


 Search

Competitions Datasets Kernels Discussion Learn

 **Applying LightGBM to Titanic dataset**  
Python notebook using data from [Titanic: Machine Learning from Disaster](#) · 1,125 views

^8

Fork 12

...

- Version 11  
11 commits
- Notebook
- Data
- Log
- Comments

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import lightgbm as lgbm
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, recall_score
, precision_score, f1_score
%matplotlib inline
```

## An implementation of LightGBM for the Titanic problem

Load and check the data:

In [2]:

```
train_df = pd.read_csv('../input/train.csv')
test_df = pd.read_csv('../input/test.csv')
train_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass        891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age           714 non-null float64
SibSp         891 non-null int64
Parch         891 non-null int64
Ticket        891 non-null object
Fare          891 non-null float64
Cabin         204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

Feature engineering

- Add any extra features that may be handy
- Encode and categoricals stores as strings
- Drop features that look difficult to get use out of

Want to make sure that everything I do to the training set happens to the test, so makes sense to run as a loop

In [3]:

```
# Not sure passenger ID is useful as a feature, but need
to save it from the test set for the submission
test_passenger_ids = test_df.pop('PassengerId')
train_df.drop(['PassengerId'], axis=1, inplace=True)

# 'Embarked' is stored as letters, so fit a label encoder
to the train set to use in the loop
embarked_encoder = LabelEncoder()
embarked_encoder.fit(train_df['Embarked'].fillna('Null'
))

# Dataframes to work on
df_list = [train_df, test_df]

for df in df_list:

    # Record anyone travelling alone
    df['Alone'] = (df['SibSp'] == 0) & (df['Parch'] == 0
)

    # Transform 'Embarked'
    df['Embarked'].fillna('Null', inplace=True)
    df['Embarked'] = embarked_encoder.transform(df['Emba
rked'])

    # Transform 'Sex'
    df.loc[df['Sex'] == 'female', 'Sex'] = 0
    df.loc[df['Sex'] == 'male', 'Sex'] = 1
    df['Sex'] = df['Sex'].astype('int8')

    # Drop features that seem unusable. Save passenger id
s if test
    df.drop(['Name', 'Ticket', 'Cabin'], axis=1, inplace
=True)
```

Prep the training set for learning

In [4]:

```
# Separate the label
```

```

y = train_df.pop('Survived')

# Take a hold out set randomly
X_train, X_test, y_train, y_test = train_test_split(train_df, y, test_size=0.2, random_state=42)

# Create an LGBM dataset for training
categorical_features = ['Alone', 'Sex', 'Pclass', 'Embarked']
train_data = lgbm.Dataset(data=X_train, label=y_train, categorical_feature=categorical_features, free_raw_data=False)

# Create an LGBM dataset from the test
test_data = lgbm.Dataset(data=X_test, label=y_test, categorical_feature=categorical_features, free_raw_data=False)

# Finally, create a dataset for the FULL training data to give us maximum amount of data to train on after performance has been calibrated
final_train_set = lgbm.Dataset(data=train_df, label=y, categorical_feature=categorical_features, free_raw_data=False)

```

Define hyperparameters for LGBM

In [5]:

```

lgbm_params = {
    'boosting': 'dart',          # dart (drop out trees) often performs better
    'application': 'binary',     # Binary classification
    'learning_rate': 0.05,       # Learning rate, controls size of a gradient descent step
    'min_data_in_leaf': 20,      # Data set is quite small so reduce this a bit
    'feature_fraction': 0.7,     # Proportion of features in each boost, controls overfitting
    'num_leaves': 41,            # Controls size of tree since LGBM uses leaf wise splits
    'metric': 'binary_logloss',  # Area under ROC curve as the evaluation metric
    'drop_rate': 0.15
}

```

Train the model

Use the held out training data to evaluate performance and early stop

Use the hold out training data to evaluate performance and early stop to control overfitting

In [6]:

```
evaluation_results = {}
clf = lgbm.train(train_set=train_data,
                 params=lgbm_params,
                 valid_sets=[train_data, test_data],
                 valid_names=['Train', 'Test'],
                 evals_result=evaluation_results,
                 num_boost_round=500,
                 early_stopping_rounds=100,
                 verbose_eval=20
                )
optimum_boost_rounds = clf.best_iteration
```

```
/opt/conda/lib/python3.6/site-packages/lightgbm/basic.py:1036: UserWarning: Using categorical_feature in Dataset.
  warnings.warn('Using categorical_feature in Dataset.')
```

```
/opt/conda/lib/python3.6/site-packages/lightgbm/basic.py:681: UserWarning: categorical_feature in param dict is overridden.
  warnings.warn('categorical_feature in param dict is overridden.')
```

```
/opt/conda/lib/python3.6/site-packages/lightgbm/basic.py:681: UserWarning: categorical_feature in param dict is overridden.
  warnings.warn('categorical_feature in param dict is overridden.')
```

```
/opt/conda/lib/python3.6/site-packages/lightgbm/basic.py:681: UserWarning: categorical_feature in param dict is overridden.
  warnings.warn('categorical_feature in param dict is overridden.')
```

Training until validation scores don't improve for 100 rounds.

```
[20] Train's binary_logloss: 0.507372
Test's binary_logloss: 0.54198
[40] Train's binary_logloss: 0.433745
Test's binary_logloss: 0.479584
[60] Train's binary_logloss: 0.413204
Test's binary_logloss: 0.461935
[80] Train's binary_logloss: 0.393179
Test's binary_logloss: 0.444887
[100] Train's binary_logloss: 0.385953
Test's binary_logloss: 0.442896
[120] Train's binary_logloss: 0.373897
Test's binary_logloss: 0.438083
[140] Train's binary_logloss: 0.36198
Test's binary_logloss: 0.431068
[160] Train's binary_logloss: 0.365415
Test's binary_logloss: 0.433233
[180] Train's binary_logloss: 0.347998
Test's binary_logloss: 0.420761
[200] Train's binary_logloss: 0.333363
Test's binary_logloss: 0.415733
[220] Train's binary_logloss: 0.326614
Test's binary_logloss: 0.409778
[240] Train's binary_logloss: 0.314509
Test's binary_logloss: 0.404197
```

```

[260] Train's binary_logloss: 0.307755
Test's binary_logloss: 0.401085
[280] Train's binary_logloss: 0.296964
Test's binary_logloss: 0.399032
[300] Train's binary_logloss: 0.292589
Test's binary_logloss: 0.399077
[320] Train's binary_logloss: 0.286967
Test's binary_logloss: 0.399104
[340] Train's binary_logloss: 0.284326
Test's binary_logloss: 0.398559
[360] Train's binary_logloss: 0.287643
Test's binary_logloss: 0.400735
[380] Train's binary_logloss: 0.282073
Test's binary_logloss: 0.399788
[400] Train's binary_logloss: 0.276178
Test's binary_logloss: 0.399436
Early stopping, best iteration is:
[301] Train's binary_logloss: 0.291464
Test's binary_logloss: 0.397995

```

### Visualise training performance

In [7]:

```

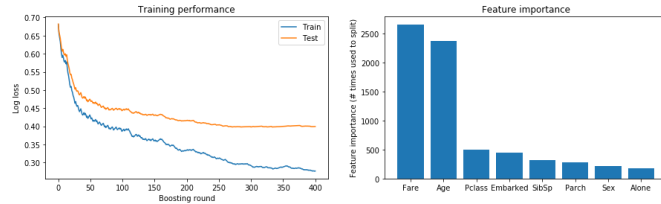
fig, axs = plt.subplots(1, 2, figsize=[15, 4])

# Plot the log loss during training
axs[0].plot(evaluation_results['Train']['binary_logloss'], label='Train')
axs[0].plot(evaluation_results['Test']['binary_logloss'], label='Test')
axs[0].set_ylabel('Log loss')
axs[0].set_xlabel('Boosting round')
axs[0].set_title('Training performance')
axs[0].legend()

# Plot feature importance
importances = pd.DataFrame({'features': clf.feature_name(),
                            'importance': clf.feature_importance()}).sort_values('importance', ascending=False)
axs[1].bar(x=np.arange(len(importances)), height=importances['importance'])
axs[1].set_xticks(np.arange(len(importances)))
axs[1].set_xticklabels(importances['features'])
axs[1].set_ylabel('Feature importance (# times used to split)')
axs[1].set_title('Feature importance')

plt.show()

```



## Examine model performance

Accuracy score can often be misleading for classifiers, so have a look at the precision and recall too

In [8]:

```
preds = np.round(clf.predict(X_test))
print('Accuracy score = \t {}'.format(accuracy_score(y_test, preds)))
print('Precision score = \t {}'.format(precision_score(y_test, preds)))
print('Recall score = \t {}'.format(recall_score(y_test, preds)))
print('F1 score = \t {}'.format(f1_score(y_test, preds)))
```

This kernel has been released under the [Apache 2.0](#) open source license.

Did you find this Kernel useful?

Show your appreciation with an upvote

8



Accuracy score = 0.8324022346368715

## Data

### Data Sources

▼ 🏆 Titanic: Machi...

... 418 x 2

t 418 x 11

t 891 x 12



## Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Last Updated: 7 years ago

### About this Competition

### Overview

The data has been split into two groups:

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

**The training set** should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each passenger. Your model will be based on “features” like passengers’

We also include **gender\_submission.csv**, a set of predictions that assume all and only female passengers survive, as an example of what a submission file should look like

Succeeded	True	Run Time	24.1 seconds
Exit Code	0	Queue Time	0 seconds
Docker Image Name	kaggle/python(Dockerfile)		
		Output Size	0
Timeout Exceeded	False	Used All Space	False
Failure Message			

**Download Log**

[illegible]



[illegible]

[illegible]

```
10.7s 38 [LightGBM] [Warning] No further splits with
```

[illegible]

```

positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
11.5s    53 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
11.5s    54 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
11.6s    55 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
11.6s    56 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
11.7s    57 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
11.7s    58 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
11.8s    59 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
11.8s    60 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
11.9s    61 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
11.9s    62 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
12.0s    63 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
12.0s    64 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
12.1s    65 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
12.1s    66 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
12.2s    67 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
12.2s    68 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
12.3s    69 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
12.3s    70 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
12.4s    71 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
12.4s    72 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
12.5s    73 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf

```

[illegible]


```
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
13.4s 91 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
13.4s 92 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
13.5s 93 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
13.5s 94 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
13.6s 95 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
13.6s 96 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
13.7s 97 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
13.7s 98 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
13.8s 99 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
13.8s 100 [LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with
positive gain, best gain: -inf
23.6s 259
23.6s 260 ...
23.6s 261 Complete. Exited with code 0.
```

Comments (3) 


Sort by

All Comments

Hotness




Click here to enter a comment...



楼上膜一下 • Posted on Latest Version • a year ago • Options • Reply 

1


It's a good kernel to help me learning usage of LightGBM and sklearn, much appreciate!



Ilya Khris... • Posted on Latest Version • 6 months ago • Options • Reply 

0

A good kernel to play with LGBM! Thank you!



Adriano ... • Posted on Latest Version • 10 days ago • Options • Reply 

0

Thanks. That's what I am looking for.

