Training better models using Automated Machine Learning

Vlad Iliescu

Tools of the trade

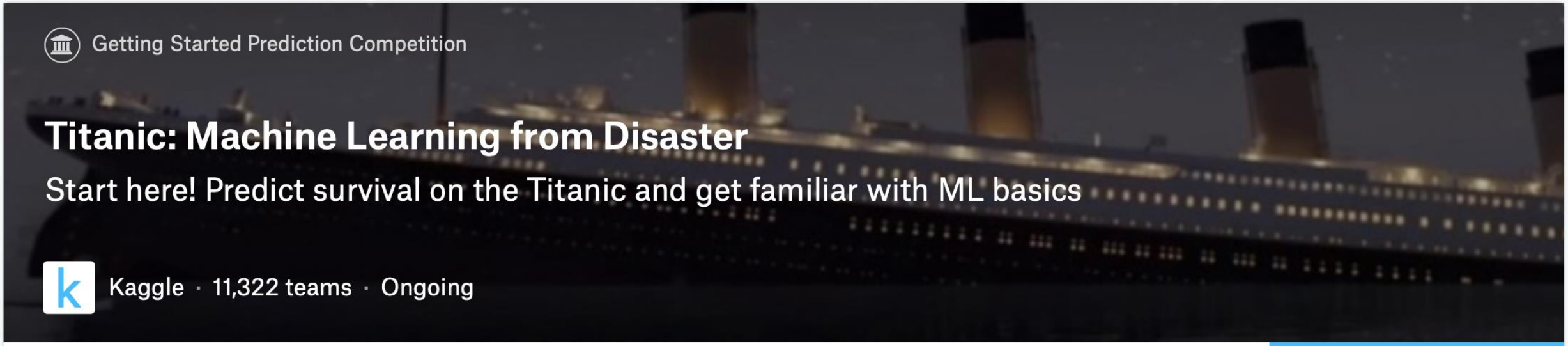
Google’s AUTOML

Auto-keras

Auto-SKLearn

Microsoft’s automated ML

Kaggle



Overview Data Kernels Discussion Leaderboard Rules Team My Submissions **Submit Predictions**

Overview

The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

Features

- PassengerId
- Survived
- Pclass
- Sex
- Age
- Sibsp
- Parch
- Ticket
- Fare
- Cabin
- Embarked

```python
import pandas as pd
from azureml.core.experiment import Experiment
from azureml.core.workspace import Workspace
from azureml.train.automl import AutoMLConfig
```

```python
# Load training data
df = pd.read_csv('./data/train.csv')

X = df.drop(['Survived', 'PassengerId'], axis=1)
y = df['Survived']
```

```
# Configure the experiment

config = AutoMLConfig(
    task='classification',
    primary_metric=AUC_weighted',
    iterations=5,
    X=X,
    y=y.values.flatten(),
    n_cross_validations=5,
    preprocess=True
)
```

```python
# Configure the experiment

config = AutoMLConfig(
    task='classification',
    primary_metric=AUC_weig
    iterations=10,
    X=X,
    y=y.values.flatten(),
    n_cross_validations=5,
    preprocess=True
)
```

- Classification
- Regression
- Forecasting

```python
# Configure the experiment

config = AutoMLConfig(
    task='classification',
    primary_metric=AUC_weighted',
    iterations=5,
    X=X,
    y=y.values.flatten(),
    n_cross_validations=5,
    preprocess=True
)
```

- AUC_Macro
- AUC_Micro
- AUC_Weighted
- accuracy
- average_precision_score_macro
- average_precision_score_micro
- average_precision_score_weighted
- balanced_accuracy
- f1_score_macro
- f1_score_micro
- f1_score_weighted
- log_loss
- norm_macro_recall
- precision_score_macro
- precision_score_micro
- precision_score_weighted
- recall_score_macro
- recall_score_micro
- recall_score_weighted
- weighted_accuracy

```
# Configure the experiment

config = AutoMLConfig(
    task='classification',
    primary_metric=AUC_weighted',
    iterations=5,
    X=X,
    y=y.values.flatten(),
    n_cross_validations=5,
    preprocess=True
)
```

```python
# Configure the experiment

config = AutoMLConfig(
    task='classification',
    primary_metric=AUC_weighted',
    iterations=5,
    X=X,
    y=y.values.flatten(),
    n_cross_validations=5,
    preprocess=True
)
```

```
# Configure the experiment

config = AutoMLConfig(
    task='classification',
    primary_metric=AUC_weighted',
    iterations=5,
    X=X,
    y=y.values.flatten(),
    n_cross_validations=5,
    preprocess=True
)
```

```python
# Configure the experiment

config = AutoMLConfig(
    task='classification',
    primary_metric=AUC_weighted',
    iterations=5,
    X=X,
    y=y.values.flatten(),
    n_cross_validations=5,
    preprocess=True
)
```

```
# Configure the experiment

config = AutoMLConfig(
    task='classification',
    primary_metric=AUC_weighted',
    iterations=5,
    X=X,
    y=y.values.flatter
    n_cross_validatior
    preprocess=False
)
```

- StandardScaleWrapper
- MinMaxScalar
- MaxAbsScaler
- RobustScalar
- PCA
- TruncatedSVDWrapper
- SparseNormalizer

```
# Configure the experiment

config = AutoMLConfig(
    task='classification',
    primary_metric=AUC_weighted',
    iterations=5,
    X=X,
    y=y.values.flatter
    n_cross_validatior
    preprocess=True
)
```

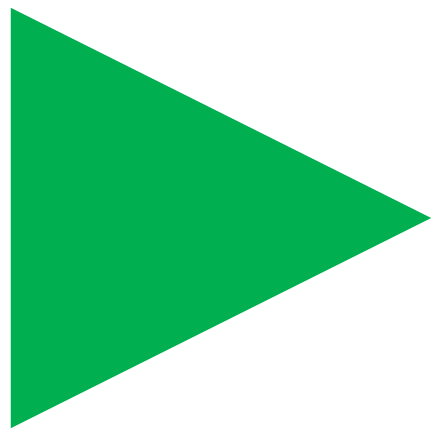
- StandardScaleWrapper
- MinMaxScalar
- MaxAbsScaler
- RobustScalar
- PCA
- TruncatedSVDWrapper
- SparseNormalizer

+

- Drop high cardinality or no variance features
- Impute missing values
- Generate additional features
- Transform and encode
- Word embeddings
- Target encodings
- Text target encoding
- Weight of Evidence (WoE)
- Cluster Distance

```python
# Run the Automated ML experiment

ws = Workspace.from_config()
experiment = Experiment(ws, 'NDR2019')

local_run = experiment.submit(config, show_output=True)
```

```
# Run the Automated ML experiment

ws = Workspace.from_config()
experiment = Experiment(ws, 'NDR2019')

local_run = experiment.submit(config, show_output=True)
```



```
Running on local machine
Parent Run ID: AutoML_c2789ab8-9f3a-45bb-85a6-a96f0cb64e65
Current status: DatasetFeaturization. Beginning to featurize the dataset.
Current status: DatasetEvaluation. Gathering dataset statistics.
Current status: FeaturesGeneration. Generating features for the dataset.
Current status: DatasetFeaturizationCompleted. Completed featurizing the dataset.
Current status: DatasetCrossValidationSplit. Generating individually featurized CV splits.
Current status: ModelSelection. Beginning model selection.
```

ITERATION	PIPELINE	DURATION	METRIC	BEST
0	MaxAbsScaler LightGBM	0:00:14	0.8526	0.8526
1	MaxAbsScaler SGD	0:00:12	0.8519	0.8526
2	StandardScalerWrapper LightGBM	0:00:12	0.8591	0.8591
3	VotingEnsemble	0:00:23	0.8615	0.8615
4	StackEnsemble	0:00:22	0.8605	0.8615

ITERATION	PIPELINE	DURATION	METRIC	BEST
0	MaxAbsScaler LightGBM	0:00:14	0.8526	0.8526
1	MaxAbsScaler SGD	0:00:12	0.8519	0.8526
2	StandardScalerWrapper LightGBM	0:00:12	0.8591	0.8591
3	VotingEnsemble	0:00:23	0.8615	0.8615
4	StackEnsemble	0:00:22	0.8605	0.8615

```python
# Run the best model on the test dataset
best_run, fitted_model = local_run.get_output()

df_test = pd.read_csv('./data/test.csv')
df_test_X = df_test.drop(['PassengerId'], axis=1)
pred_y = fitted_model.predict(df_test_X)
```

```
# Generate the submission file
output = pd.DataFrame({
        'PassengerId': df_test['PassengerId’],
        'Survived': pred_y
        })

output.to_csv('./output.csv', index=False)
```

Make Submission

Top 19%

2225	**Daniel Tham**		0.79425	2	13m
2226	**fulgerica**		0.79425	1	now

Your Best Entry ↑

Your submission scored 0.79425, which is not an improvement of your best score. Keep trying!

2227	**Luis Riveros**		0.78947	6	2mo

Turning it up to 11

```python
# Configure the experiment

config = AutoMLConfig(
    task='classification',
    primary_metric='AUC_weighted',
    iterations=50,
    X=X,
    y=y.values.flatten(),
    n_cross_validations=5,
    preprocess=True
)
```

1	MaxAbsScaler SGD	0:00:11	0.8515	0.8526
2	StandardScalerWrapper LightGBM	0:00:12	0.8591	0.8591
3	MaxAbsScaler RandomForest	0:00:11	0.8269	0.8591
4	MaxAbsScaler SGD	0:00:11	0.8321	0.8591
5	MaxAbsScaler ExtremeRandomTrees	0:00:11	0.7098	0.8591
6	MaxAbsScaler SGD	0:00:11	0.8480	0.8591
7	StandardScalerWrapper ExtremeRandomTrees	0:00:14	0.8356	0.8591
8	MaxAbsScaler RandomForest	0:00:13	0.8106	0.8591
9	MaxAbsScaler LightGBM	0:00:12	0.8669	0.8669
10	MaxAbsScaler BernoulliNaiveBayes	0:00:11	0.8380	0.8669
11	MaxAbsScaler LightGBM	0:00:11	0.8467	0.8669
12	MaxAbsScaler ExtremeRandomTrees	0:00:12	0.8537	0.8669
13	MaxAbsScaler ExtremeRandomTrees	0:00:12	0.8573	0.8669
14	MaxAbsScaler SGD	0:00:11	0.8310	0.8669
15	MaxAbsScaler LightGBM	0:00:13	0.8640	0.8669
16	StandardScalerWrapper BernoulliNaiveBayes	0:00:11	0.7009	0.8669
17	MaxAbsScaler RandomForest	0:00:12	0.7151	0.8669
18	StandardScalerWrapper LightGBM	0:00:14	0.8548	0.8669
19	SparseNormalizer LightGBM	0:00:14	0.8589	0.8669
20	MaxAbsScaler LightGBM	0:00:16	0.8490	0.8669
21	TruncatedSVDWrapper ExtremeRandomTrees	0:00:24	0.7959	0.8669

2225	**Daniel Tham**		0.79425	2	34m
2226	**fulgerica**		0.79425	2	~10s

Your Best Entry ↑

Your submission scored 0.78947, which is not an improvement of your best score. Keep trying!

2227	**Luis Riveros**		0.78947	6	2mo
2228	**Chen Tan**		0.78947	2	2mo
2229	**aaaaakiyama**		0.78947	2	2mo

```python
# Configure the experiment

config = AutoMLConfig(
    task='classification',
    primary_metric='accuracy',
    iterations=10,
    X=X,
    y=y.values.flatten(),
    n_cross_validations=5,
    preprocess=True
)
```

ITERATION	PIPELINE	DURATION	METRIC	BEST
0	MaxAbsScaler SGD	0:00:11	0.8024	0.8024
1	MaxAbsScaler SGD	0:00:11	0.8024	0.8024
2	MaxAbsScaler ExtremeRandomTrees	0:00:11	0.6017	0.8024
3	MaxAbsScaler SGD	0:00:11	0.8092	0.8092
4	MaxAbsScaler RandomForest	0:00:11	0.7094	0.8092
5	MaxAbsScaler SGD	0:00:14	0.8035	0.8092
6	MaxAbsScaler RandomForest	0:00:11	0.7890	0.8092
7	StandardScalerWrapper RandomForest	0:00:11	0.7084	0.8092
8	VotingEnsemble	0:00:27	0.8148	0.8148
9	StackEnsemble	0:00:27	0.8114	0.8148

What just happened?

Questions?

The END

Vlad Iliescu