Competitions overview

WINNING A KAGGLE COMPETITION IN PYTHON

Yauhen Babakhin
Kaggle Grandmaster

Instructor

Yauhen Babakhin

- Master's Degree in Applied Data Analysis
- 5 years of working experience in Data Science
- Kaggle competitions Grandmaster
- Gold medals in both classic Machine Learning and Deep Learning competitions

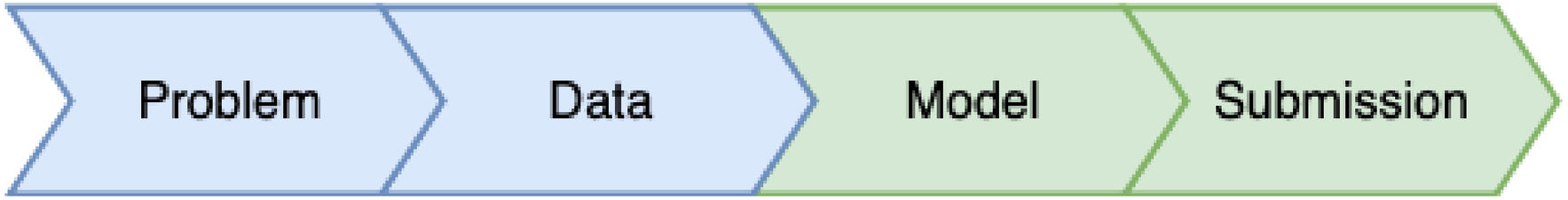

kaggle™

Kaggle benefits

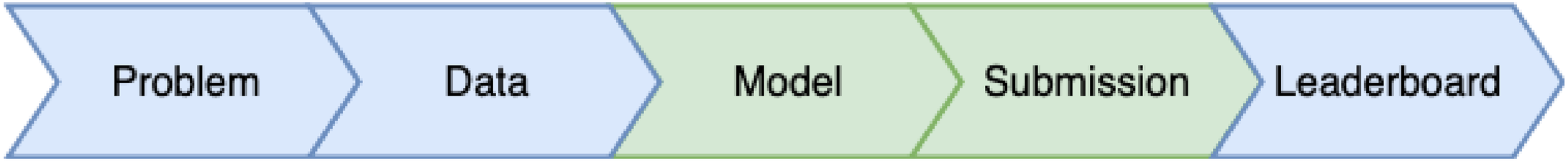
1. Get practical experience on the real-world data
2. Develop portfolio projects
3. Meet a great Data Science community
4. Try new domain or model type
5. Keep up-to-date with the best performing methods

Competition process

Competition process



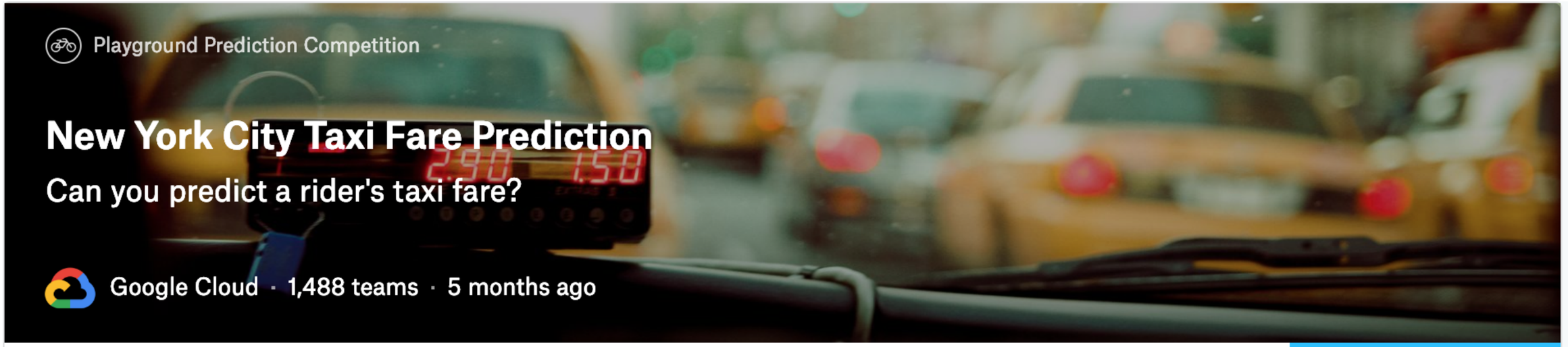
Competition process



How to participate

1. Go to **http://kaggle.com** website and select the competition
2. Download the data
3. Start building the models!

New York city taxi fare prediction



Overview | Data | Kernels | Discussion | Leaderboard | Rules | Join Competition

Overview

Description

Evaluation

Timeline

In this playground competition, hosted in partnership with Google Cloud and Coursera, you are tasked with predicting the fare amount (inclusive of tolls) for a taxi ride in New York City given the pickup and dropoff locations. While you can get a basic estimate based on just the distance between the two points, this will result in an RMSE of $5-$8, depending on the model used (see the starter code for an example

Train and Test data

```
import pandas as pd

# Read train data
taxi_train = pd.read_csv('taxi_train.csv')
taxi_train.columns.to_list()
```

```
['key',
 'fare_amount',
 'pickup_datetime',
 'pickup_longitude',
 'pickup_latitude',
 'dropoff_longitude',
 'dropoff_latitude',
 'passenger_count']
```

```
# Read test data
taxi_test = pd.read_csv('taxi_test.csv')
taxi_test.columns.to_list()
```

```
['key',
 'pickup_datetime',
 'pickup_longitude',
 'pickup_latitude',
 'dropoff_longitude',
 'dropoff_latitude',
 'passenger_count']
```

Sample submission

```
# Read sample submission
taxi_sample_sub = pd.read_csv('taxi_sample_submission.csv')
taxi_sample_sub.head()
```

```
                            key     fare_amount
0     2015-01-27 13:08:24.0000002     11.35
1     2015-01-27 13:08:24.0000003     11.35
2     2011-10-08 11:53:44.0000002     11.35
3     2012-12-01 21:12:12.0000002     11.35
4     2012-12-01 21:12:12.0000003     11.35
```

Let's practice!

WINNING A KAGGLE COMPETITION IN PYTHON

datacamp

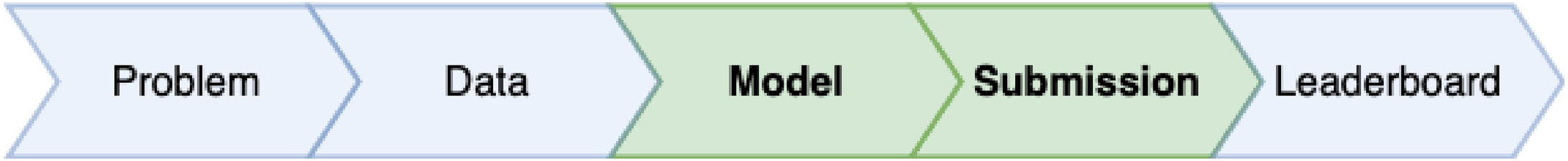
Prepare your first submission

WINNING A KAGGLE COMPETITION IN PYTHON

Yauhen Babakhin
Kaggle Grandmaster

datacamp

What is submission



New York city taxi fare prediction

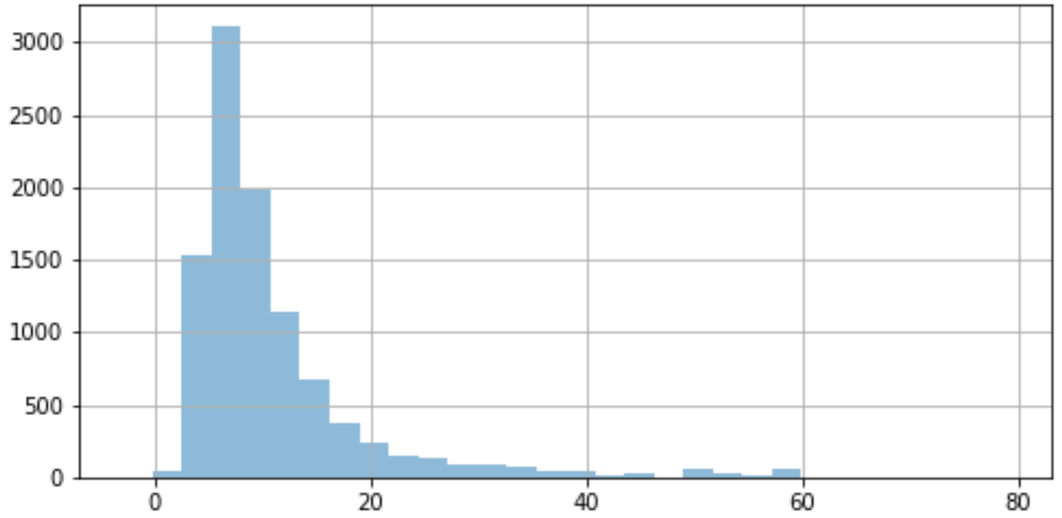
```
# Read train data
taxi_train = pd.read_csv('taxi_train.csv')
taxi_train.columns.to_list()
```

```
['key',
 'fare_amount',
 'pickup_datetime',
 'pickup_longitude',
 'pickup_latitude',
 'dropoff_longitude',
 'dropoff_latitude',
 'passenger_count']
```

Problem type

```
import matplotlib.pyplot as plt

# Plot a histogram
taxi_train.fare_amount.hist(bins=30, alpha=0.5)
plt.show()
```



Build a model

```
from sklearn.linear_model import LinearRegression


# Create a LinearRegression object
lr = LinearRegression()
```

```
# Fit the model on the train data
lr.fit(X=taxi_train[['pickup_longitude', 'pickup_latitude', 'dropoff_longitude',
                     'dropoff_latitude', 'passenger_count']],
       y=taxi_train['fare_amount'])
```

Predict on test set

```
# Select features
features = ['pickup_longitude', 'pickup_latitude',
            'dropoff_longitude', 'dropoff_latitude',
            'passenger_count']
```

```
# Make predictions on the test data
taxi_test['fare_amount'] = lr.predict(taxi_test[features])
```

Prepare submission

```
# Read a sample submission file
taxi_sample_sub = pd.read_csv('taxi_sample_submission.csv')
taxi_sample_sub.head(1)
```

```
                              key     fare_amount
0     2015-01-27 13:08:24.0000002     11.35
```

```
# Prepare a submission file
taxi_submission = taxi_test[['key', 'fare_amount']]

# Save the submission file as .csv
taxi_submission.to_csv('first_sub.csv', index=False)
```

Let's practice!

WINNING A KAGGLE COMPETITION IN PYTHON

datacamp

Public vs Private leaderboard

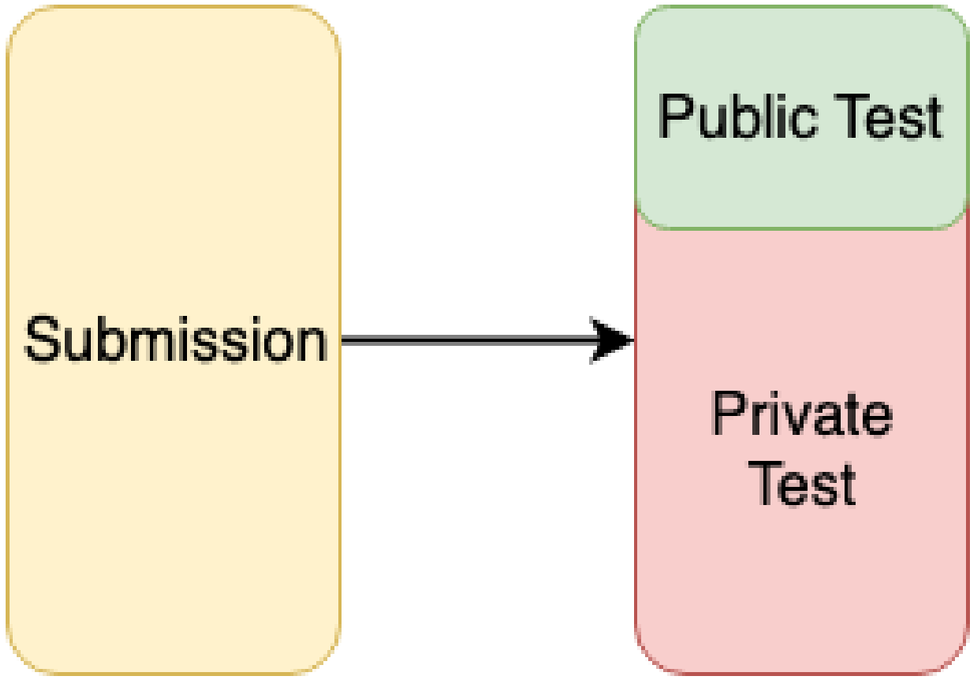
WINNING A KAGGLE COMPETITION IN PYTHON

Yauhen Babakhin
Kaggle Grandmaster

Competition metric

Evaluation metric	Type of problem
Area Under the ROC (AUC)	Classification
F1 Score (F1)	Classification
Mean Log Loss (LogLoss)	Classification
Mean Absolute Error (MAE)	Regression
Mean Squared Error (MSE)	Regression
Mean Average Precision at K (MAPK, MAP@K)	Ranking

Test split

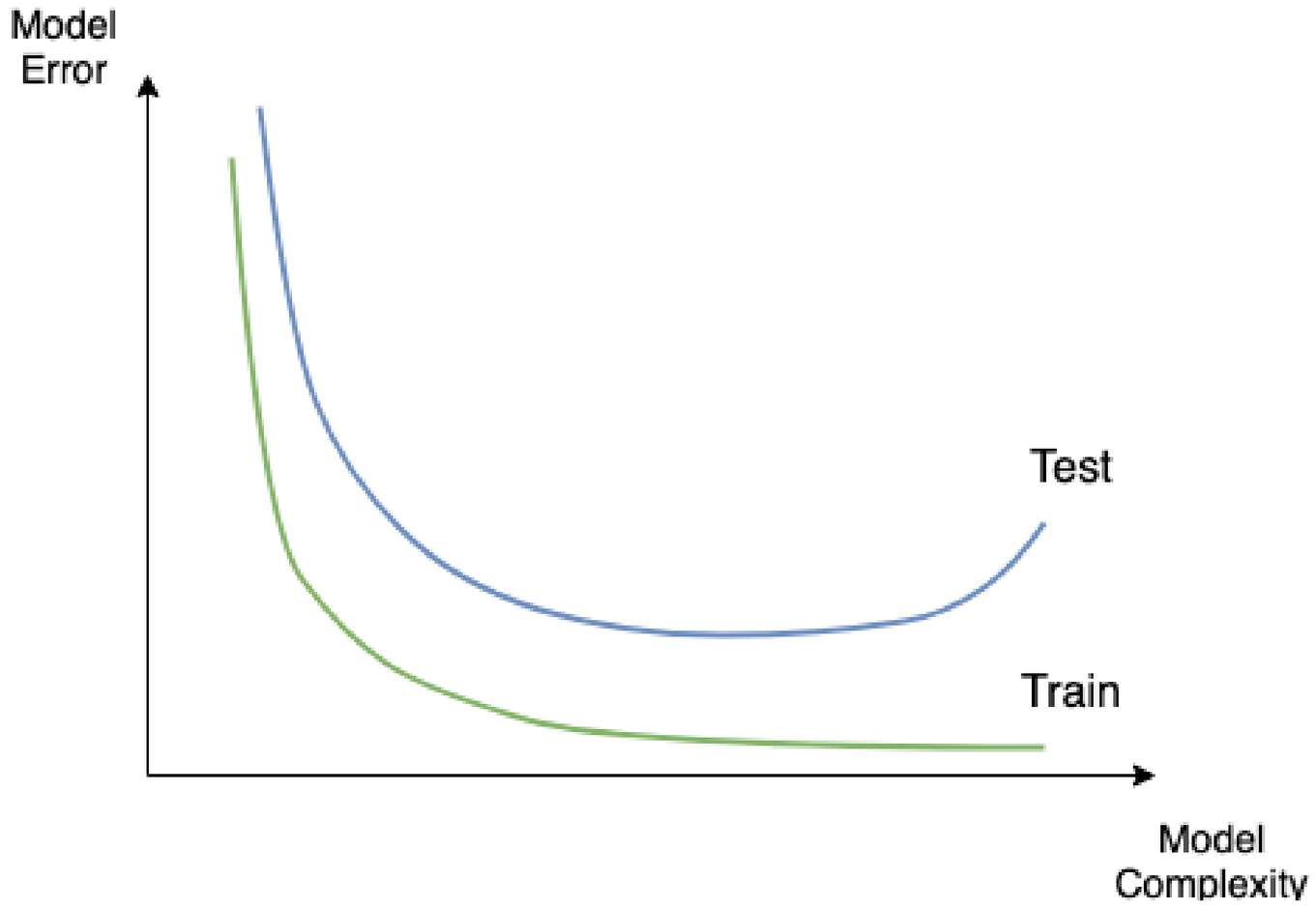


Leaderboards

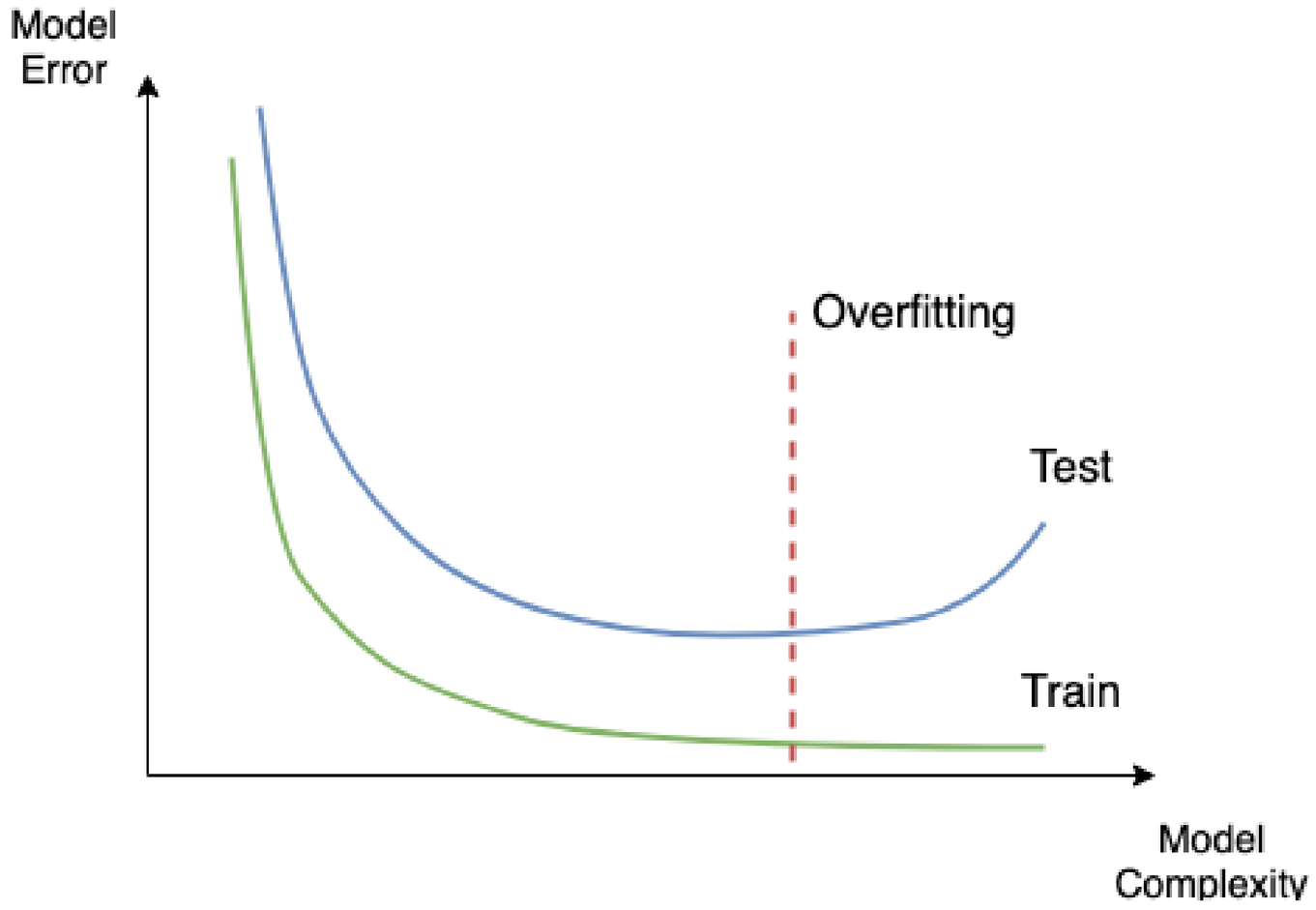
```
# Write a submission file to the disk
submission[['id', 'target']].to_csv('submission_1.csv', index=False)
```

Submission	Public LB MSE	Private LB MSE
submission_1.csv	2.895	?

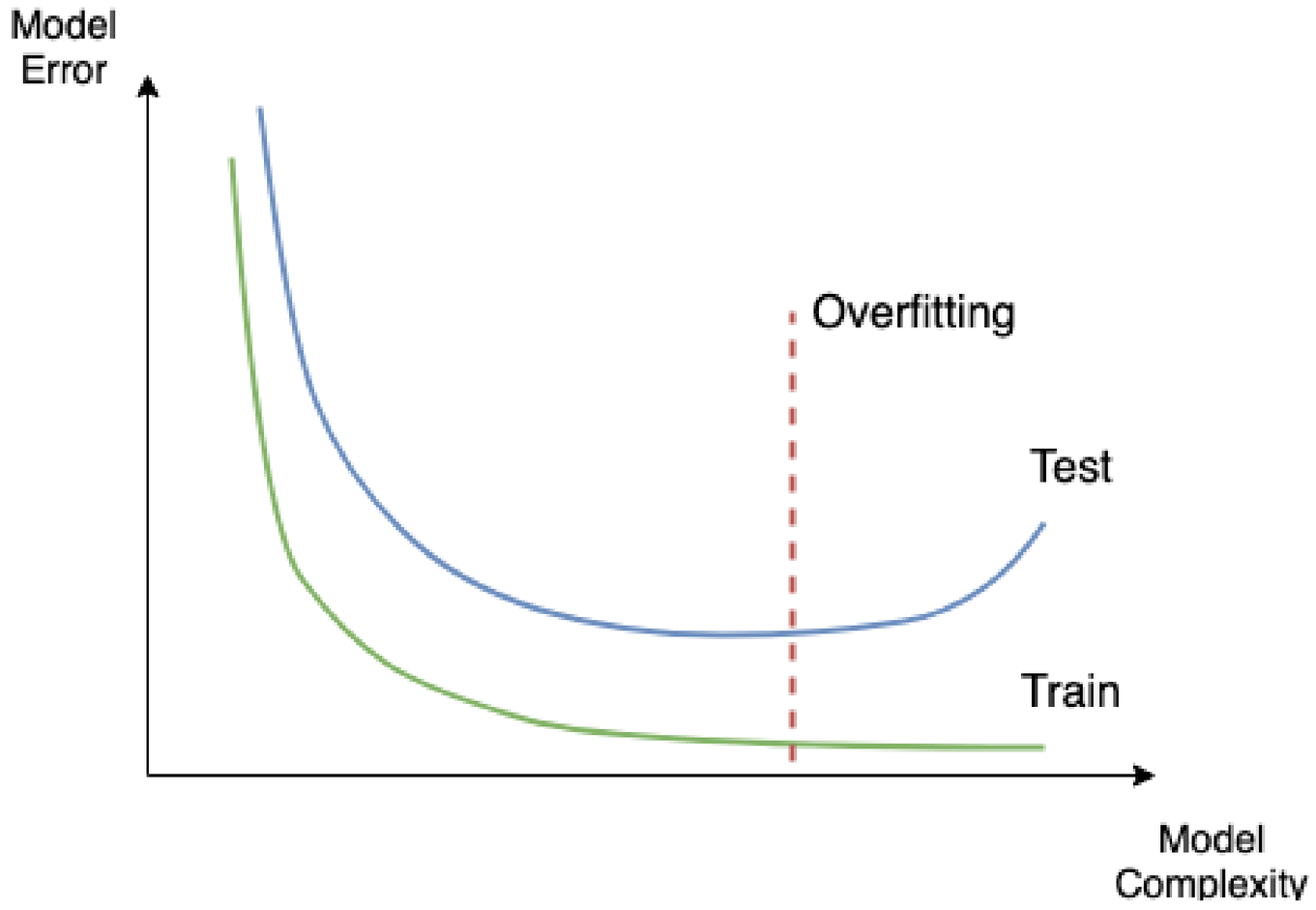
Overfitting



Overfitting



Overfitting



Public vs Private leaderboard shake-up

#	△pub	Team Name
1	—	Kyle Boone
2	▲ 2	Mike & Silogram
3	▼ 1	Major Tom
4	▼ 1	AhmetErdem
5	—	SKZ Lost in Translation
6	▲ 2	Stefan Stefanov
7	▲ 3	hklee
8	▼ 1	rapids.ai
9	▼ 3	Three Musketeers
10	▲ 3	J&J

#	△pub	Team Name
1	▲ 1484	gmobaz
2	▲ 414	RHINODAVEB
3	▲ 1784	Jayden Tan
4	▲ 1599	mchahhou
5	▲ 2753	R.elsharawy
6	▲ 1132	DDgg
7	▲ 772	Maverix
8	▲ 115	dil-bert
9	▲ 213	zr17
10	▲ 1211	KG123

Let's practice!

WINNING A KAGGLE COMPETITION IN PYTHON

datacamp