aws machine learning

Architecting for ML, On AWS

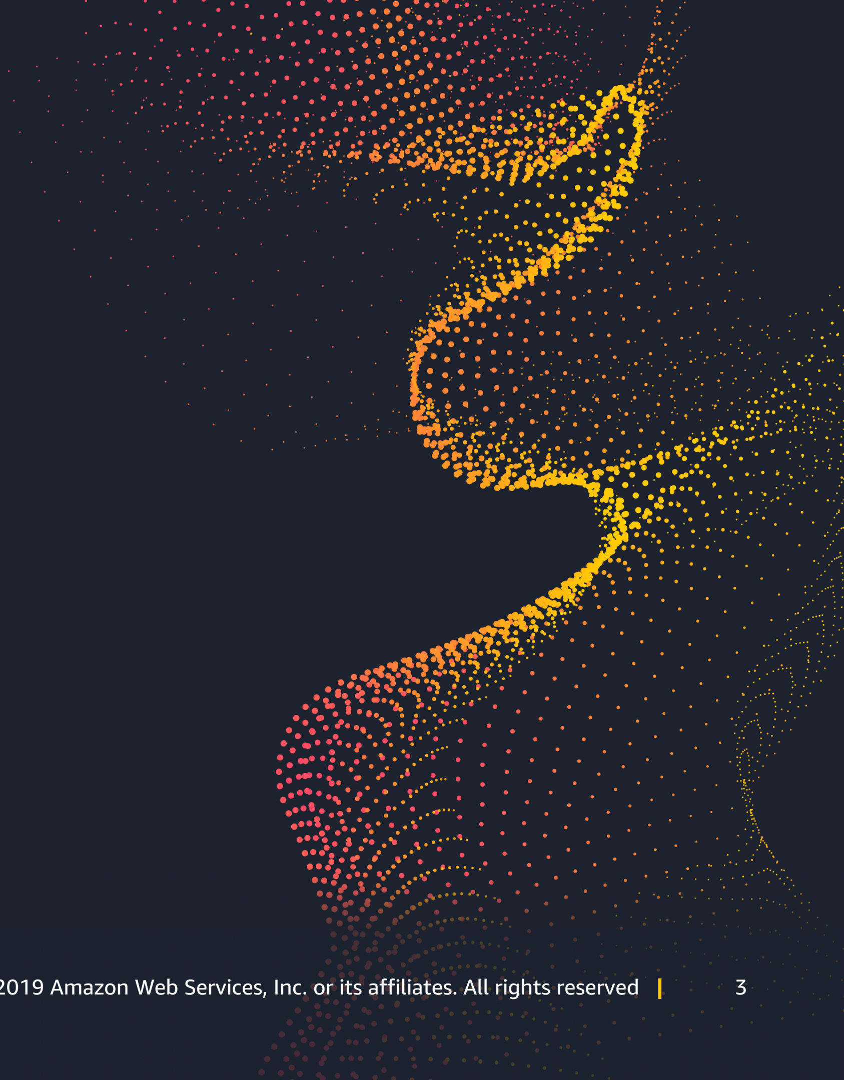
Day 2

Ibrahim Gabr & Oliver Steffmann
Machine Learning & Artificial Intelligence Solutions Architect

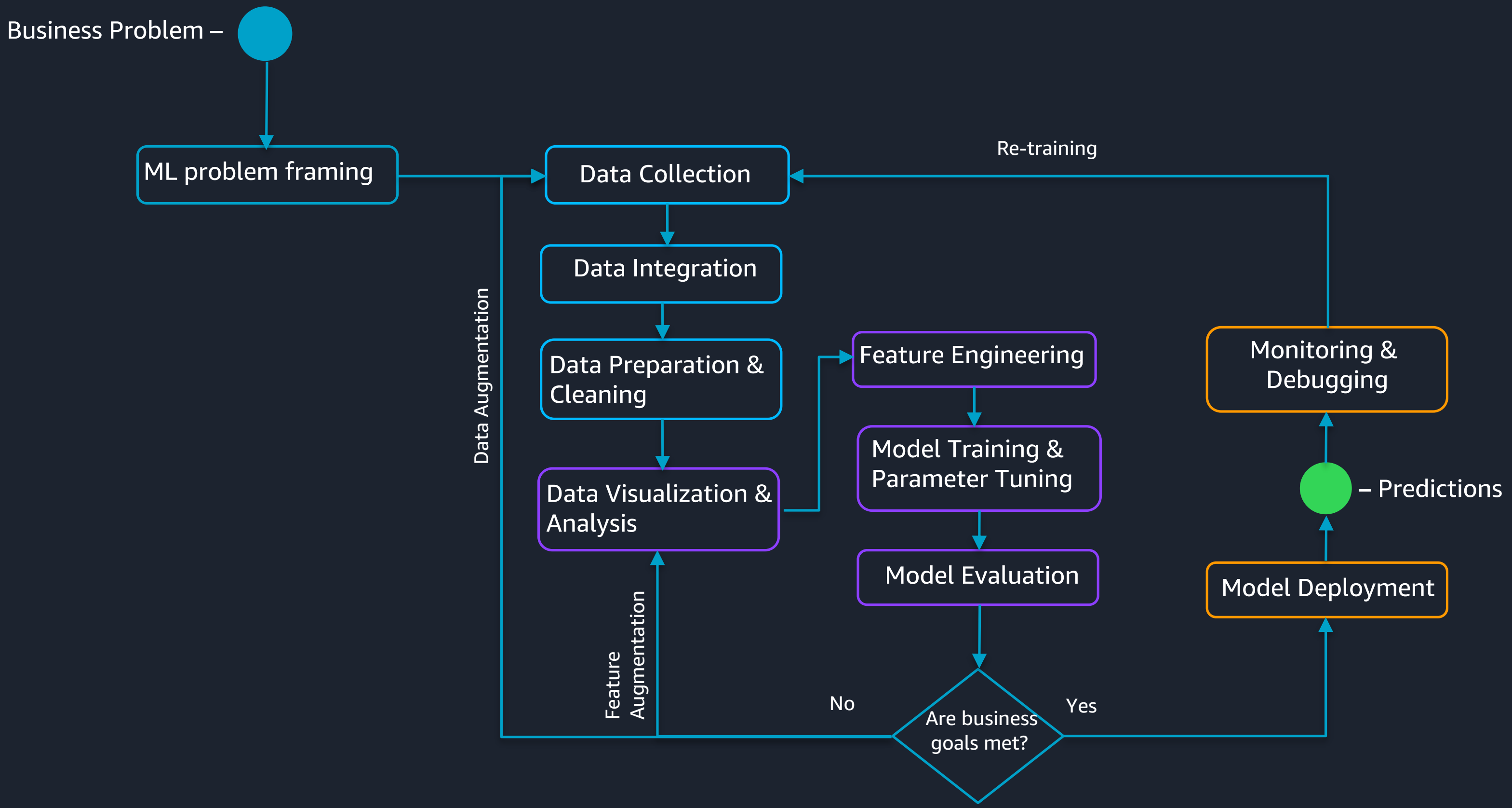
Agenda

Day 1	Day 2	Day 3
AI/ML on AWS Intro lab	Feature engineering Model evaluation	Moving to production Build
Team up Define problem	Build	Present
Write-up	Working model	Solution architecture

Feature Engineering



Machine learning process



Eighty percent of data science work

is data preparation

Prepare Training Data

Data Selection

- Fully explore available data
- Consider more data sources
- Think about what missing data
- Exclude data you don’t need
- Look at feature correlations

Data Processing

- Clean the data to remove bad data, fix missing data
- Format the data to feed ML algorithms
- Sample a subset of data for initial experiments

Feature Engineering

- Scale the data to a consistent scale
- Rounding, binning
- Aggregate features to single values
- Encode data, reduce dimensions
- Remove outliers
- Derive new features

Why?

- You can isolate and highlight key information, which helps your algorithms "focus" on what’s important.
- You can bring in your own **domain expertise**.
- Most importantly, once you understand the "vocabulary" of feature engineering, you can bring in other people’s domain expertise!

Feature Engineering - Conceptual

What rows and columns are in my data set already?

Do those actually represent the real world?

How are they going to interact with my model?

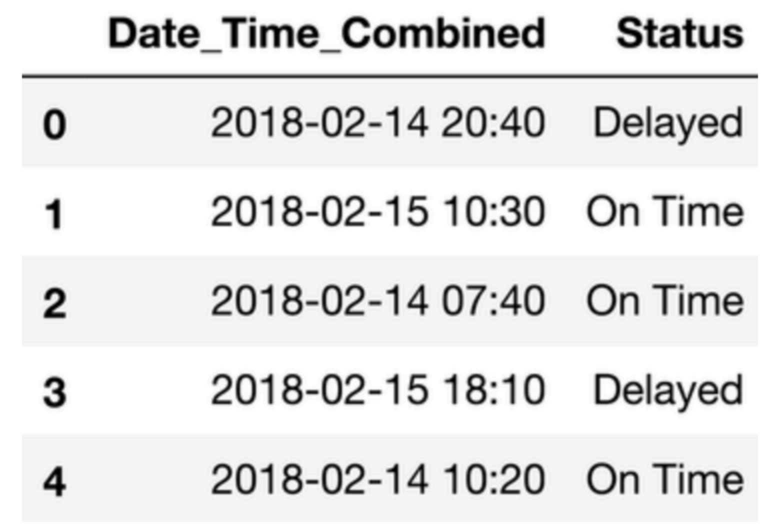
Do I need to transform any columns? Normalize? Scale?

Do I need to remove any outliers?

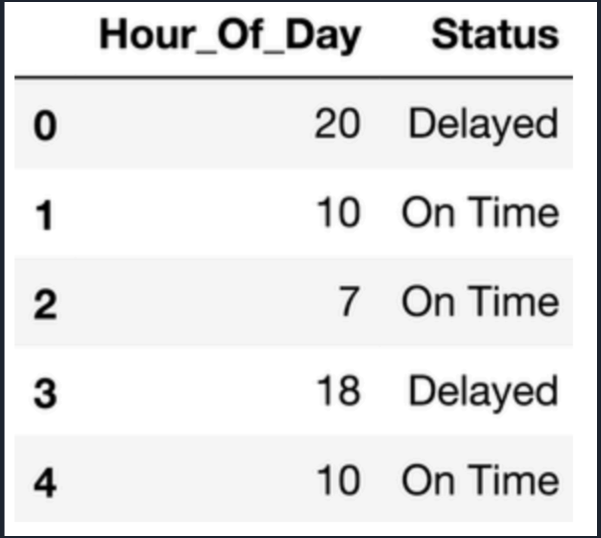
Do I need to combine any columns?

Do I need to add additional features?

Simple example



	Date_Time_Combined	Status
0	2018-02-14 20:40	Delayed
1	2018-02-15 10:30	On Time
2	2018-02-14 07:40	On Time
3	2018-02-15 18:10	Delayed
4	2018-02-14 10:20	On Time



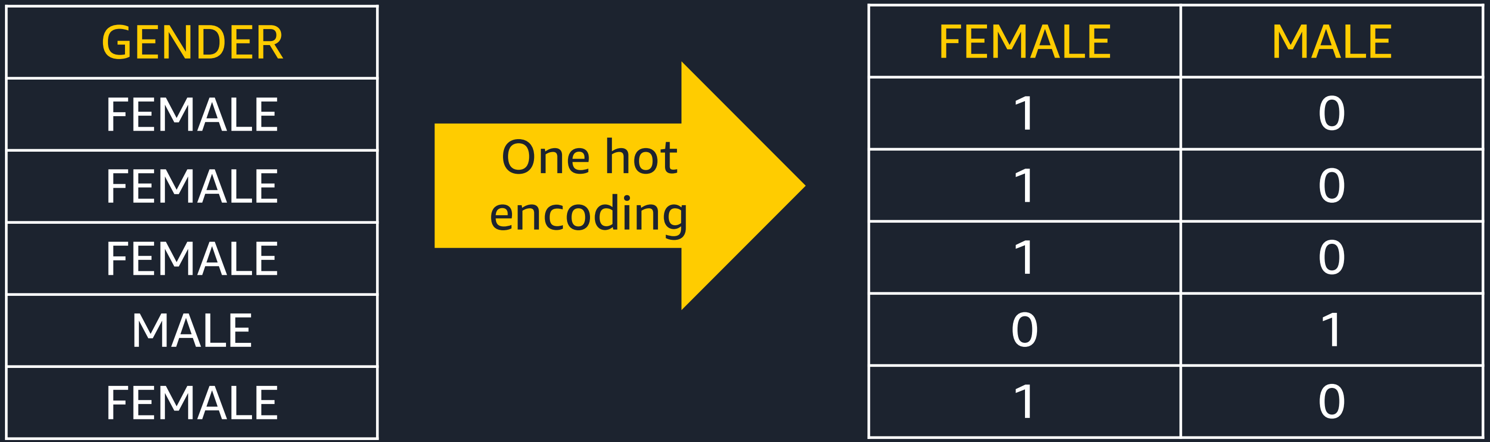
	Hour_Of_Day	Status
0	20	Delayed
1	10	On Time
2	7	On Time
3	18	Delayed
4	10	On Time

Other considerations

- Are my column data types appropriate?
- Do I need any one hot encoding of categorical features?
- Do I need to transform any of my columns?
- Should I be using data augmentation?
- Are there any specific data requirements imposed by the ML algorithm?
- Do I have sufficient Pandas and Python functions to prepare the data?

One hot encoding

```
# perform one-hot encoding of a specific column
tmp_df = pd.get_dummies(df['GENDER'])
tmp_df.head()
```

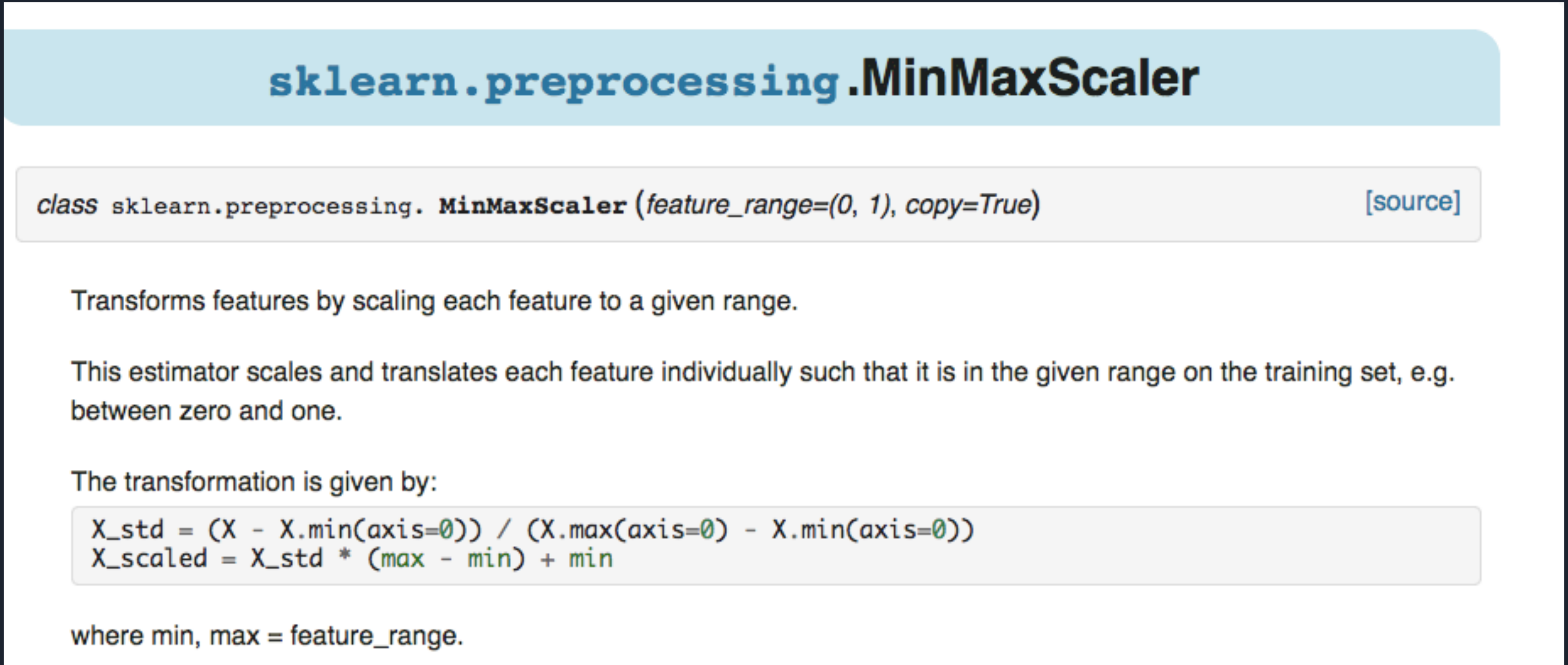


GENDER
FEMALE
FEMALE
FEMALE
MALE
FEMALE

One hot encoding

FEMALE	MALE
1	0
1	0
1	0
0	1
1	0

Min Max Scaling



sklearn.preprocessing.MinMaxScaler

class `sklearn.preprocessing.` **`MinMaxScaler`** (*feature_range=(0, 1), copy=True*) [source]

Transforms features by scaling each feature to a given range.

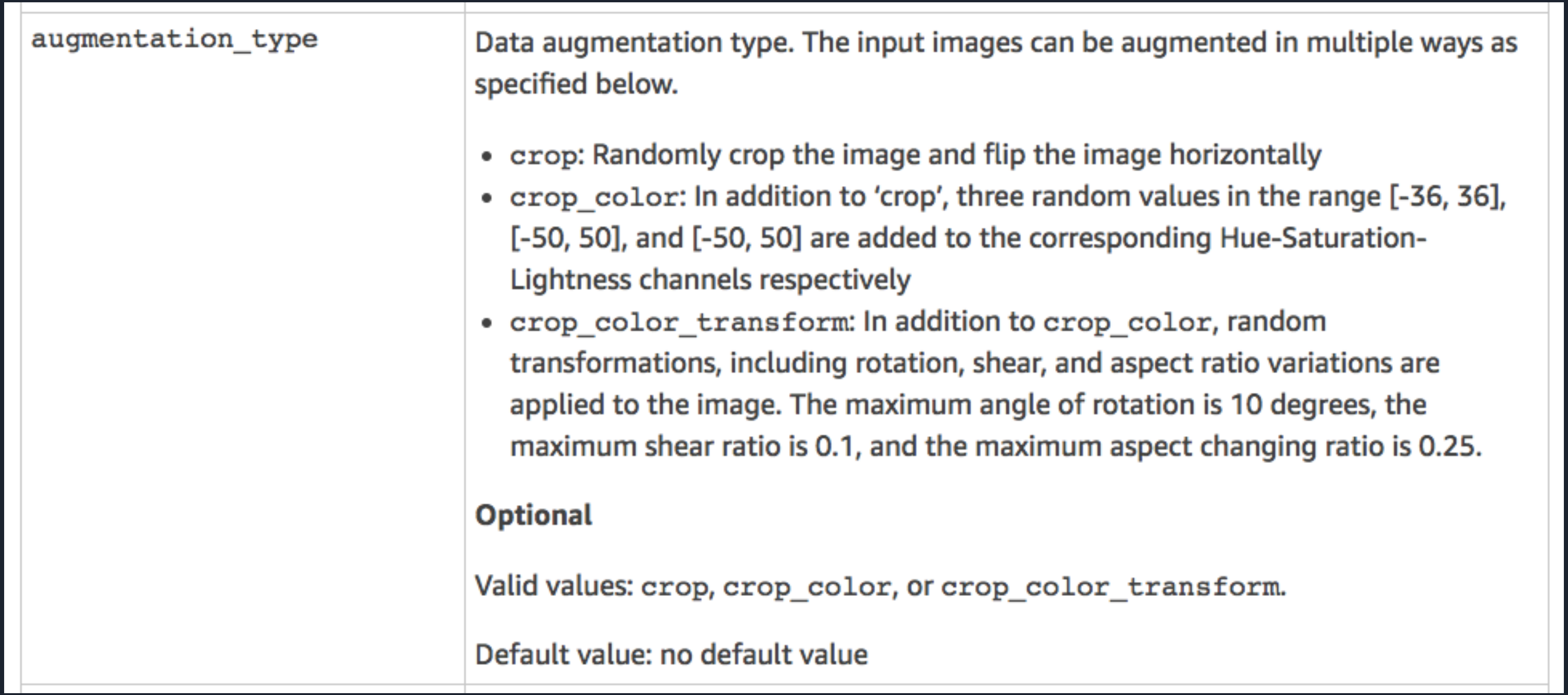
This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

The transformation is given by:

```
X_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))
X_scaled = X_std * (max - min) + min
```

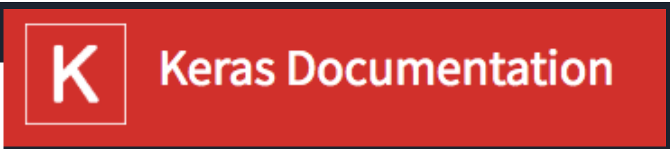
where min, max = feature_range.

Should I use data augmentation?



`augmentation_type`	Data augmentation type. The input images can be augmented in multiple ways as specified below. • `crop`: Randomly crop the image and flip the image horizontally • `crop_color`: In addition to 'crop', three random values in the range [-36, 36], [-50, 50], and [-50, 50] are added to the corresponding Hue-Saturation-Lightness channels respectively • `crop_color_transform`: In addition to `crop_color`, random transformations, including rotation, shear, and aspect ratio variations are applied to the image. The maximum angle of rotation is 10 degrees, the maximum shear ratio is 0.1, and the maximum aspect changing ratio is 0.25. **Optional** Valid values: `crop`, `crop_color`, or `crop_color_transform`. Default value: no default value

Image Classification Hyperparameters



K Keras Documentation

Feature Engineering - Practical

How do I change the type of data I'm working with?

What Python and Pandas functions will I use?

Are my X's and Y's actually lining up? If not, why?

Do I have the mechanics properly set up?

Do I have everything I need to launch my training job?

Feature Selection

- Statistical -> Correlation, Chi-Square etc.
- Recursive Feature Elimination
- Automatic
  - Lasso
  - Tree's

Some basics

```
import pandas as pd
# read and write CSV files
df = pd.read_csv(‘file_name.csv’)
df.to_csv('fraud_train.csv', sep='\t’,
           index=False, header=False)
# plot histograms of values
df.hist()
df[‘column_name’].hist()
```

Filtering

```
# 2 syntax options for filtering
tmp_df = df[(df[‘col’] > 2) &
            (df[‘otherCol’] < 10)]
tmp_df = df[(df.col > 2) &
            (df.otherCol < 10)]
# selecting rows with column value in set
tmp_df[tmp_df[‘col’].isin([10,15,20])]
```

Counts of values

```
# get value counts for each unique value
tmp_df[’col7’].value_counts()

# get values
tmp_df[‘col7’].value_counts().index.tolist()

# get number of occurrences
tmp_df[‘col7’].value_counts().values.tolist()
```

Miscellaneous

```
# add a new column
tmp_df[‘newCol’] = ‘some value’

# get shape, basic statistics
tmp_df.shape
tmp_df.describe()

# concatenate dataframes
trans_df  = pd.DataFrame(X_train)
target_df = pd.DataFrame(y_train)
train_df  = pd.concat([target_df, trans_df],
                      axis=1)
```

Train / test split

```
# use sklearn to split dataframe content to
# arrays of data for training and testing
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = \
  train_test_split(df[['AGE','GENDER','LOCATION']],
                   df['SALARY’],
                   test_size=0.25,
                   random_state=1)
```

Some S3 commands

```
# copy an s3 file locally
!aws s3 cp s3://source_bucket/file_name.csv .

# copy an entire s3 folder locally
!aws s3 sync s3://source_bucket/folder .

# upload a folder to s3
!aws s3 sync source_folder s3://source_bucket/folder
```

Model Evaluation

Confusion matrix

Predictions

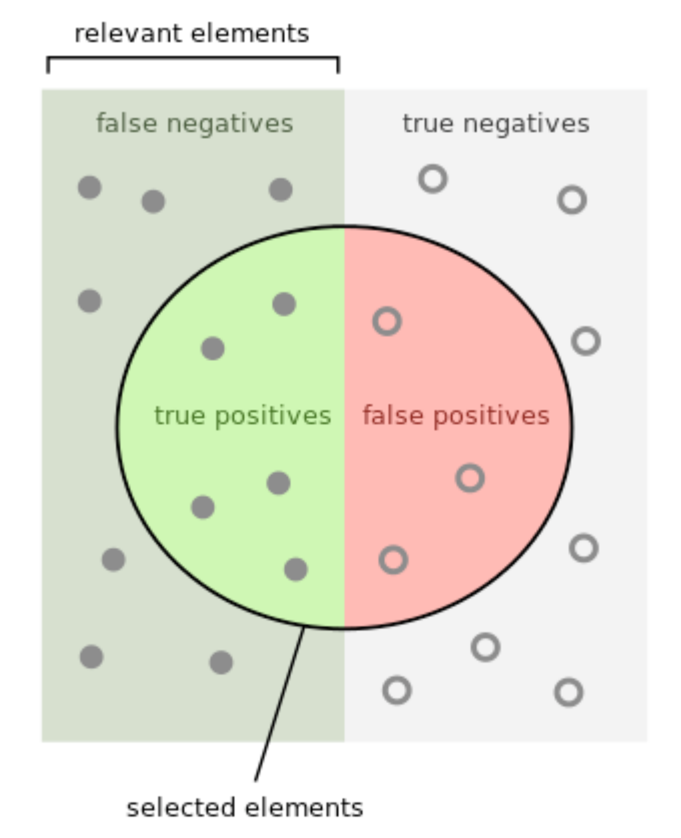
Actuals (labeled data)

	Positive	Negative
Positive	True Positive ✓	False Negative ✗
Negative	False Positive ✗	True Negative ✓

Recall

Precision

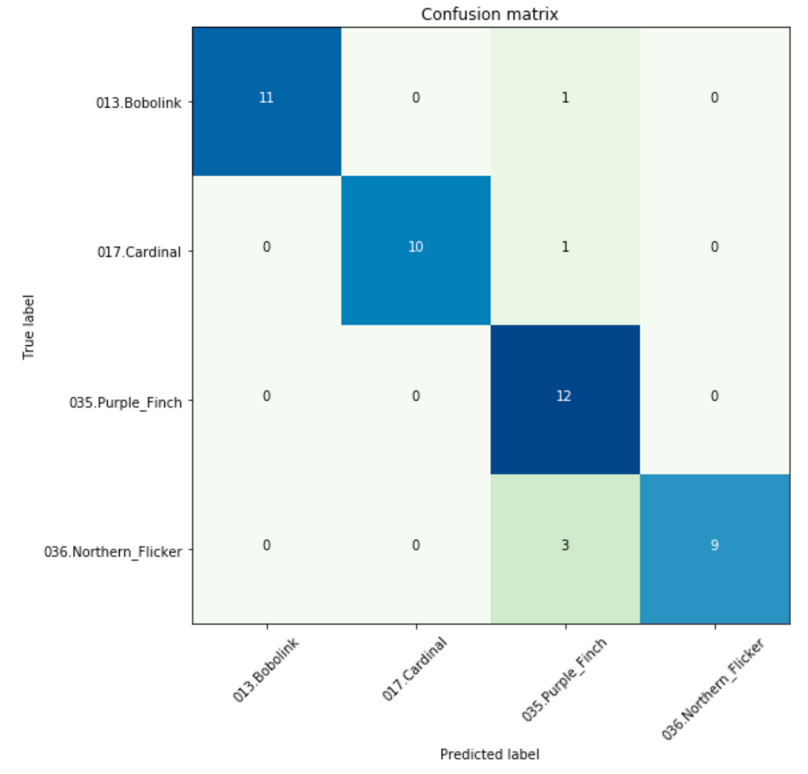
Classification model results



- Actual positive samples on the left, negatives on the right
- Predicted positive samples in the circle, predicted negatives outside
- In each case, some are correct (true), others are incorrect (false)
- Precision is? 5 / 8 = 62.5%
- Recall is? 5 / 12 = 41.7%

https://en.wikipedia.org/wiki/Precision_and_recall

Multi-class confusion matrix



Confusion matrix

True label \ Predicted label	013.Bobolink	017.Cardinal	035.Purple_Finch	036.Northern_Flicker
013.Bobolink	11	0	1	0
017.Cardinal	0	10	1	0
035.Purple_Finch	0	0	12	0
036.Northern_Flicker	0	0	3	9

Common binary classification model evaluation metrics

$$\frac{\text{TP} + \text{TN}}{\text{Total Predictions}}$$

Accuracy

$$\frac{\text{TP}}{\text{Positive Predictions}}$$

Precision

$$\frac{\text{TP}}{\text{Positive Samples}}$$

Recall

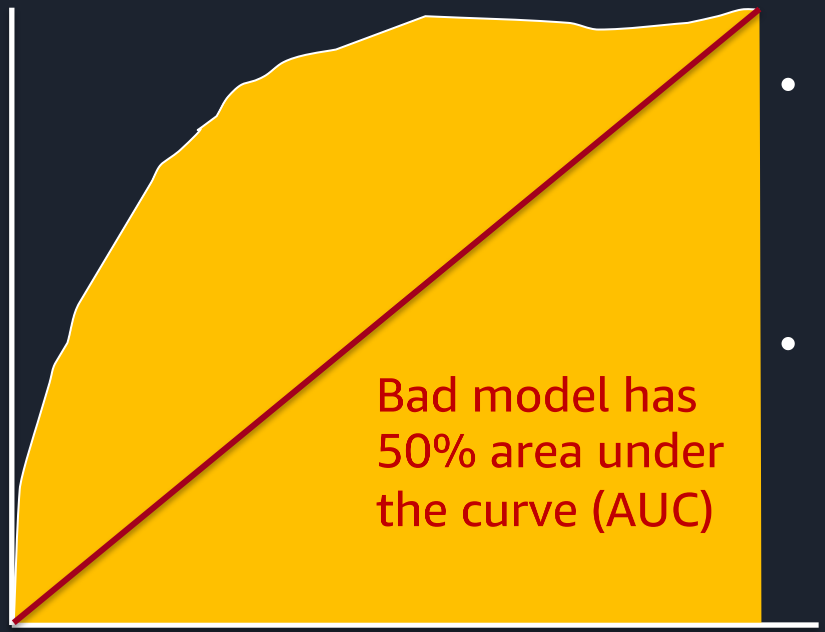
AUC

(area under ROC curve)

Receiver Operator Curve

True Positive Rate

(aka recall, sensitivity)

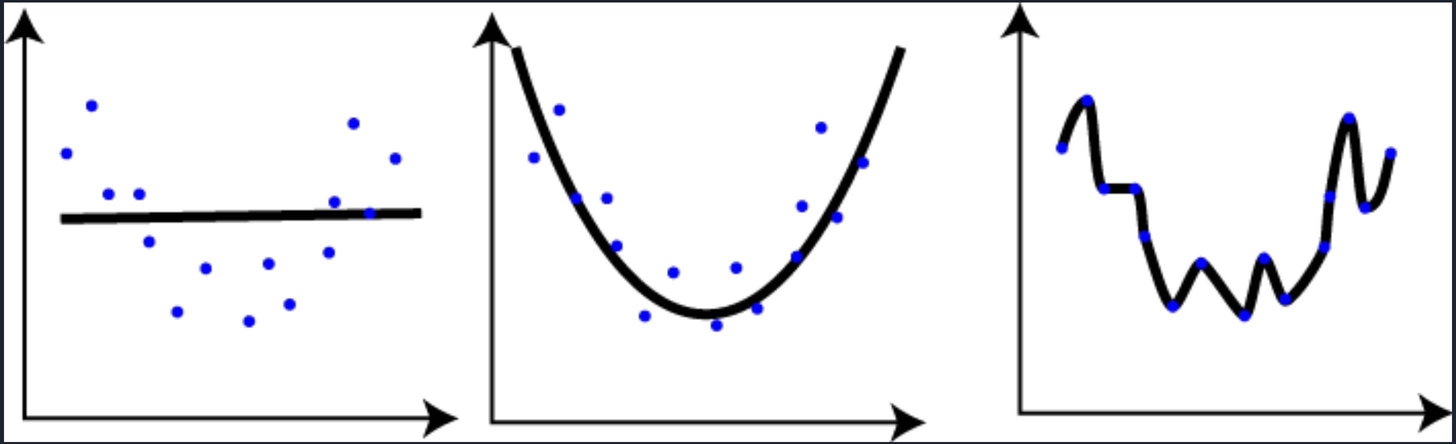


False Positive Rate

- Measure true positives and false positives given different probability cut-offs
- Best models get closest to area of 1.0

Good ROC blog post: https://medium.com/greyatom/lets-learn-about-auc-roc-curve-4a94b4d88152

Overfitting vs. Underfitting



Underfitting

Normal

Overfitting

It is very helpful to align model evaluation directly with business goals

Loss Function – Give different impact to different errors when training the model

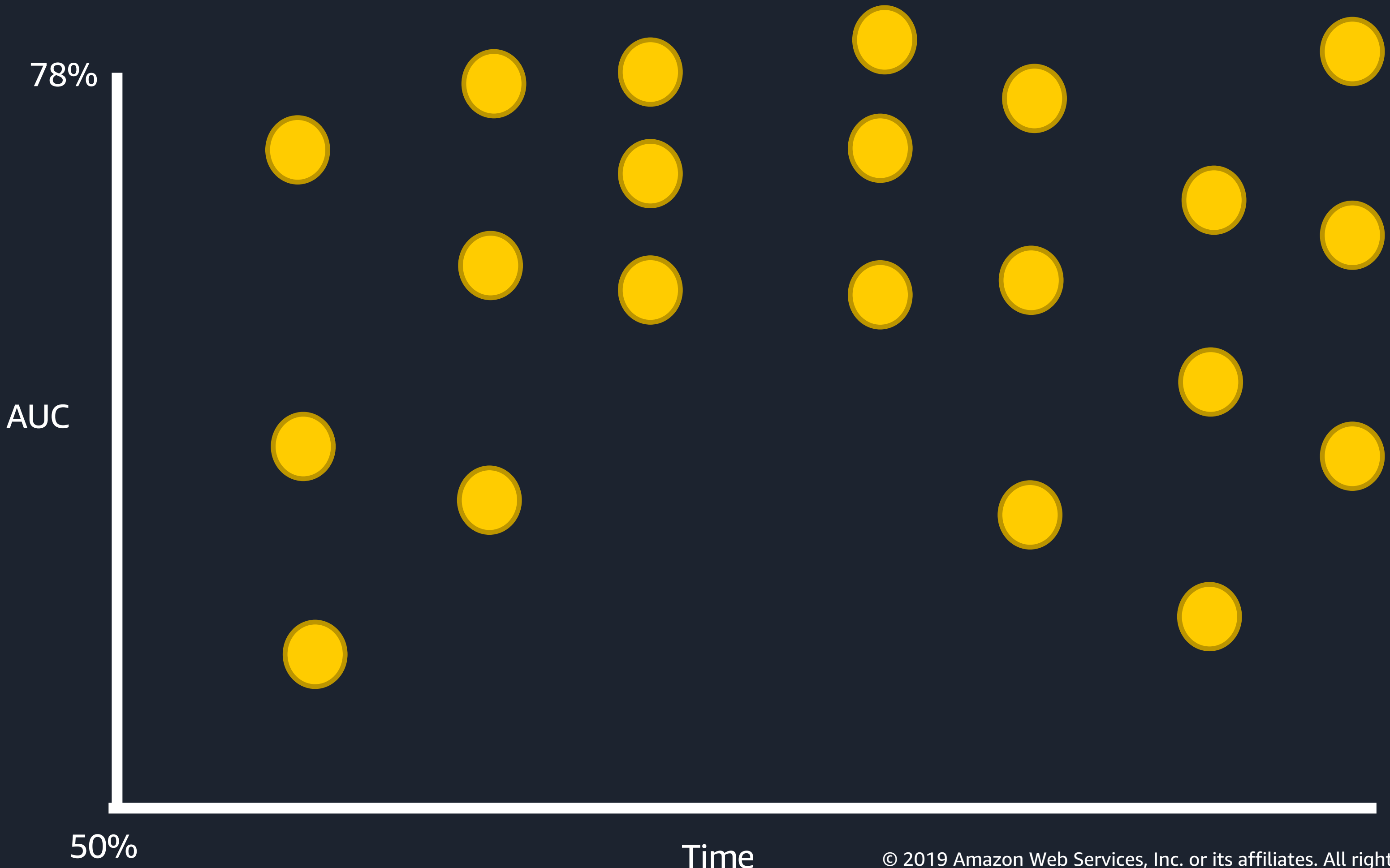
Economic Weights – Quantify the economic impact of the model outputs to evaluate results https://aws.amazon.com/blogs/machine-learning/training-models-with-unequal-economic-error-costs-using-amazon-sagemaker/

How do I eek out a few more percentage points of model accuracy?

Hyperparameter Optimization (HPO)

a.k.a.

Automatic Model Tuning


78%
AUC
50%
Time

Use the evaluation questions to ask interesting questions about about your project.

Over time, as you become a machine learning practitioner, you should be able to answer them.

Today, don't stress yourself out.

aws machine learning