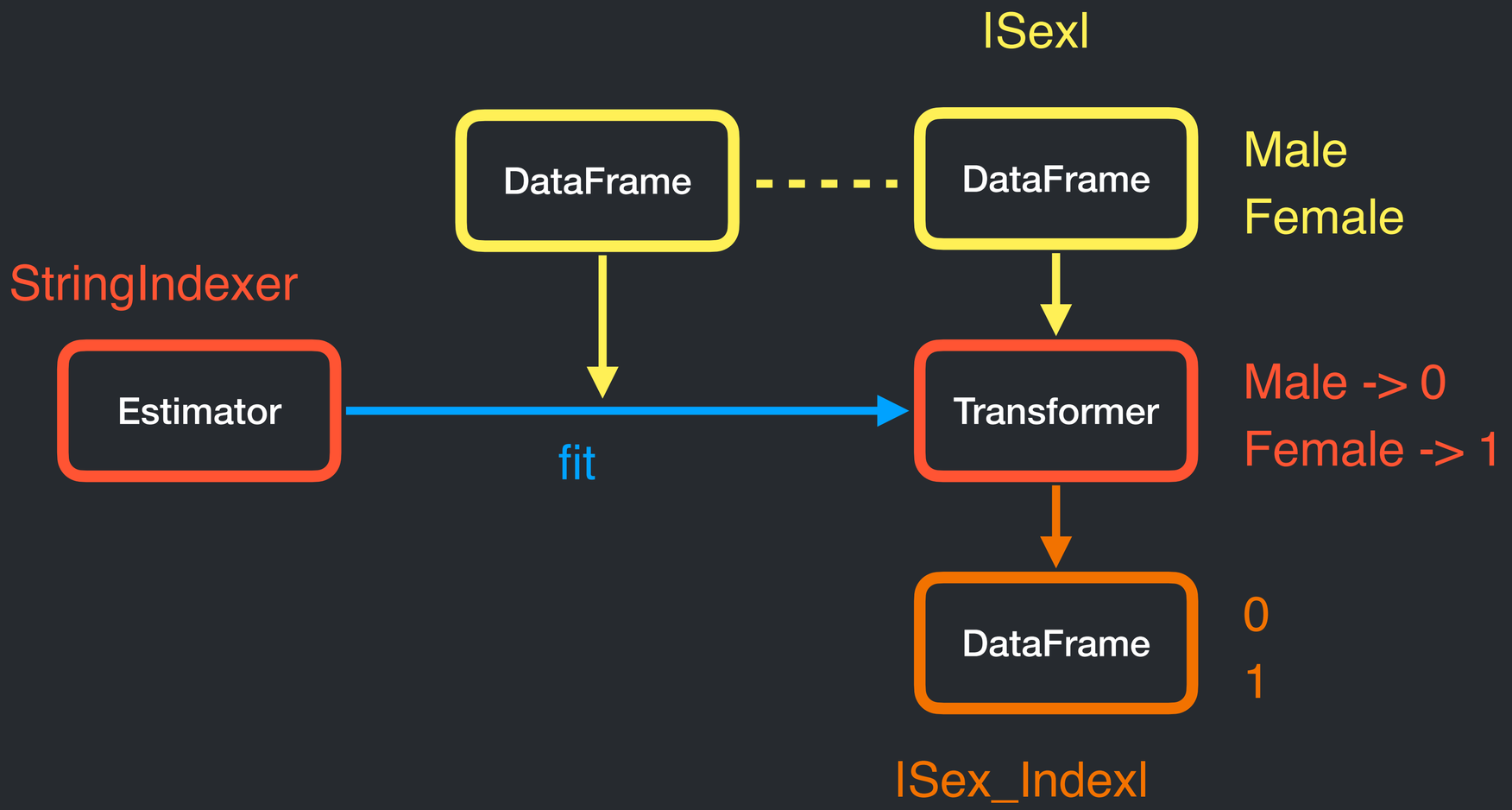
Estimator - Алгоритм создания Transformer на основе данных.

transformer = Estimator.fit(df)



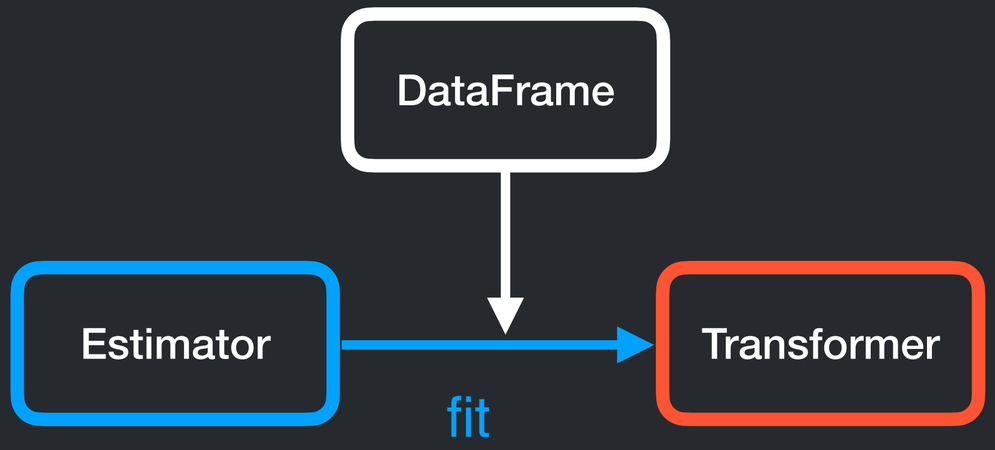
ENCODING

Encoding - процесс, с помощью которого признаки преобразуются в подходящую алгоритмам форму.



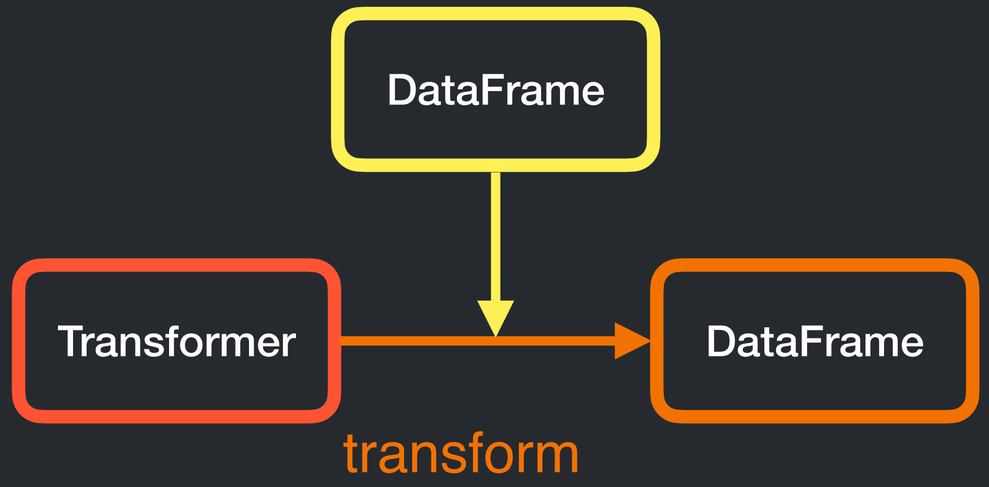
МАШИННОЕ ОБУЧЕНИЕ В SPARK ML

ОБУЧЕНИЕ МОДЕЛИ




```
model = LinearRegression().fit(df)
```

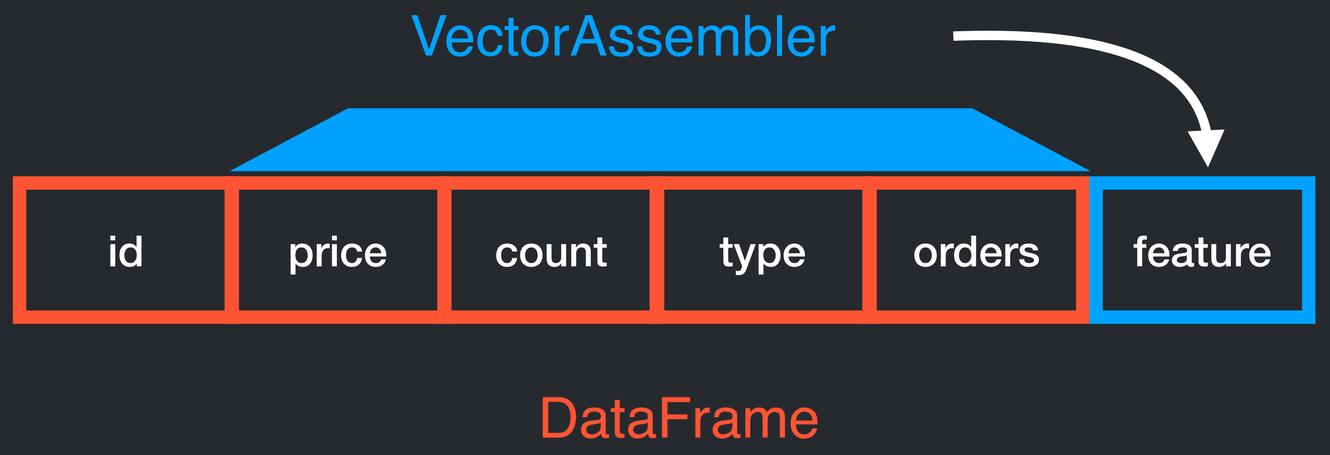
ПРИМЕНЕНИЕ МОДЕЛИ




```
prediction_df = model.transform(df)
```

ВЕКТОРИЗАЦИЯ

Векторизация (Embedding) - векторное представление данных.



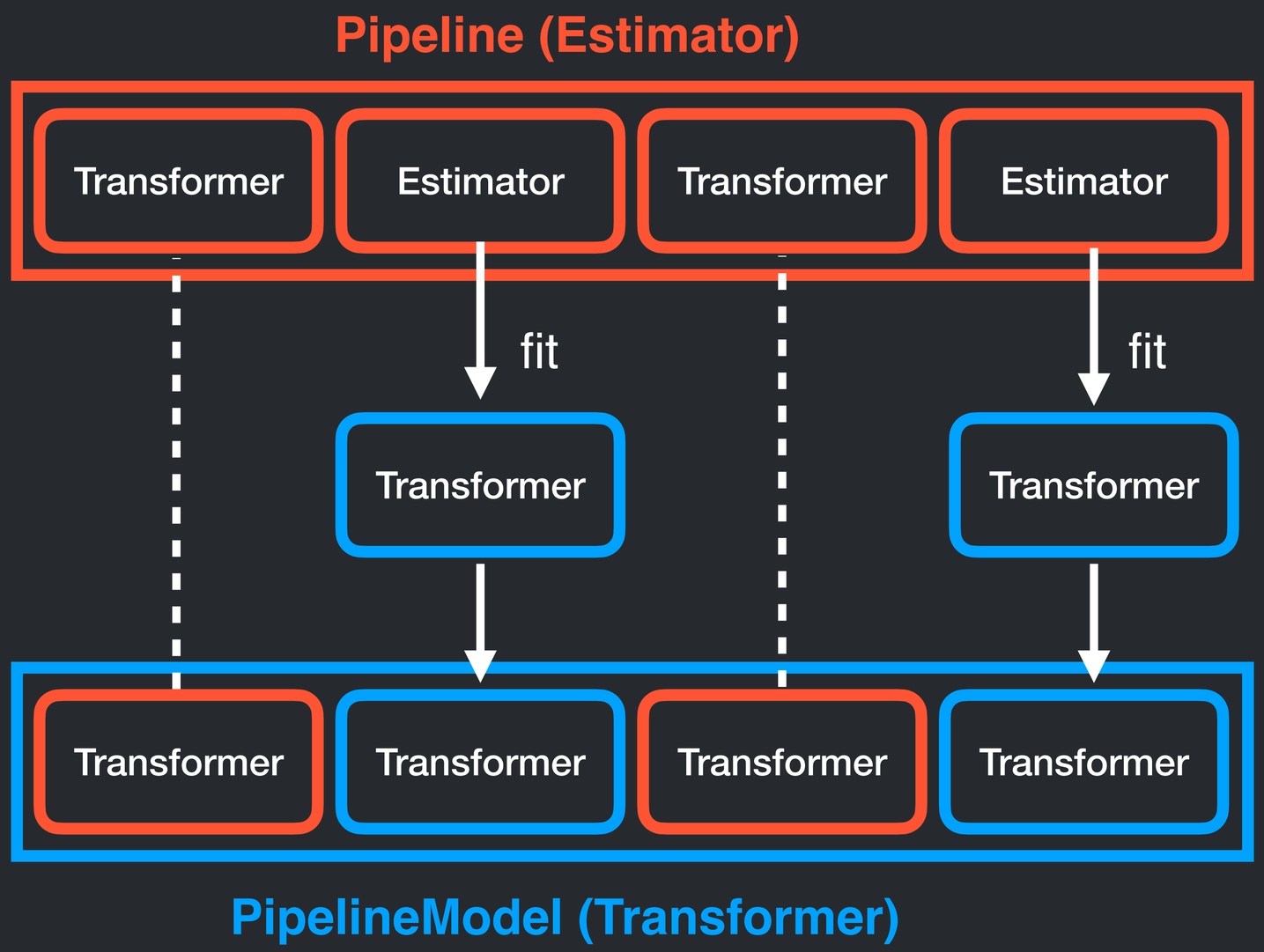
КОМПОНЕНТ PIPELINE

Pipeline - конвейер, объединяющий любое количество Transformer и Estimator для создания процесса машинного обучения.

Свойства:

- Задается в виде последовательности из Transformer или Estimator.
- Любой Transformer созданный в результате работы Estimator, автоматически становятся частью Pipeline.
- Все компоненты являются Stateless, т.е не хранят состояние.

КОМПОНЕНТ PIPELINE

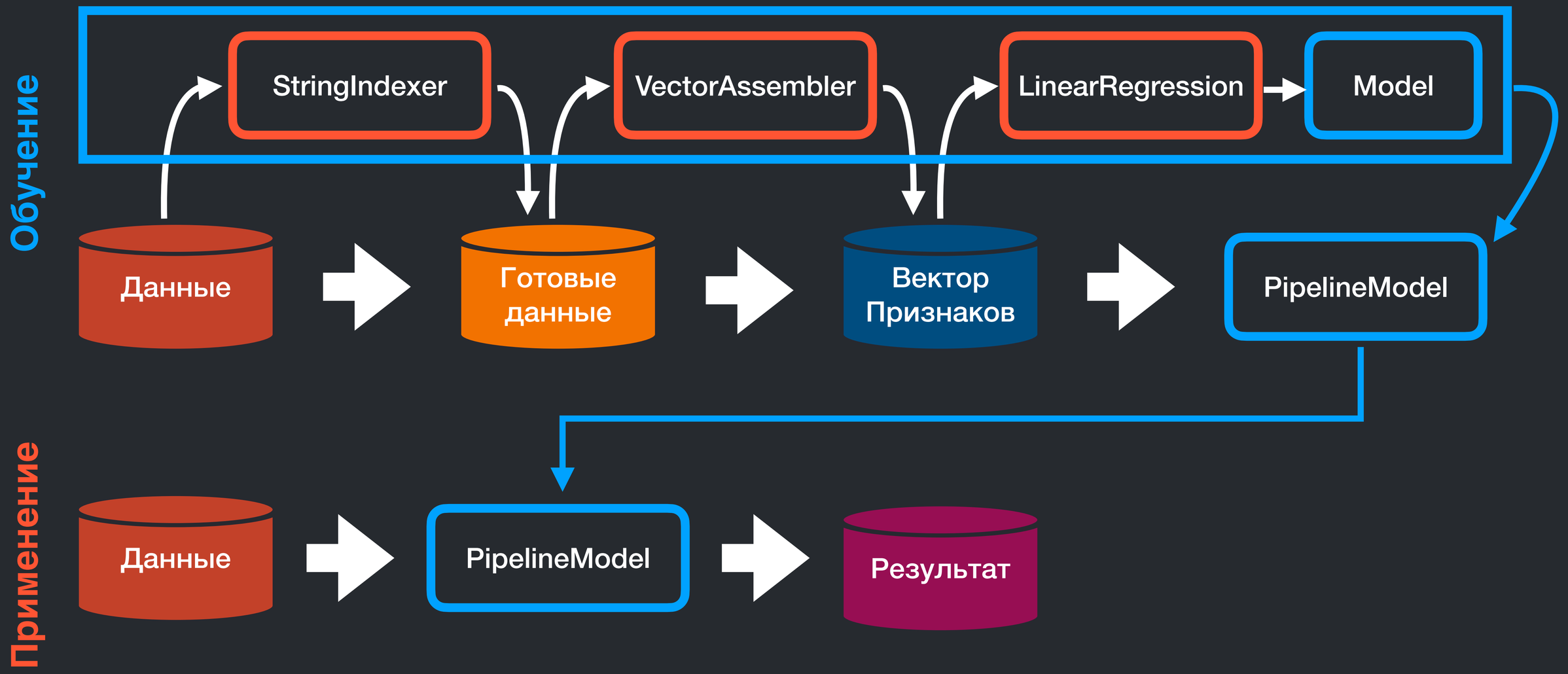



```
pipeline = Pipeline(stages=[. . .])
```

```
model = pipeline.fit(train_df)
```

```
predict_df = model.transform(test_df)
```

ПРИМЕНЕНИЕ PIPELINE



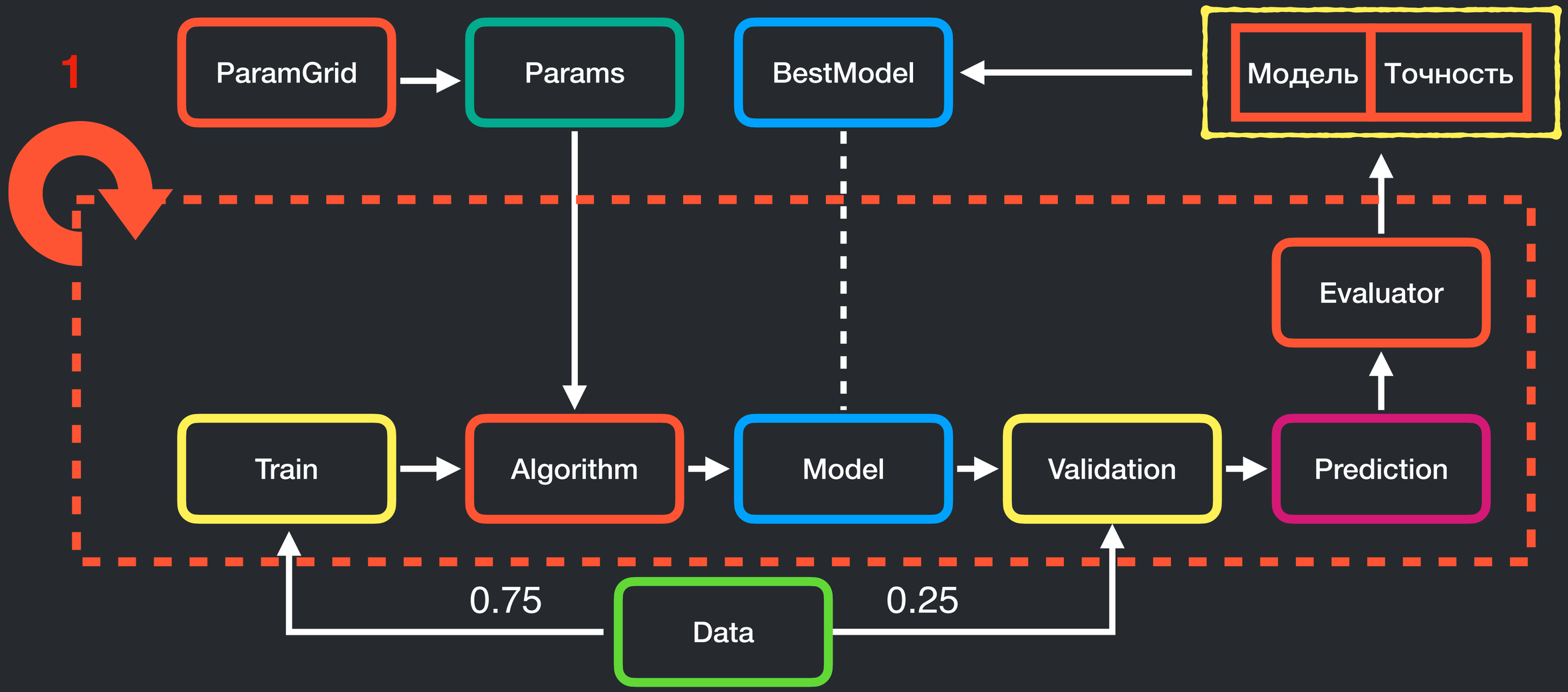
КОМПОНЕНТ EVALUATOR

Evaluator - оценщик качества модели согласно указанному алгоритму.

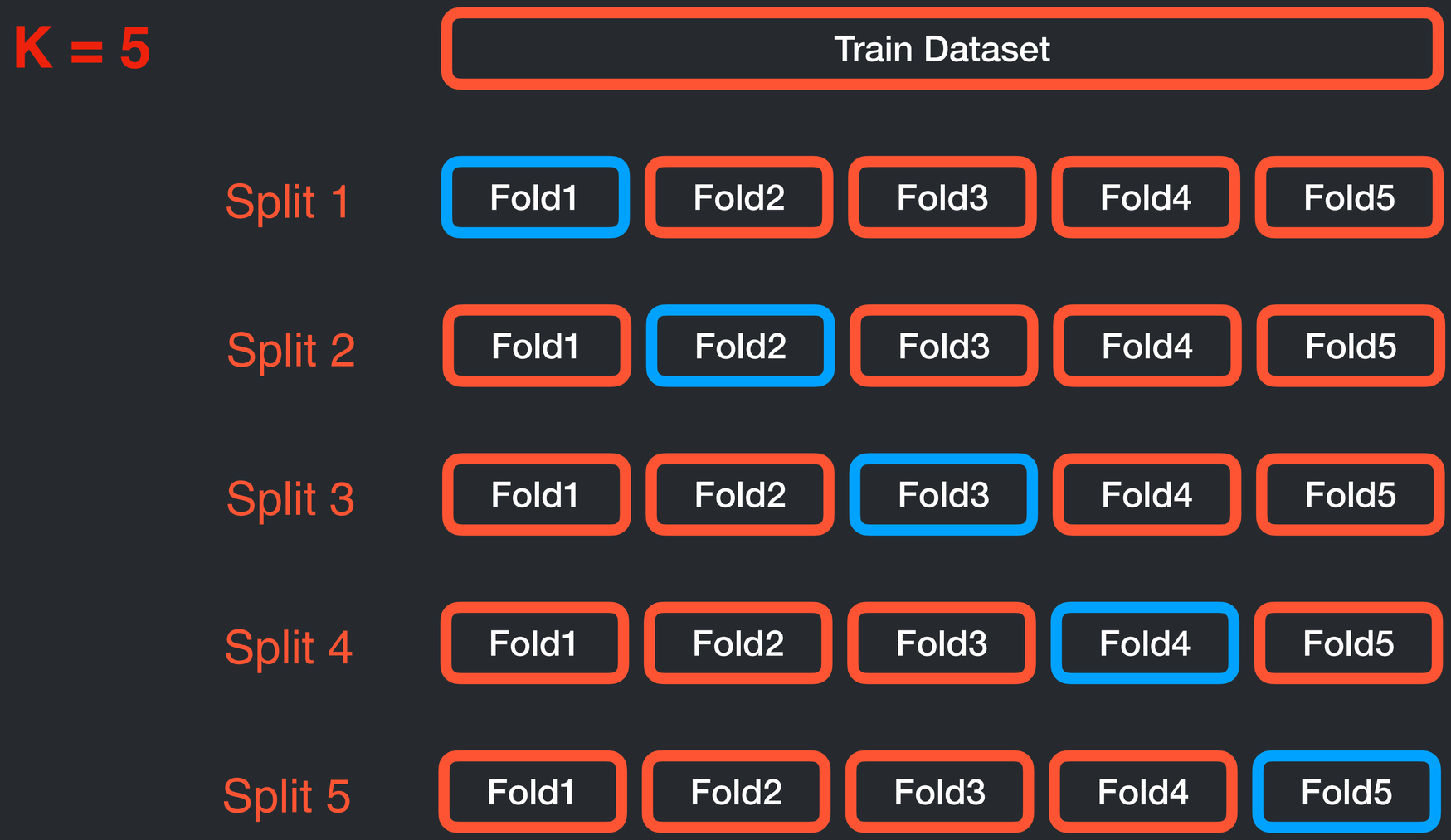
- Регрессия
  MSE, RMSE, MAE, R2 (Коэффициент детерминации)
- Классификация
  Accuracy, Precision, Recall, F-measure, ROC, AUROC, AUPRC
- Ранжирование
  Precision at K, MAP, NDCG

ОПТИМИЗАЦИЯ ГИПЕРПАРАМЕТРОВ

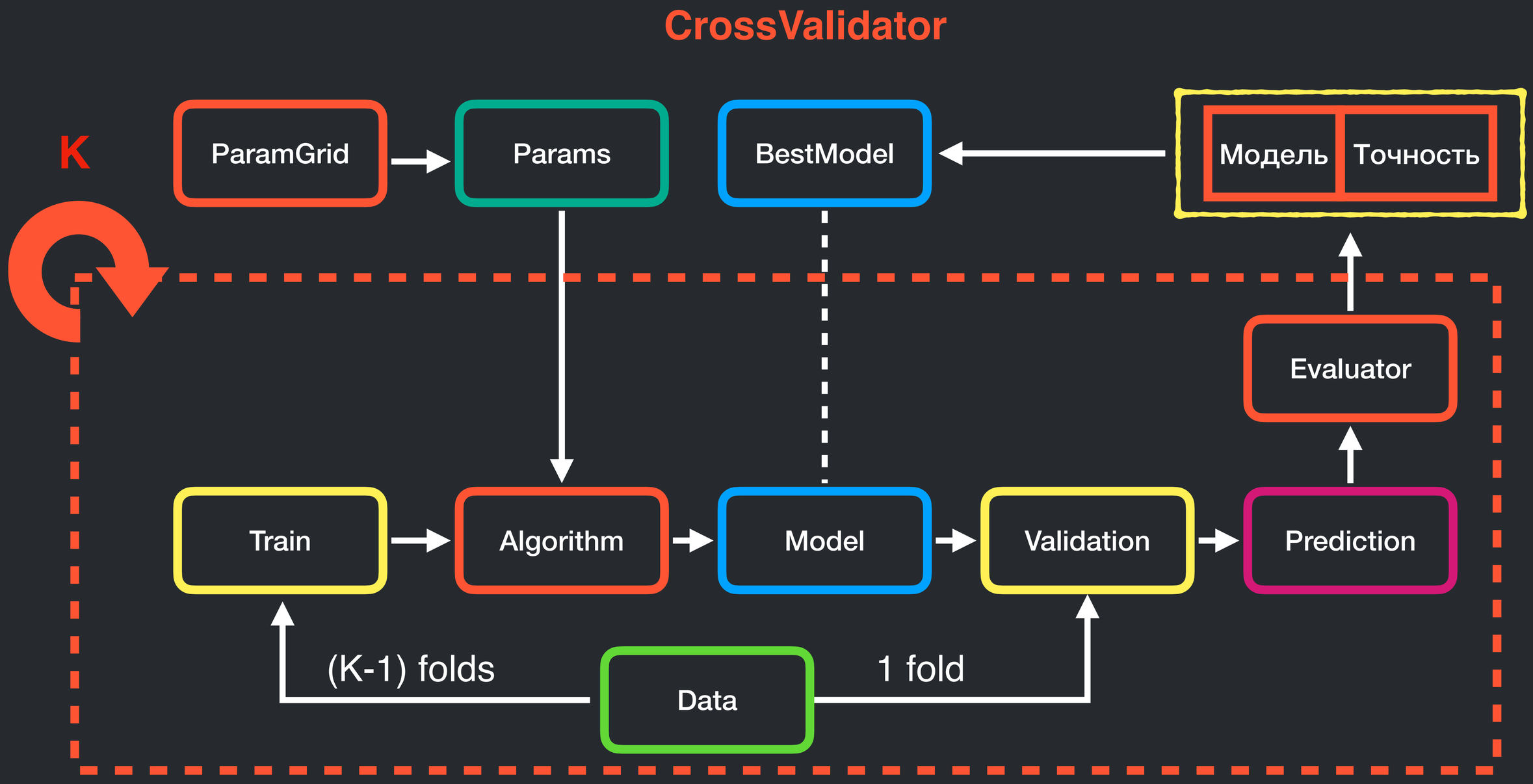
TrainValidationSplit



КРОСС-ВАЛИДАЦИЯ



ОПТИМИЗАЦИЯ ГИПЕРПАРАМЕТРОВ



ПОДДЕРЖИВАЕМЫЕ КЛАССЫ МОДЕЛЕЙ

Поддерживаемые классы задач:

- Регрессия
- Кластеризация
- Классификация
- Деревья решений
- Ансамбли деревьев (GBT)
- Collaborative filtering (ALS)