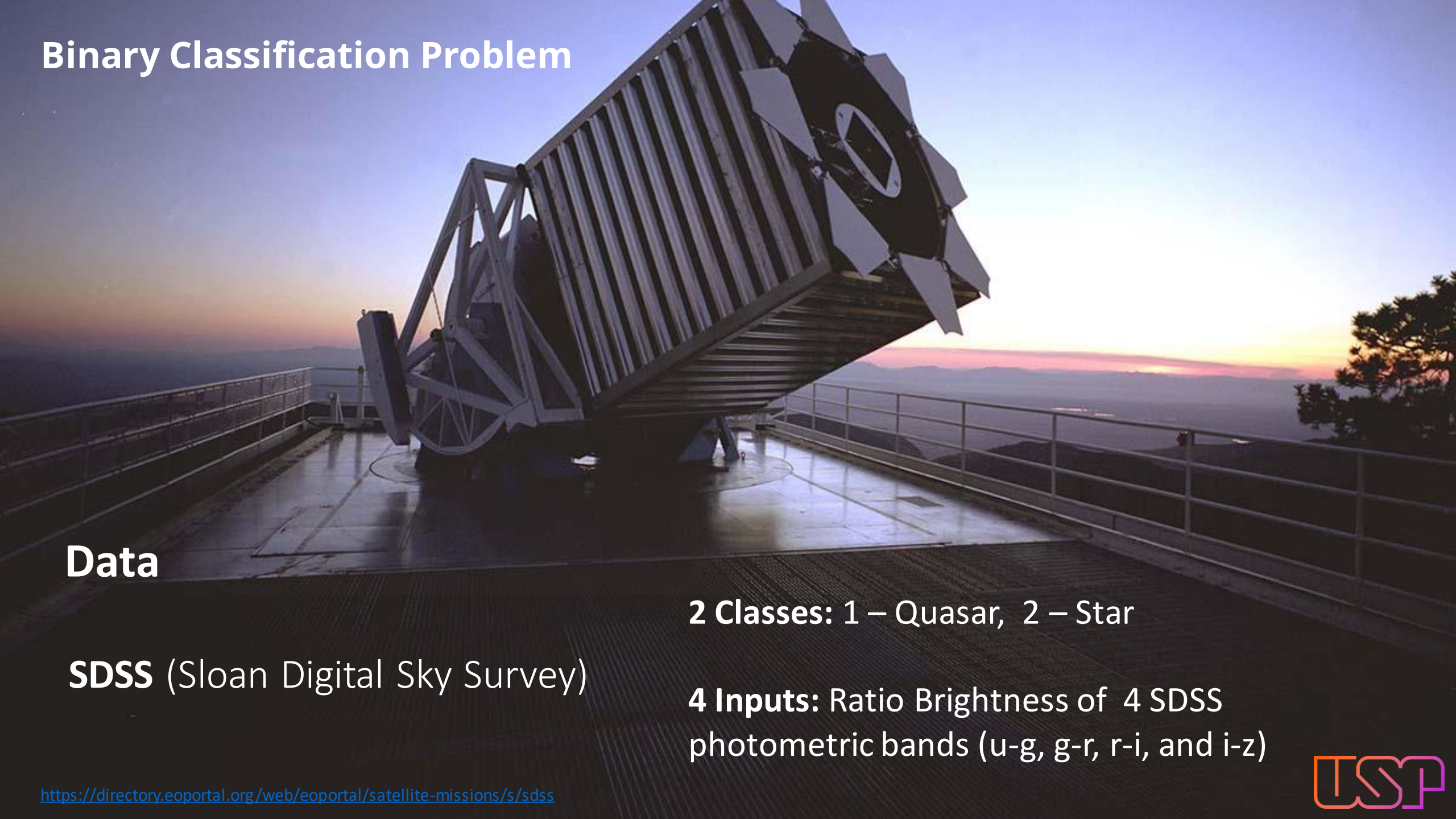
Machine Learning Model Evaluation Metrics for Classification

Data Analysis in Astronomy – AGA0505

Rodrigo da Motta Cabral de Carvalho

Binary Classification Problem



Data

SDSS (Sloan Digital Sky Survey)

2 Classes: 1 – Quasar, 2 – Star

4 Inputs: Ratio Brightness of 4 SDSS photometric bands (u-g, g-r, r-i, and i-z)

https://directory.eoportal.org/web/eoportal/satellite-missions/s/sdss

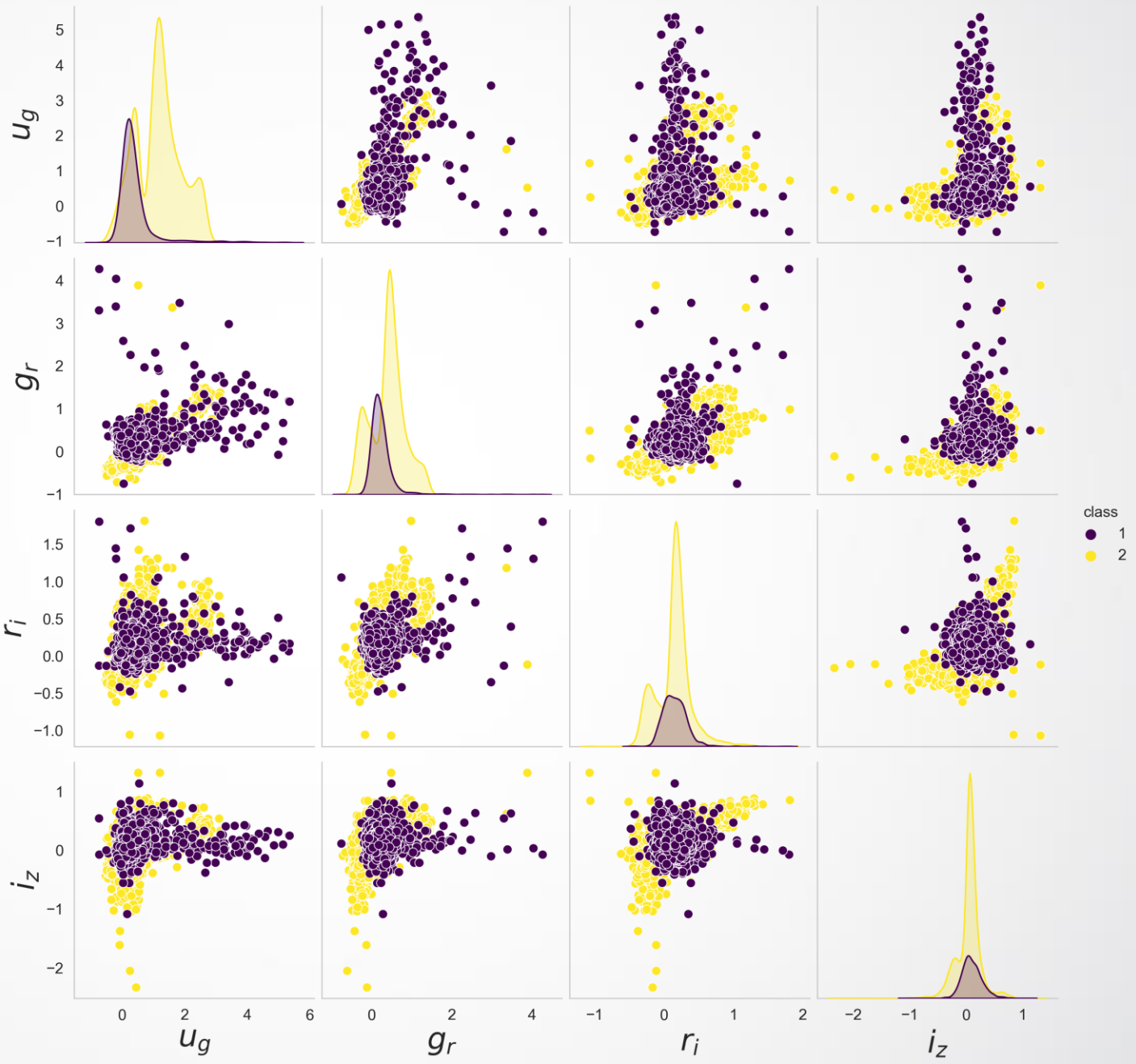
Data Distribution

7000 Star sample

2000 Quasar sample

Ratio 7:2
Usually a ratio bigger than 4:1 results in bias

(Changing Your Performance Metric, Resampling Your Dataset, Generate Synthetic Samples, Penalized Models)



Algorithm

Support Vector Machine (SVM)

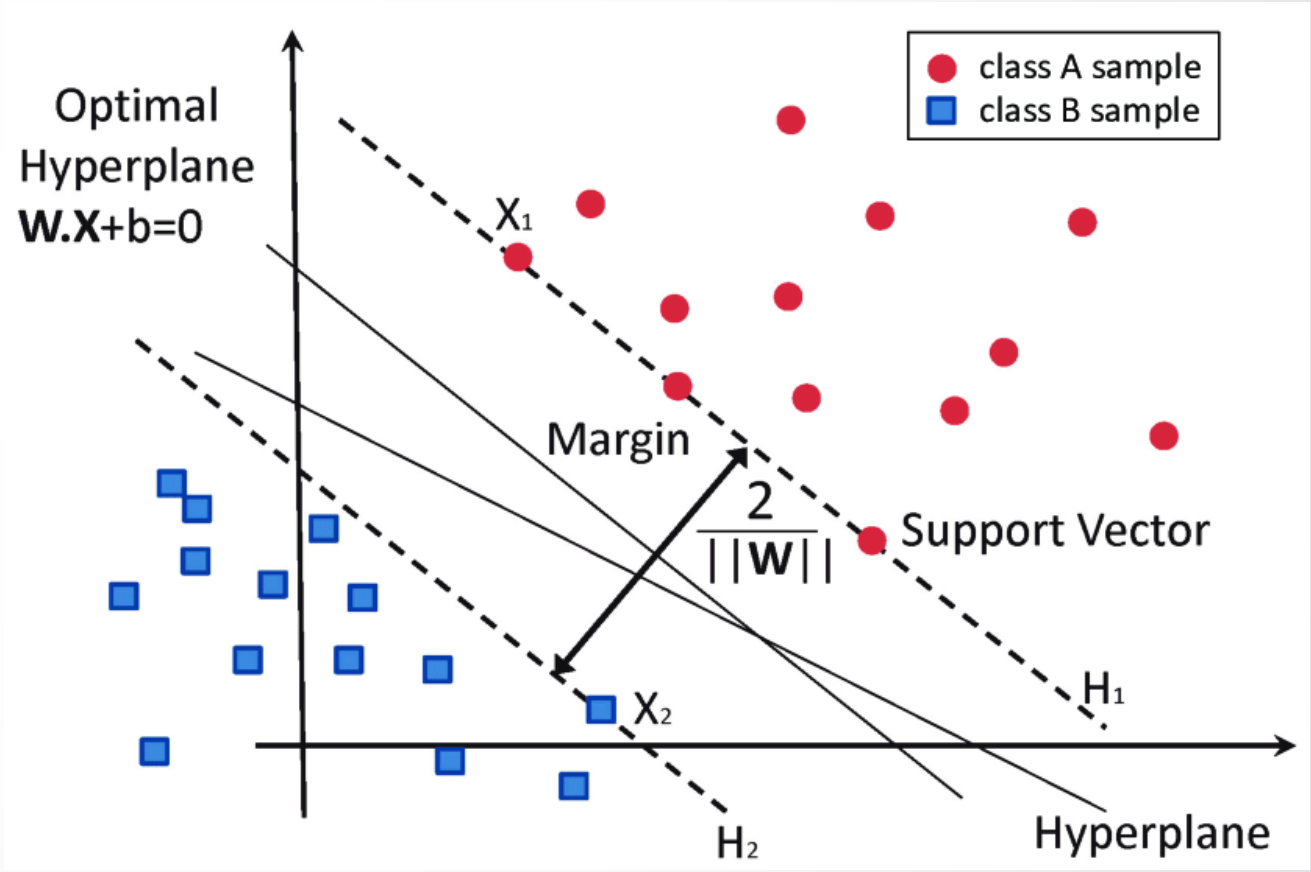
Find a hyperplane that separate our data

Penalty parameter **C**

$$J(\mathbf{w}) + C||\mathbf{w}||^2$$

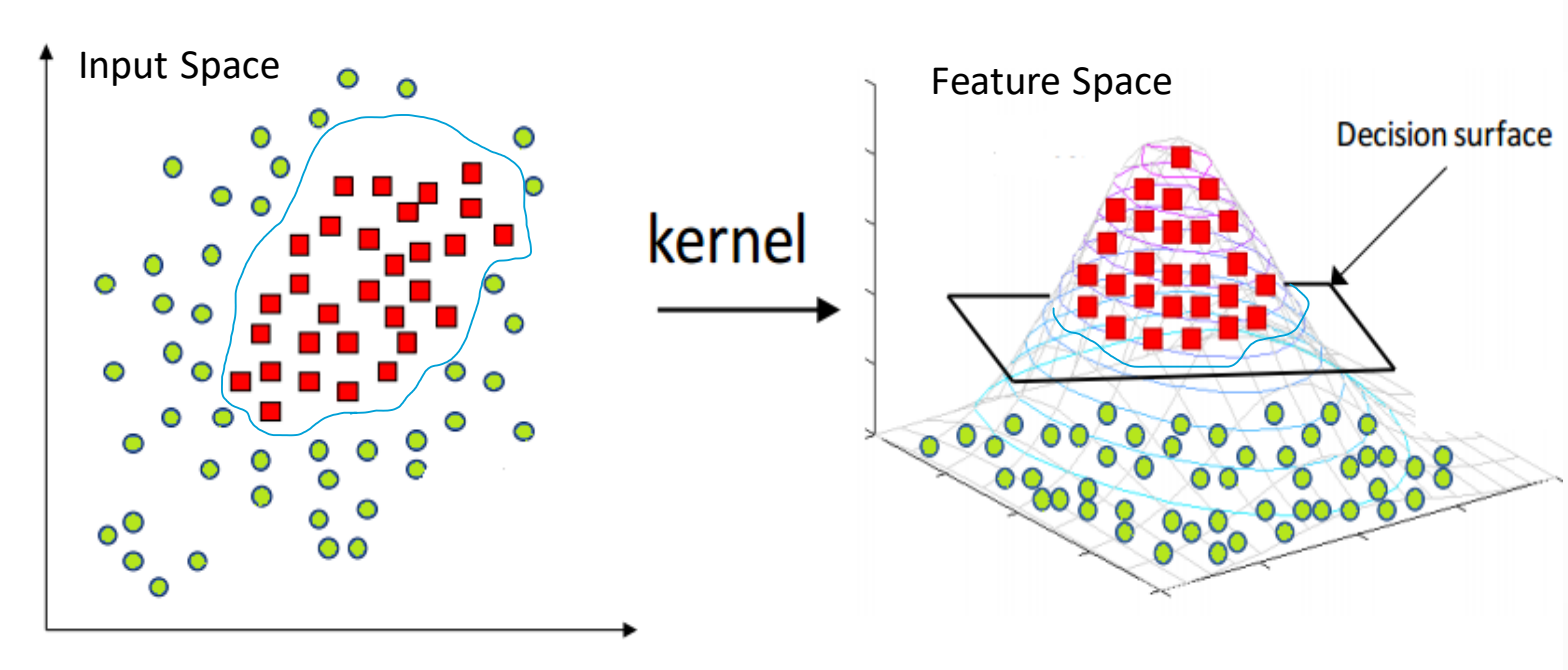
J(w): loss function

It's not the best algorithm in this case since our data is very overlapped



https://www.researchgate.net/figure/Classification-of-data-by-support-vector-machine-SVM_fig8_304611323

Kernelized Support Vector Machine



Make another dimension on the **feature space** with a **non-linear function**

Gamma parameter
Distance of influence of a single point

https://medium.com/analytics-vidhya/how-to-classify-non-linear-data-to-linear-data-bb2df1a6b781

Classification Metrics

For Binary classification

Classification Accuracy

Precision

Recall

Precision/Recall Curve

ROC

AUC

Classification Accuracy

$$\text{Accuracy} = \frac{n \text{ Correct predictions}}{N \text{All predictions}}$$

For the best **Random State**, **C** parameter, **Gamma** parameter

$$\text{Accuracy}_{train} \approx 0.987$$

$$\text{Accuracy}_{test} \approx 0.986$$

Dummy Classifier*

It's not a metric but a good way to evaluate if the data is easily separable and to understand better your data.

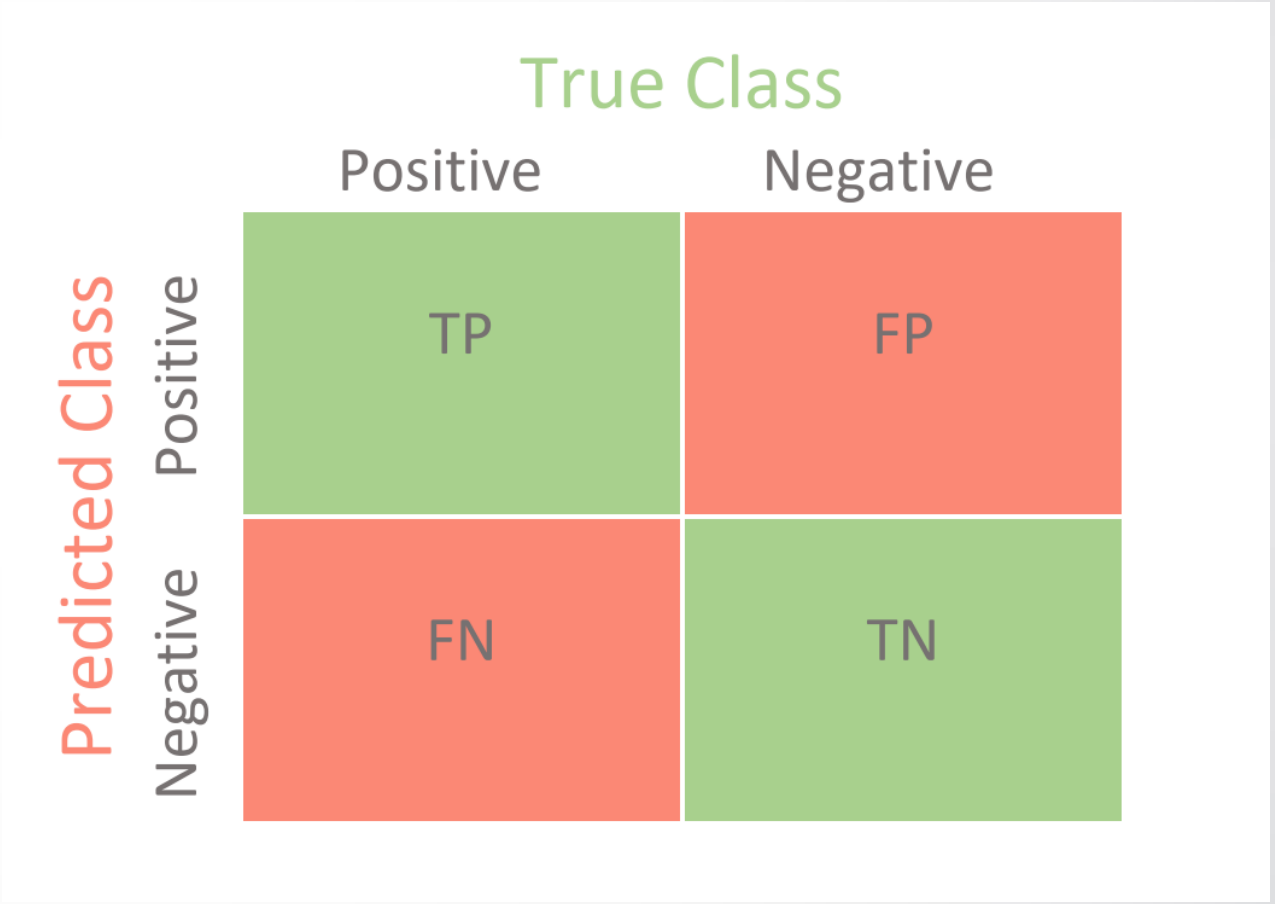
- **Most Frequent:** The classifier always predicts the most frequent class label in the training data.
- **Stratified:** It generates predictions by respecting the class distribution of the training data. It is different from the "most frequent" strategy as it instead associates a probability with each data point of being the most frequent class label.
- **Uniform:** It generates predictions uniformly at random.

Most Frequent:

$$\text{Dummy}_{test} \approx 0.781$$

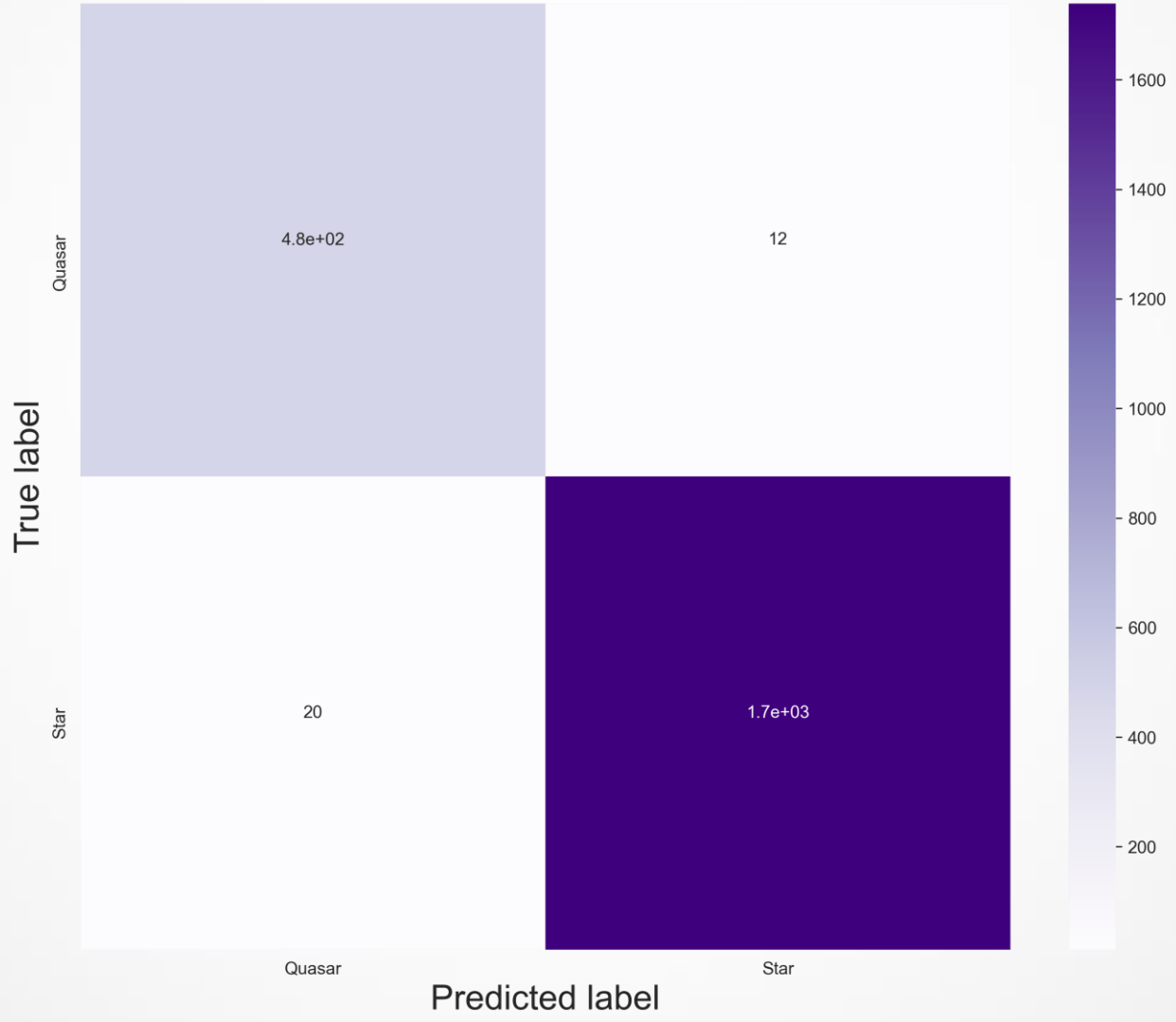
Confusion Matrix

- Insights about the errors a model is making
- Can reveal biases



https://www.r-bloggers.com/2020/12/weighting-confusion-matrices-by-outcomes-and-observations/

Confusion Matrix



Precision and Recall

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

To **minimize False Positives** we want to **maximize Recall**

To **minimize False Negatives** we want to **maximize Precision**

$$Recall_{star} \approx 0.99 \qquad Recall_{quasar} \approx 0.98$$

$$Precision_{star} \approx 0.99 \qquad Precision_{quasar} \approx 0.96$$

Predict Probability*

It's not a metric! But it's important to ROC.
Some models can assign a probability to their predictions

I used **Linear Support Vector Machine**

Non Linear Support Vector Machine can't predict the probabilities

Predicted Probabilities

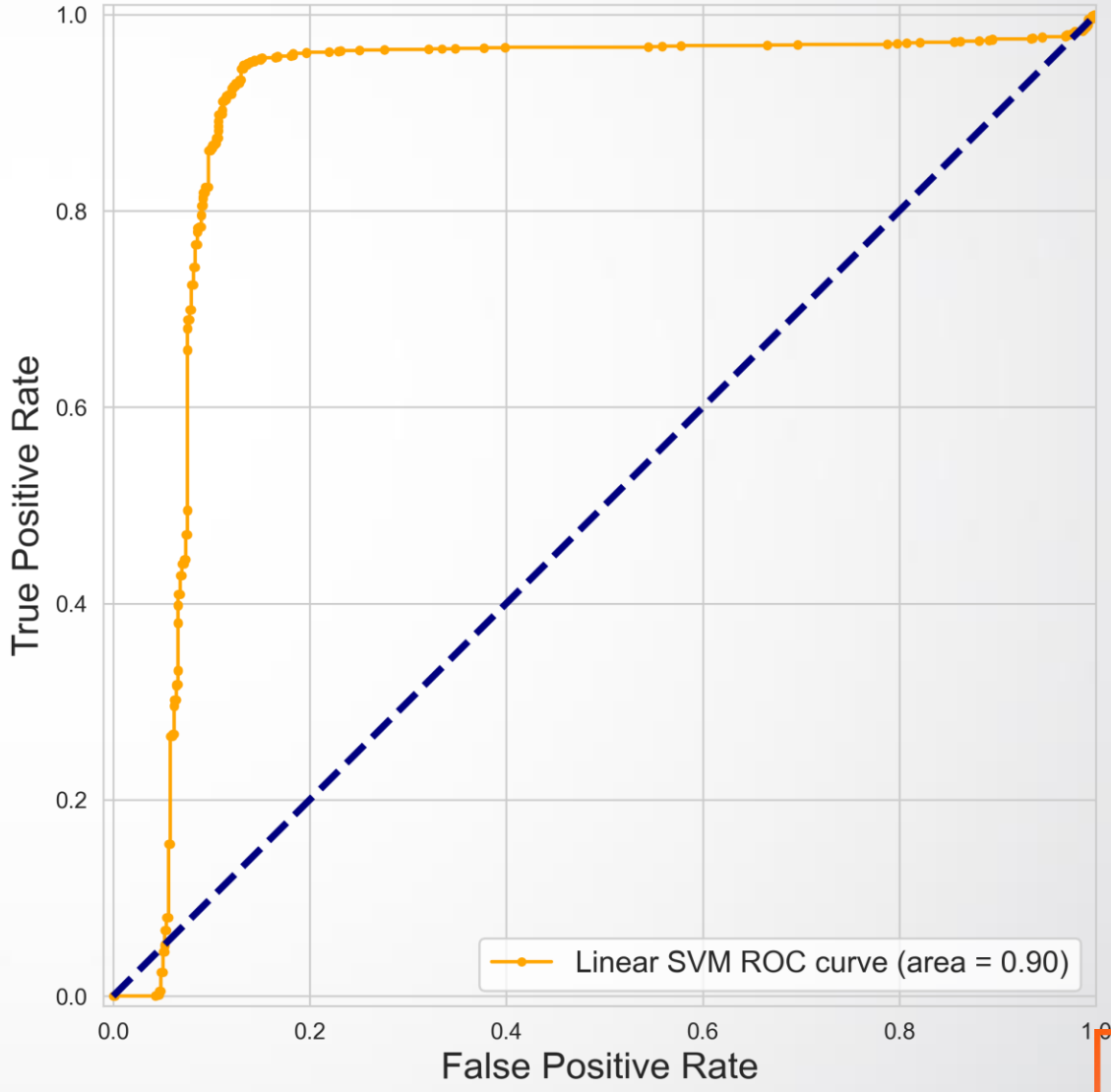
1	2
0.59	0.41
0.20	0.80
0.06	0.94
0.12	0.88
0.16	0.84

ROC Curve

(Receiver Operating Characteristic) curve

$$\text{True Positive Rate} = \frac{TP}{TP + FN}$$

$$\text{False Positive Rate} = \frac{FP}{FP + TN}$$



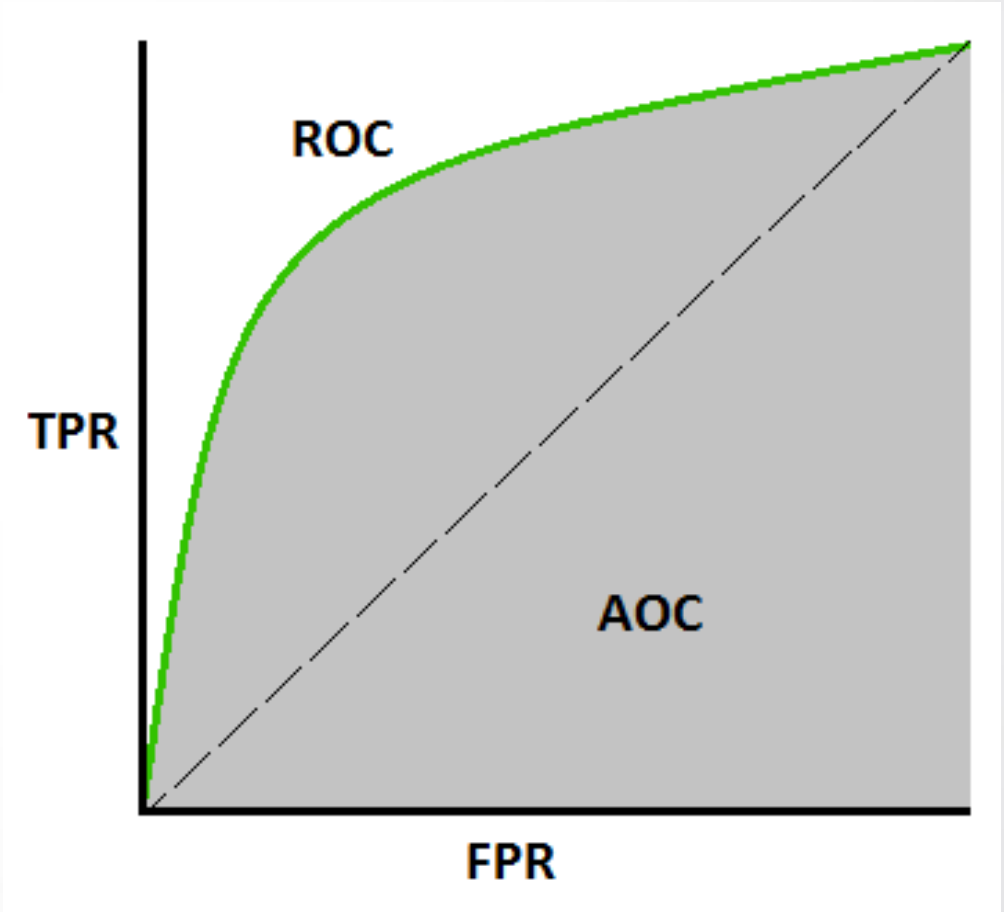
AUC
(Area Under Curve)

The name says my itself

Higher the **AUC**, the better is at predicting the classes

In our case **AUC(area) = 0.9**

AUC(area) = 0.5 represents no distinction between the classes

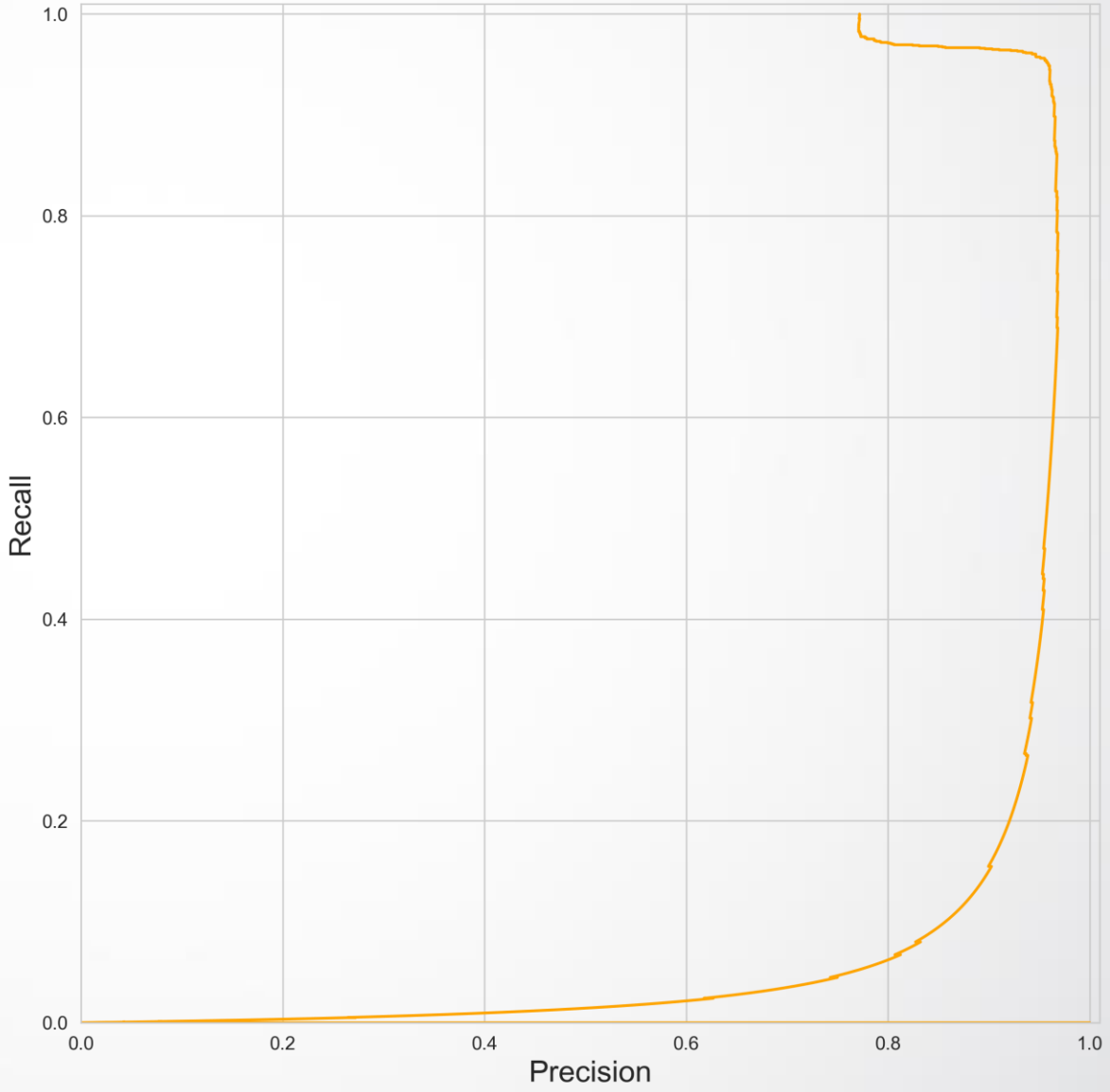


https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Precision x Recall Curve

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$



Reminders

It's very important to **know your data!**

Keep your **objetives very clear** in our mind

To choose the model and the metric you need to know very well your objetive and data

References and Recommendations

- https://www.youtube.com/watch?v=wpQiEHYkBys&t=618s (Evaluation Metric for classification and regression)
- https://towardsdatascience.com/hyperparameter-tuning-for-support-vector-machines-c-and-gamma-parameters-6a5097416167 (Kernelized Support Vector Machine parameters)
- https://medium.com/analytics-vidhya/how-to-classify-non-linear-data-to-linear-data-bb2df1a6b781 (Non linear SVM (aka. Kernelized SVM) explained)
- https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 (Confusion matrix explained)
- https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (AUC and ROC Cover explained)
- https://www.r-bloggers.com/2020/12/weighting-confusion-matrices-by-outcomes-and-observations/ (R approach to confusion matrix and interpretation)