Metrics optimization

Lesson overview

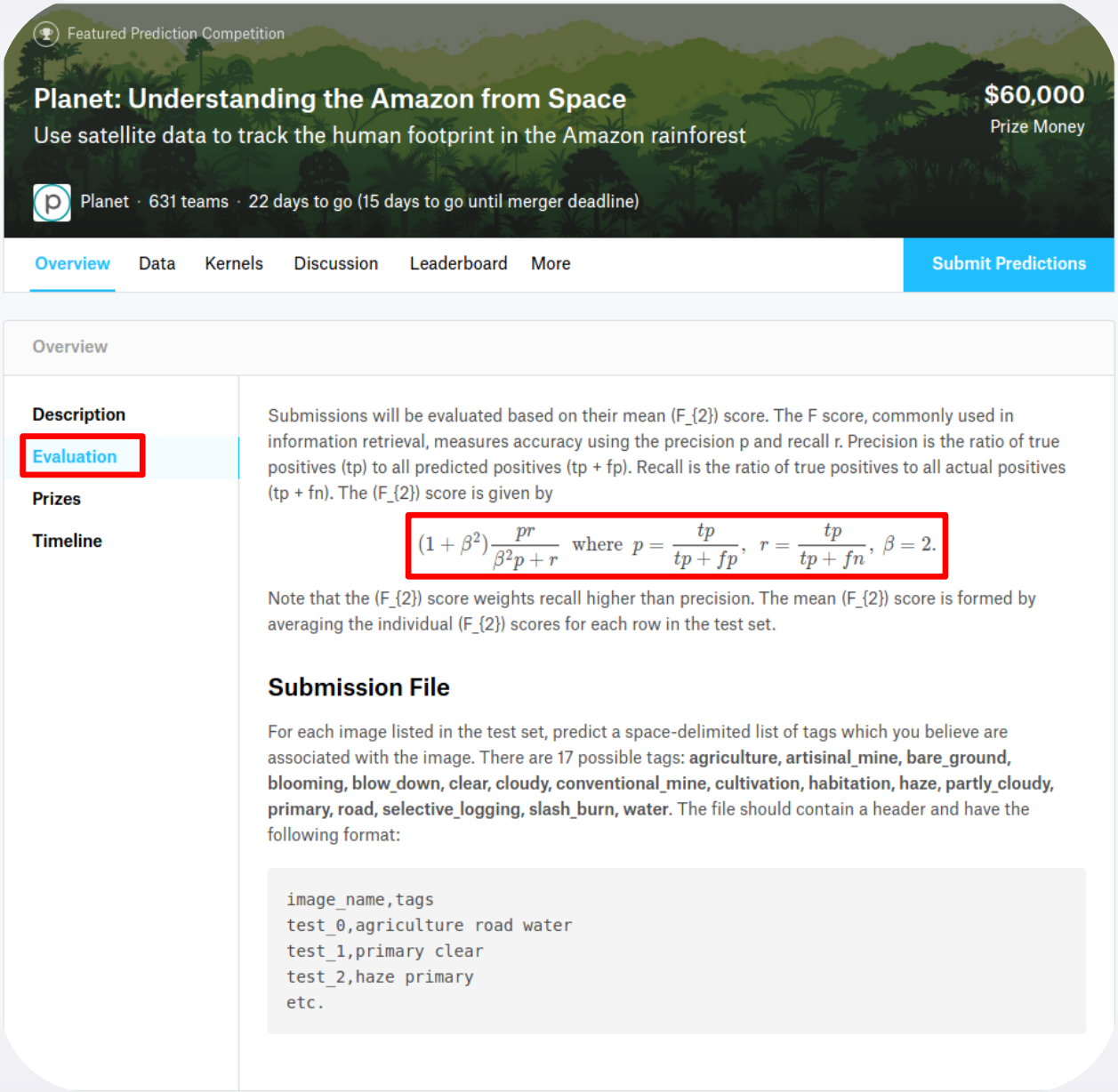
In this video:

- **Metrics:**
  - Why there are so many
  - Why should we care about them in competitions

In the following videos:

- **Loss versus metric**
- **Review the most important metrics**
  - For classification and regression tasks
  - Discuss baseline solutions for their optimization
- **Optimization techniques for the metrics**

Metrics



Featured Prediction Competition

Planet: Understanding the Amazon from Space

Use satellite data to track the human footprint in the Amazon rainforest

$60,000
Prize Money

Planet · 631 teams · 22 days to go (15 days to go until merger deadline)

Overview | Data | Kernels | Discussion | Leaderboard | More | Submit Predictions

Overview

- **Description**
- **Evaluation**
- **Prizes**
- **Timeline**

Submissions will be evaluated based on their mean (F_{2}) score. The F score, commonly used in information retrieval, measures accuracy using the precision p and recall r. Precision is the ratio of true positives (tp) to all predicted positives (tp + fp). Recall is the ratio of true positives to all actual positives (tp + fn). The (F_{2}) score is given by

$$(1+\beta^2)\frac{pr}{\beta^2 p + r} \text{ where } p = \frac{tp}{tp+fp}, \quad r = \frac{tp}{tp+fn}, \quad \beta = 2.$$

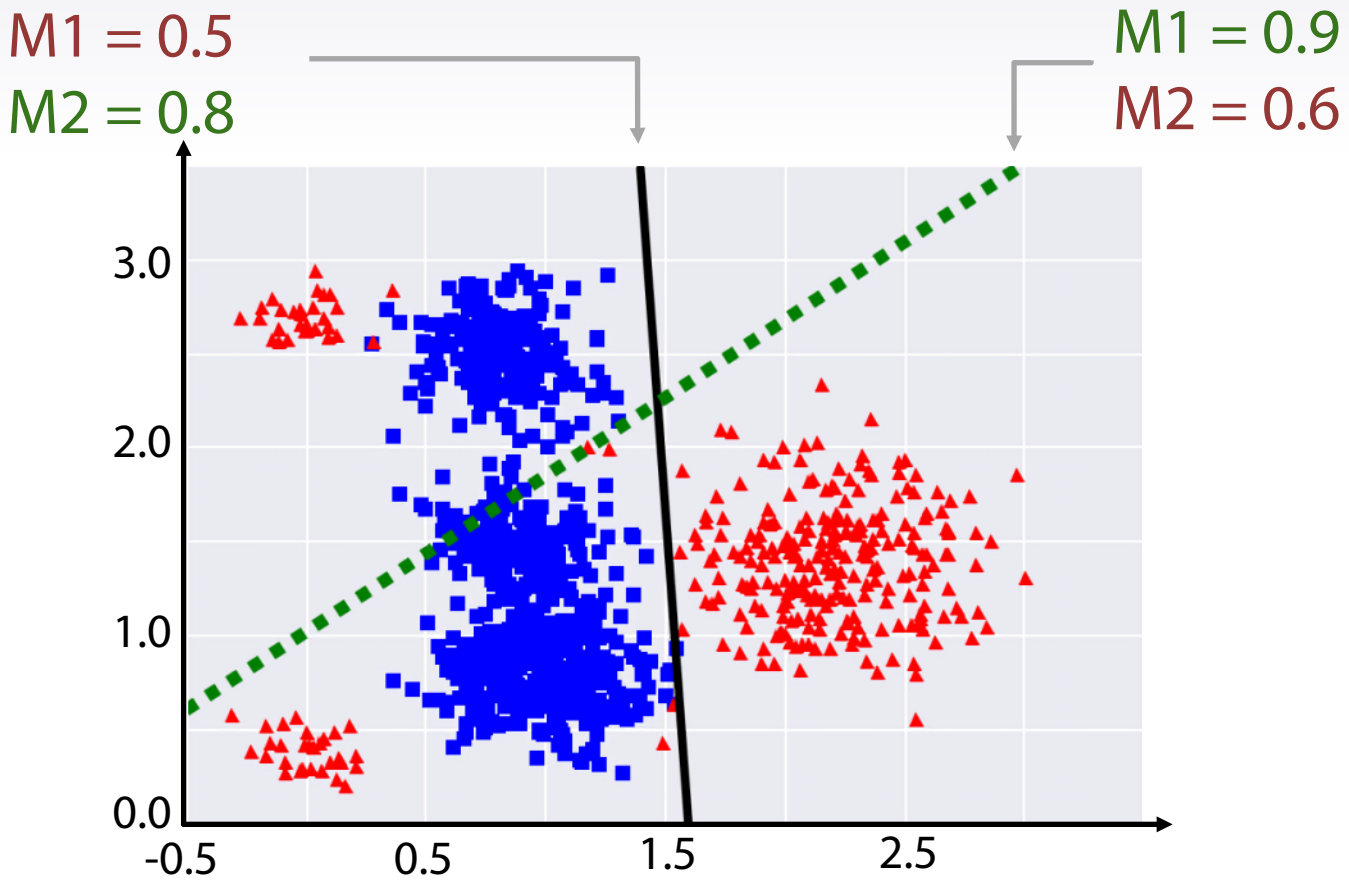
Note that the (F_{2}) score weights recall higher than precision. The mean (F_{2}) score is formed by averaging the individual (F_{2}) scores for each row in the test set.

Submission File

For each image listed in the test set, predict a space-delimited list of tags which you believe are associated with the image. There are 17 possible tags: **agriculture, artisinal_mine, bare_ground, blooming, blow_down, clear, cloudy, conventional_mine, cultivation, habitation, haze, partly_cloudy, primary, road, selective_logging, slash_burn, water**. The file should contain a header and have the following format:

```
image_name,tags
test_0,agriculture road water
test_1,primary clear
test_2,haze primary
etc.
```

Motivation



Chosen metric determines optimal decision boundary

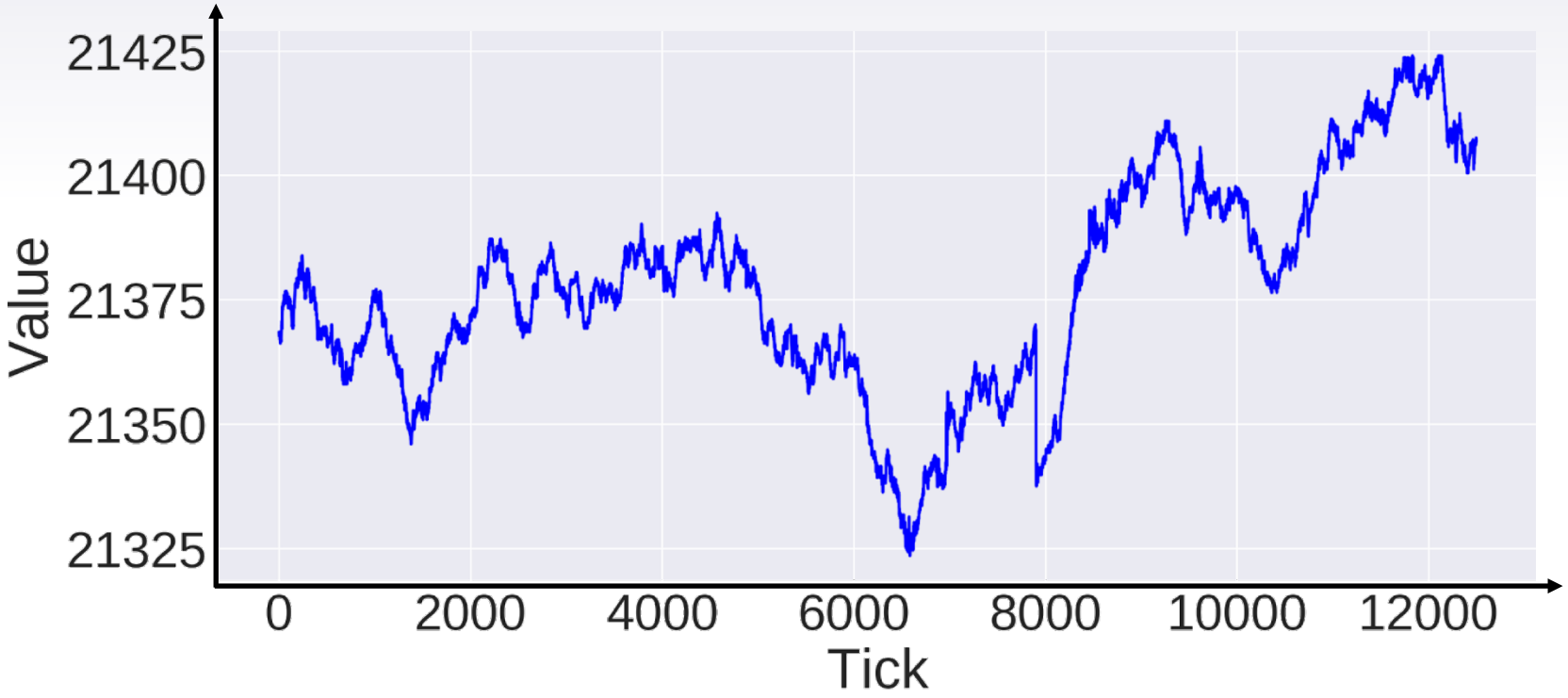
Image from "Scalable Learning of Non-Decomposable Objectives", Eban et al. AISTATS, 2017

Take-away point

If your model is scored with some metric, you get best results by optimizing exactly that metric

In the lesson about leaks, we'll discuss leader board probing. That is, we can check, for example, if the mean target value on public part of test set is the same as on train. If it's not, we would need to adapt our predictions to suit rest set better.

Motivation 2

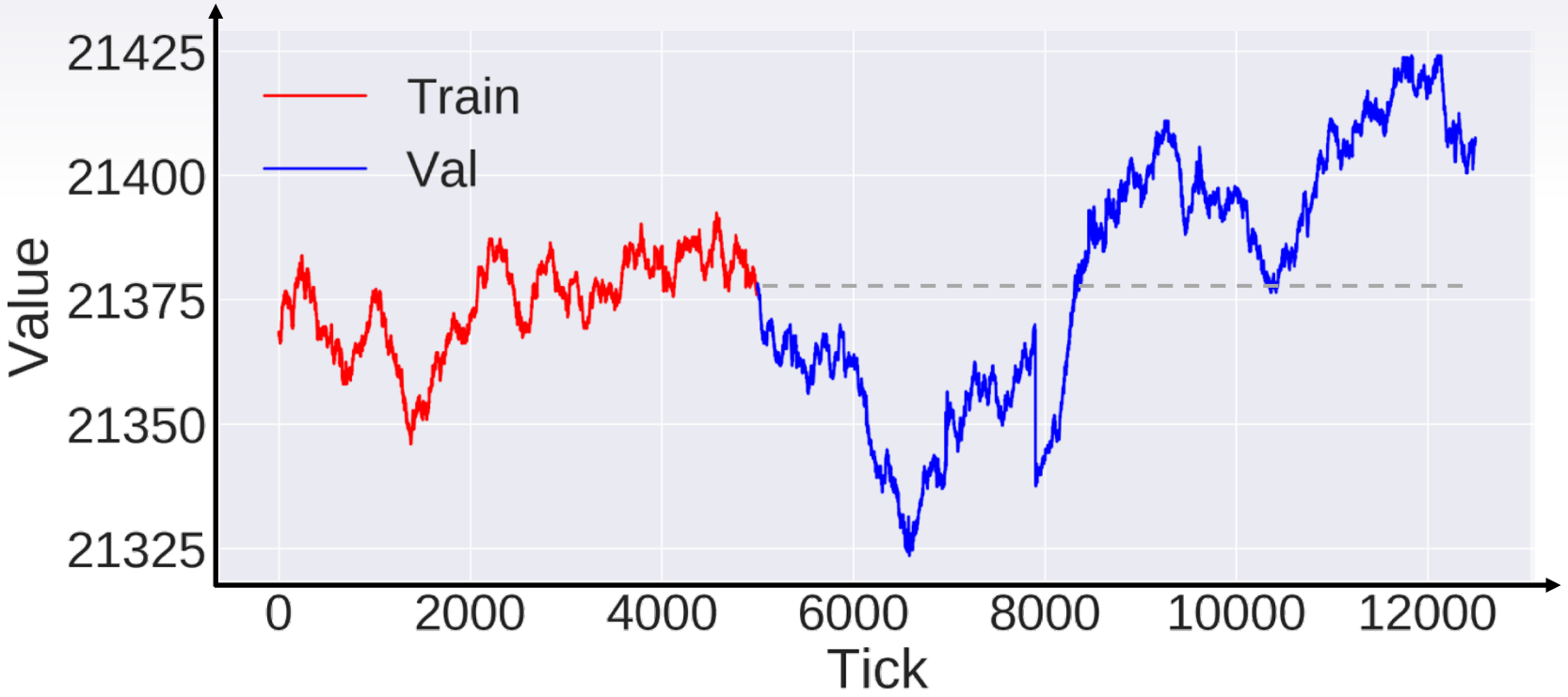


In particular, time series can be very challenging to forecast.

Even if you did a validation just right. Split by time, rolling windows, fill the distribution in the future can be much different from what we had in the train set.

Or sometimes, there's just not enough training data, so a model cannot capture the patterns.

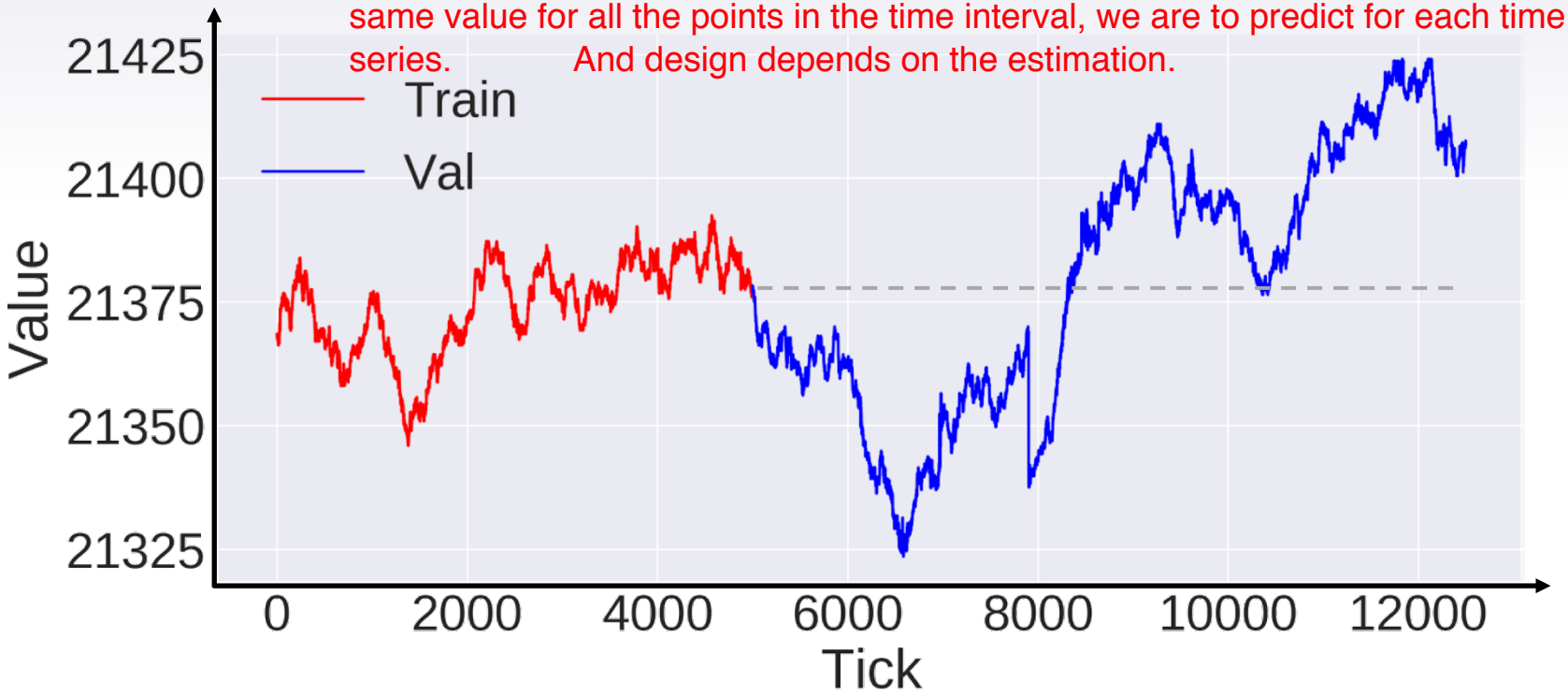
Motivation 2



$$Loss(\hat{y}_i; y_i) = \begin{cases} |y_i - \hat{y}_i|, & \text{if trend predicted correctly} \\ (y_i - \hat{y}_i)^2, & \text{if trend predicted incorrectly} \end{cases}$$

Motivation 2

So I realized that it would be much better to set all the predictions to either last value plus a very tiny constant, or last value minus very tiny constant. The same value for all the points in the time interval, we are to predict for each time series. And design depends on the estimation.



$$Loss(\hat{y}_i; y_i) = \begin{cases} |y_i - \hat{y}_i|, & \text{if trend predicted correctly} \\ (y_i - \hat{y}_i)^2, & \text{if trend predicted incorrectly} \end{cases}$$

Predict *trend* instead of the values:

Predict $y_{last} + 10^{-6}$

or $y_{last} - 10^{-6}$

Conclusion

- **Why there are so many metrics?**
  - Different metrics for different problems

- **Why should we care about metric in competitions?**
  - It is how the competitors are ranked!