livelessons
video instruction from technology experts

Lesson 5: Advanced Applications

5.9 Tuning Models: Features, Cross Validation, and Grid Search

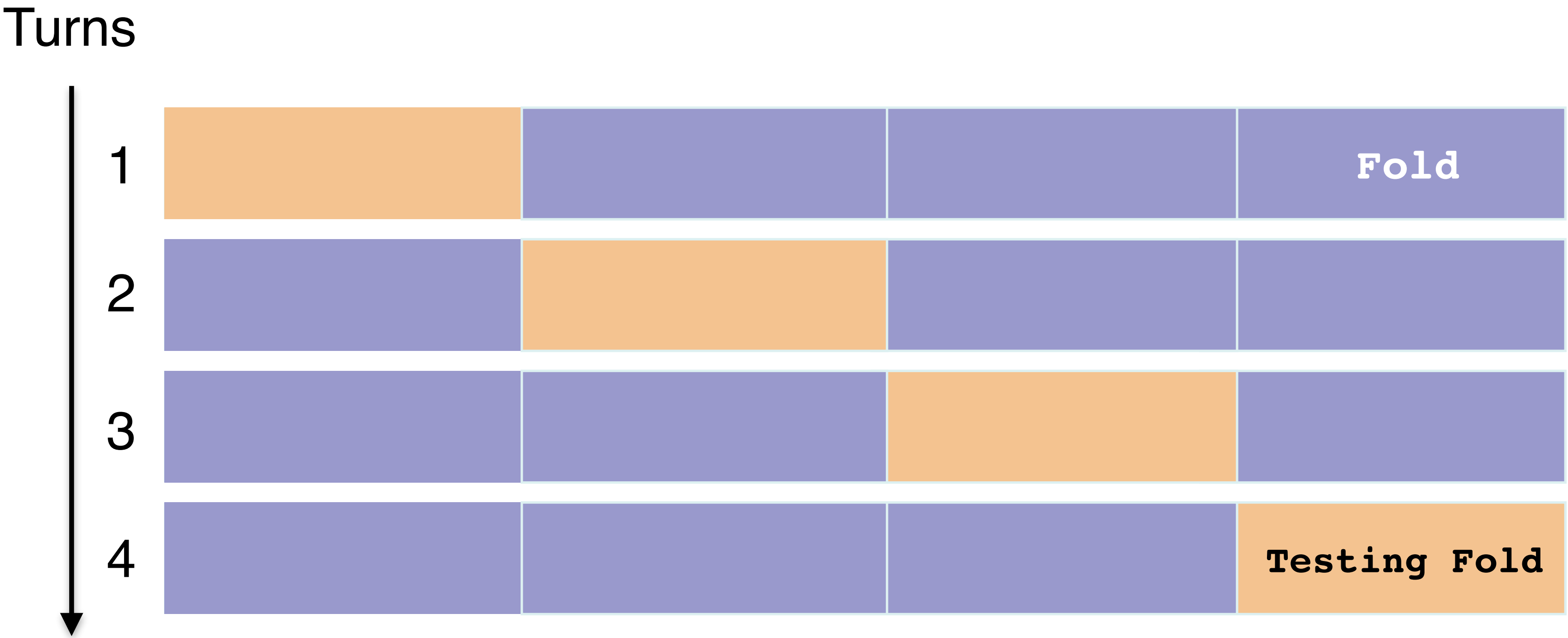
livelessons

Feature Transformers

- Text: Tokenizer, TF-IDF, and Word2Vec
- Transformation: StandardScaler, Normalizer, PolynomialExpansion
- Categorical: StringIndexer, OneHotEncoder, VectorIndexer

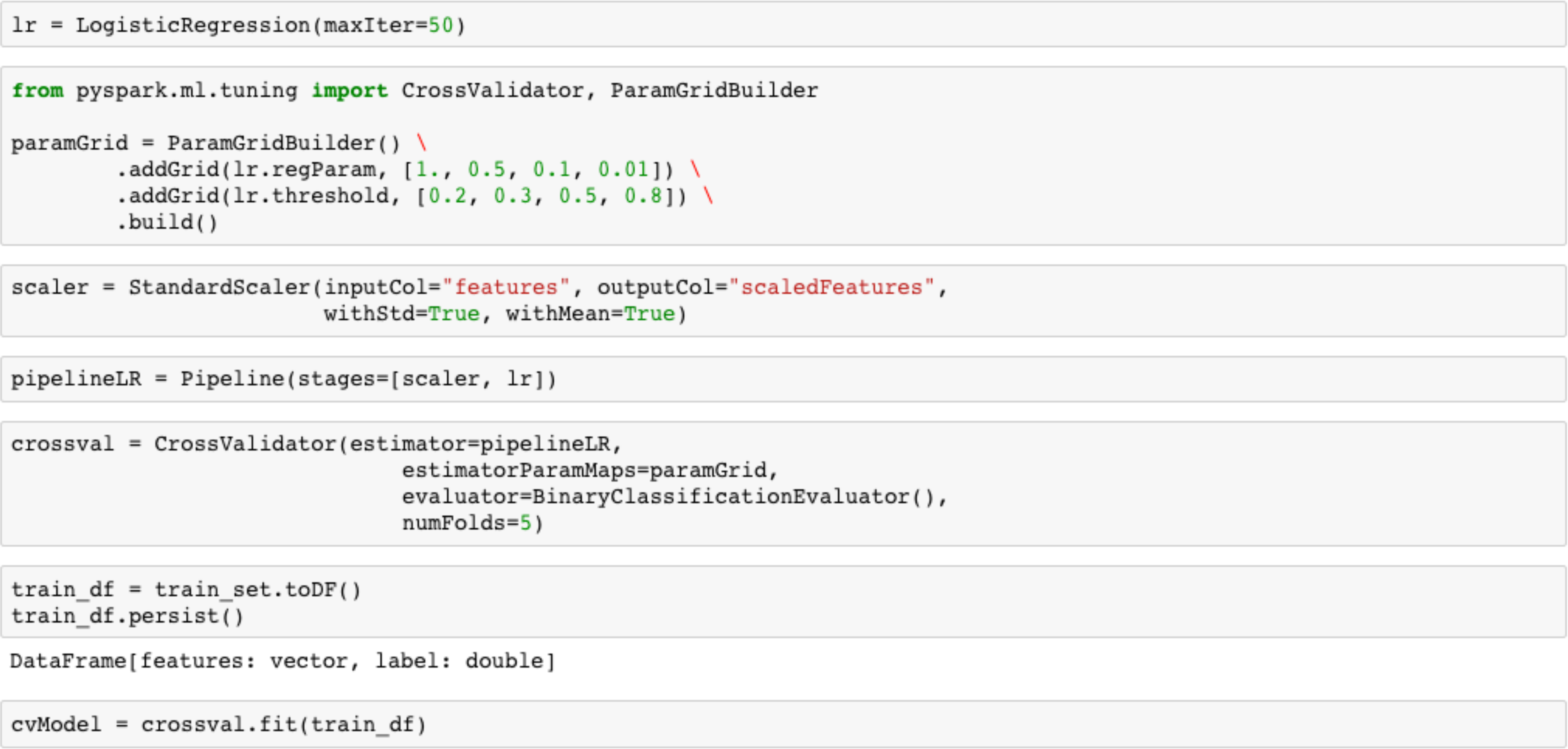
http://spark.apache.org/docs/latest/ml-features.html#feature-transformers

k-fold Cross Validation



Grid Search

- Exhaustive brute force search
- Find optimal hyperparameters or models
- Computationally costly
- But embarrassingly parallel!



```
lr = LogisticRegression(maxIter=50)
```

```
from pyspark.ml.tuning import CrossValidator, ParamGridBuilder

paramGrid = ParamGridBuilder() \
        .addGrid(lr.regParam, [1., 0.5, 0.1, 0.01]) \
        .addGrid(lr.threshold, [0.2, 0.3, 0.5, 0.8]) \
        .build()
```

```
scaler = StandardScaler(inputCol="features", outputCol="scaledFeatures",
                         withStd=True, withMean=True)
```

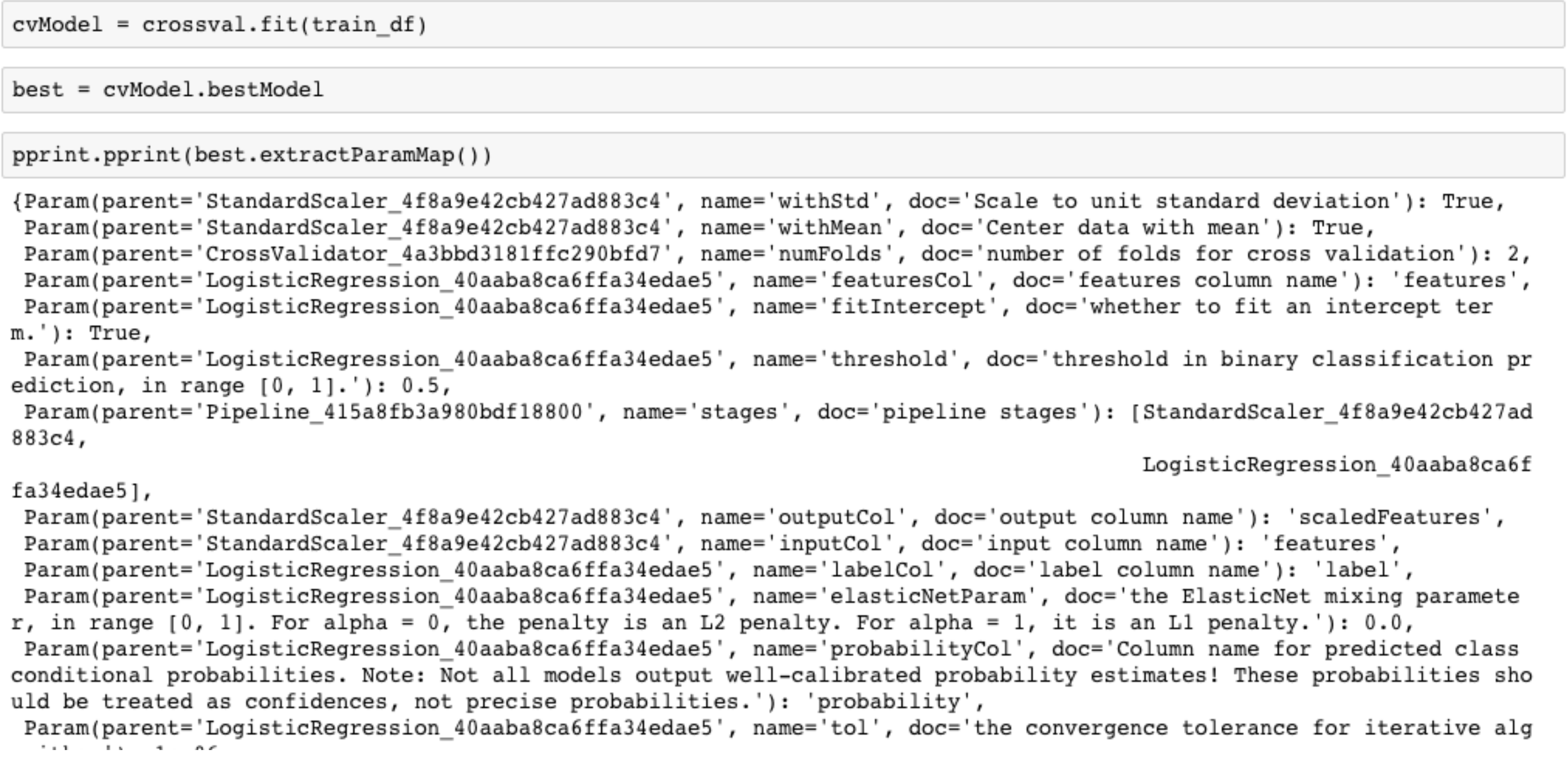
```
pipelineLR = Pipeline(stages=[scaler, lr])
```

```
crossval = CrossValidator(estimator=pipelineLR,
                          estimatorParamMaps=paramGrid,
                          evaluator=BinaryClassificationEvaluator(),
                          numFolds=5)
```

```
train_df = train_set.toDF()
train_df.persist()
```

```
DataFrame[features: vector, label: double]
```

```
cvModel = crossval.fit(train_df)
```



```
cvModel = crossval.fit(train_df)
```

```
best = cvModel.bestModel
```

```
pprint.pprint(best.extractParamMap())
```

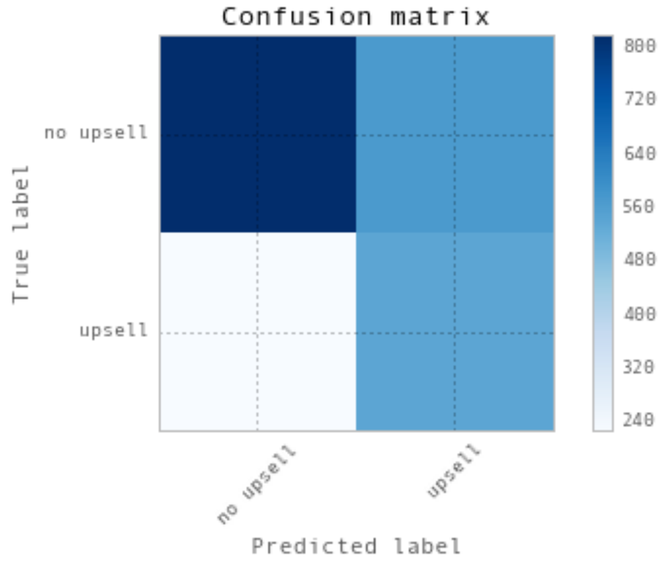
```
{Param(parent='StandardScaler_4f8a9e42cb427ad883c4', name='withStd', doc='Scale to unit standard deviation'): True,
 Param(parent='StandardScaler_4f8a9e42cb427ad883c4', name='withMean', doc='Center data with mean'): True,
 Param(parent='CrossValidator_4a3bbd3181ffc290bfd7', name='numFolds', doc='number of folds for cross validation'): 2,
 Param(parent='LogisticRegression_40aaba8ca6ffa34edae5', name='featuresCol', doc='features column name'): 'features',
 Param(parent='LogisticRegression_40aaba8ca6ffa34edae5', name='fitIntercept', doc='whether to fit an intercept ter
m.'): True,
 Param(parent='LogisticRegression_40aaba8ca6ffa34edae5', name='threshold', doc='threshold in binary classification pr
ediction, in range [0, 1].'): 0.5,
 Param(parent='Pipeline_415a8fb3a980bdf18800', name='stages', doc='pipeline stages'): [StandardScaler_4f8a9e42cb427ad
883c4,
                                                                                       LogisticRegression_40aaba8ca6f
fa34edae5],
 Param(parent='StandardScaler_4f8a9e42cb427ad883c4', name='outputCol', doc='output column name'): 'scaledFeatures',
 Param(parent='StandardScaler_4f8a9e42cb427ad883c4', name='inputCol', doc='input column name'): 'features',
 Param(parent='LogisticRegression_40aaba8ca6ffa34edae5', name='labelCol', doc='label column name'): 'label',
 Param(parent='LogisticRegression_40aaba8ca6ffa34edae5', name='elasticNetParam', doc='the ElasticNet mixing paramete
r, in range [0, 1]. For alpha = 0, the penalty is an L2 penalty. For alpha = 1, it is an L1 penalty.'): 0.0,
 Param(parent='LogisticRegression_40aaba8ca6ffa34edae5', name='probabilityCol', doc='Column name for predicted class
conditional probabilities. Note: Not all models output well-calibrated probability estimates! These probabilities sho
uld be treated as confidences, not precise probabilities.'): 'probability',
 Param(parent='LogisticRegression_40aaba8ca6ffa34edae5', name='tol', doc='the convergence tolerance for iterative alg
orithms.'): 1e-06,
             precision    recall  f1-score   support

        0.0       0.78      0.59      0.67      1391
        1.0       0.49      0.71      0.58       770

avg / total       0.68      0.63      0.64      2161

Test Error = 0.370198981953
Accuracy: 62.9801018047


Confusion matrix Grid Search LR
[[818 573]
 [227 543]]
```



Review

- When evaluating models, always split into a testing and training dataset
- Accuracy can be a very misleading measure, especially when errors do not have equal weight
- `spark.ml` is a higher level API for constructing ML pipelines
- To tune a model, we can grid search over possible parameter values

Next Up: Deploying Models