SLU13 - Bias-Variance trade-off & Model Selection

December 3rd, 2023

1. Introduction

Context

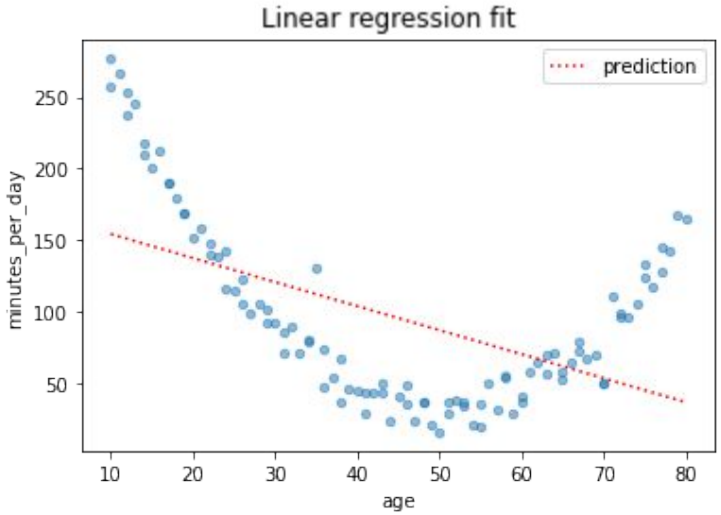
SLU01 to SLU06 → Data

SLU07 to SLU12 → Models

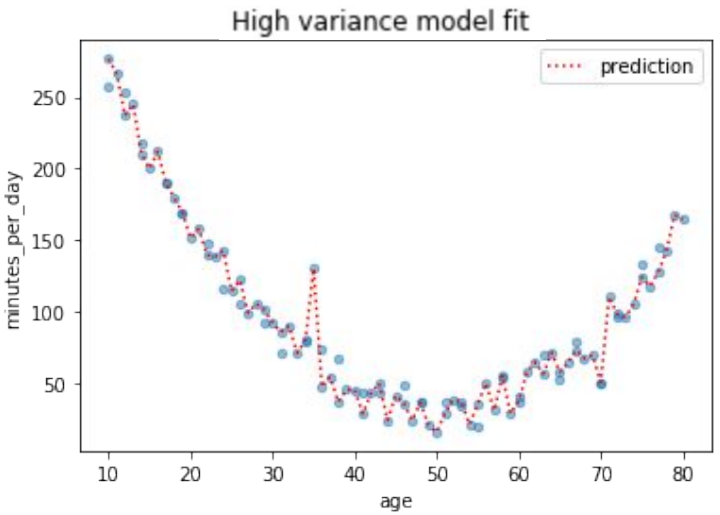
SLU13 to SLU17 → Model Improvement

Motivation

Which model is better?



$R^2 = 0.05$



$R^2 = 0.99$

Motivation

Answer: They are both bad!

In this SLU we will try to give you the tools to answer the following questions:

- How do we evaluate our models?
    - (Hint: it's more than just maximizing R^2, accuracy, etc.)
- How do we make sure our models generalize well?

Overview

- **Bias-variance trade-off**
    a. Overfitting and underfitting
- **Model Selection**
    a. Train-test split
    b. Train-validation-test split
    c. Cross-validation
- **Learning curves**

2. Topic Explanation

Bias-variance Tradeoff → **Model Selection** → **Learning Curves**

Bias-variance Tradeoff

- What is our main goal?
  - To approximate the one true, *general* target function
- Bias-variance decomposition is a way of analyzing the generalization error

Bias:

- *Always learning the same wrong thing*
- Results from simplistic assumptions and a lack of model adaptability

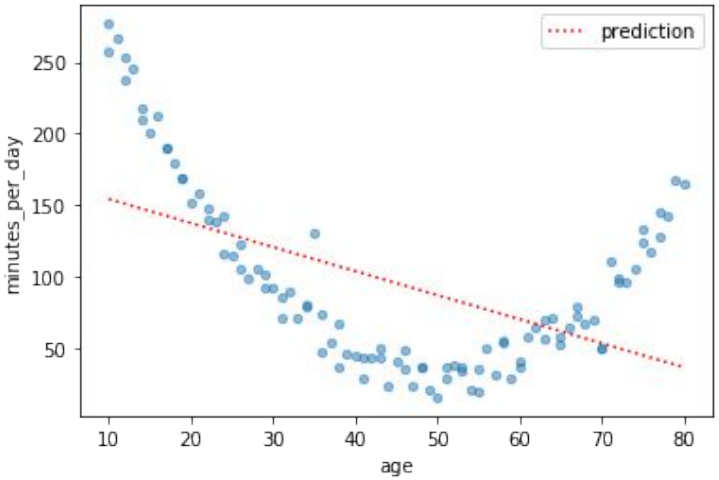
Variance:

- *Learning random things*
- Results from excessive flexibility, a lack of underlying structure to anchor predictions

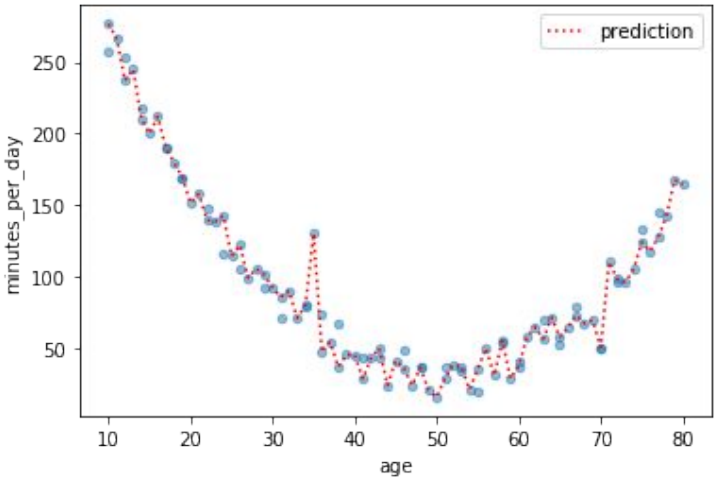
Bias-variance Tradeoff

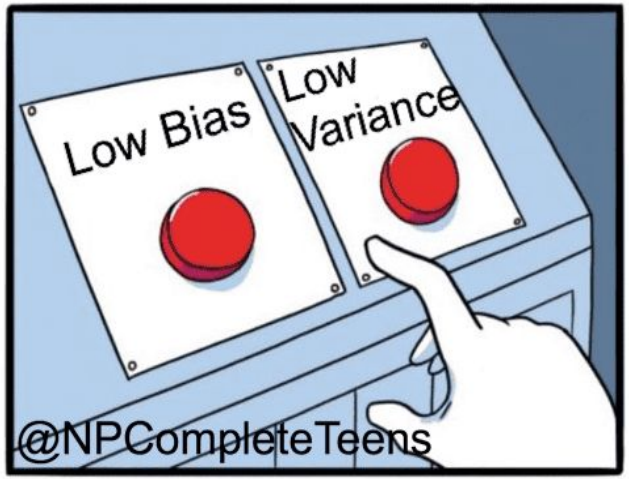
- Regression example:

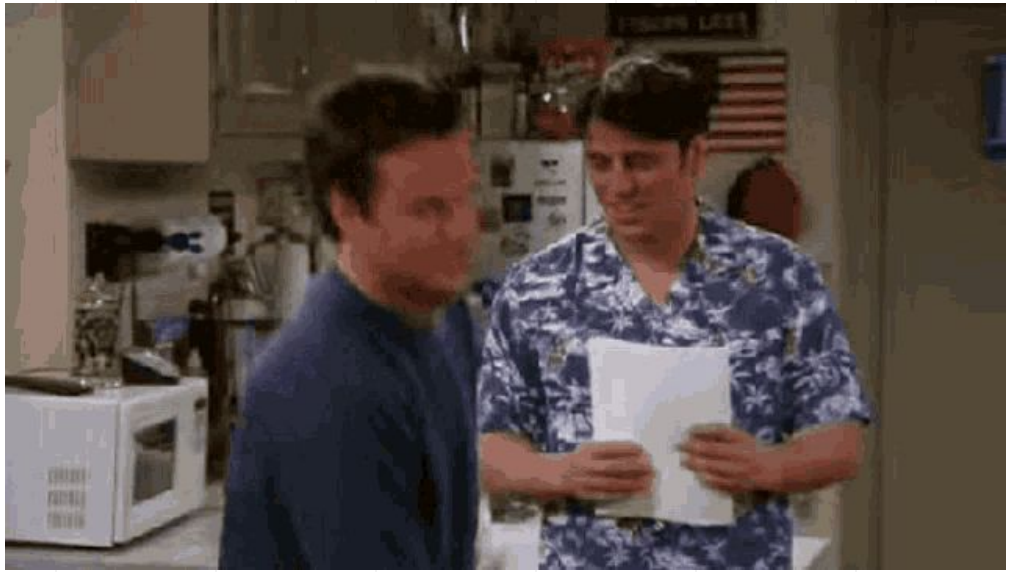
High bias == **Underfitting**



High variance == **Overfitting**



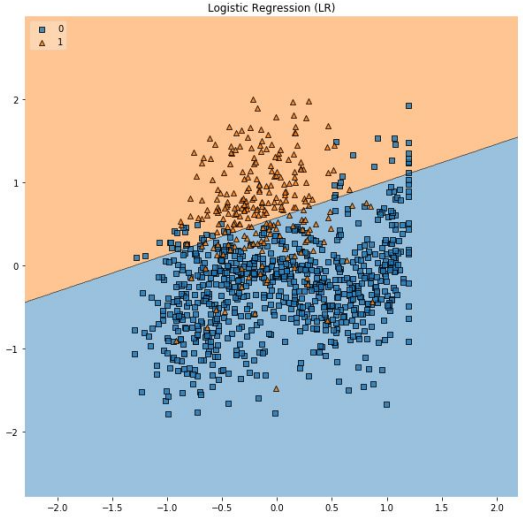

Low Bias
Low Variance
@NPCompleteTeens



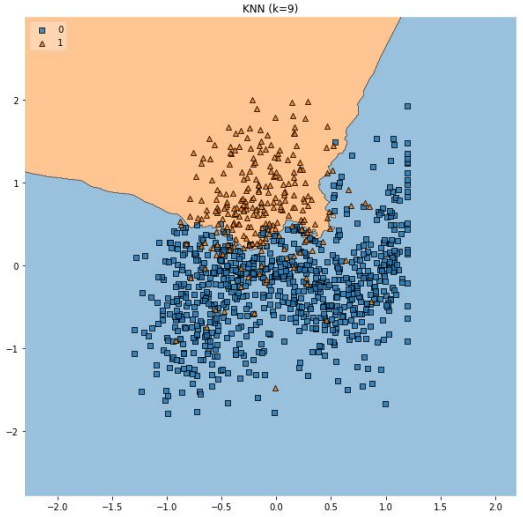
Bias-variance Tradeoff

- Classification example:

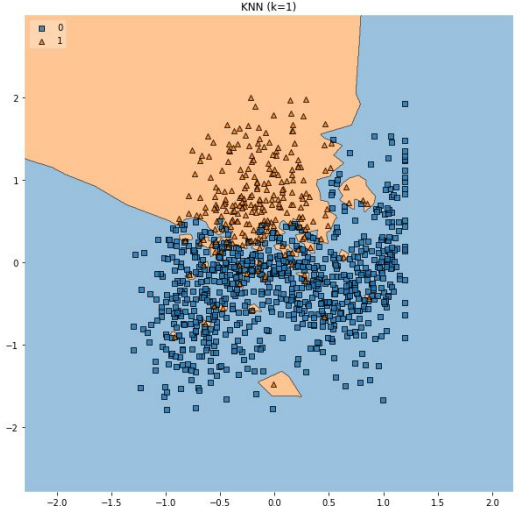
High bias (*underfitting*)



Just right

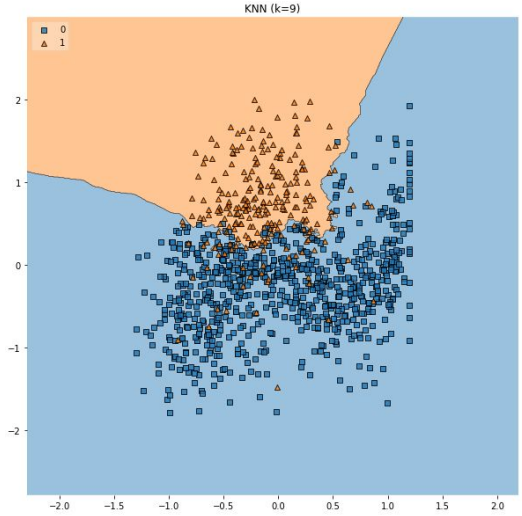


High variance (*overfitting*)

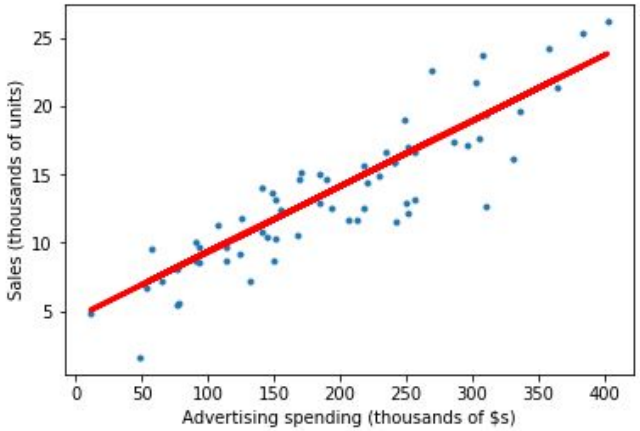


Bias-variance Tradeoff

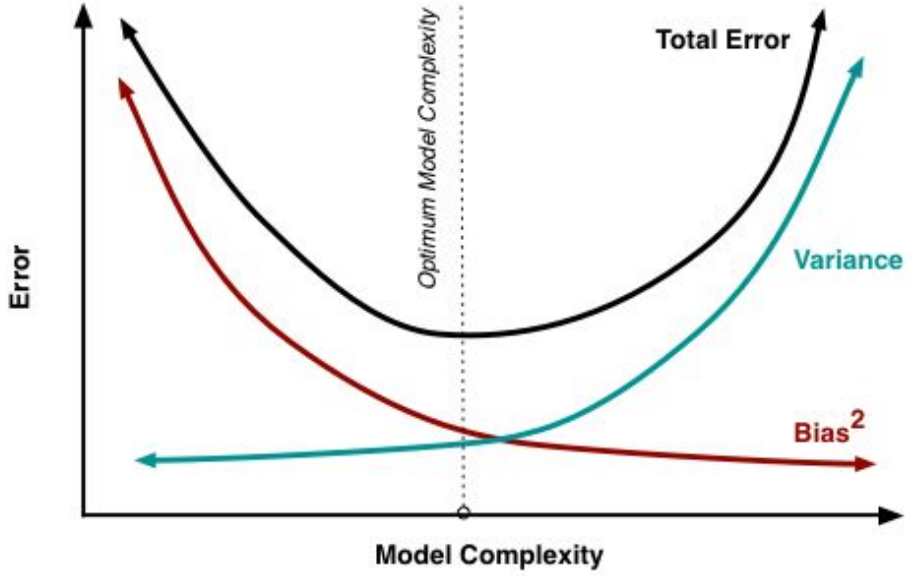
Just right



- But the “just right” model still isn’t perfect!
- There will always be some **irreducible error**:
  - Error that cannot be eliminated by building good models
  - Noise and outliers will exist



Bias-variance Tradeoff

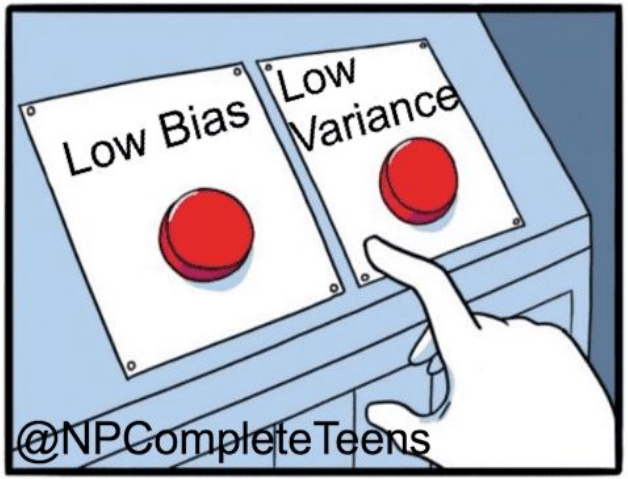


In theory, we reach the right level of complexity when the increase in bias is equivalent to the reduction in variance:

$$\frac{dBias}{dComplexity} = -\frac{dVariance}{dComplexity}$$

But there is no analytical way to find this optimum...

So we need to test our hypothesis!

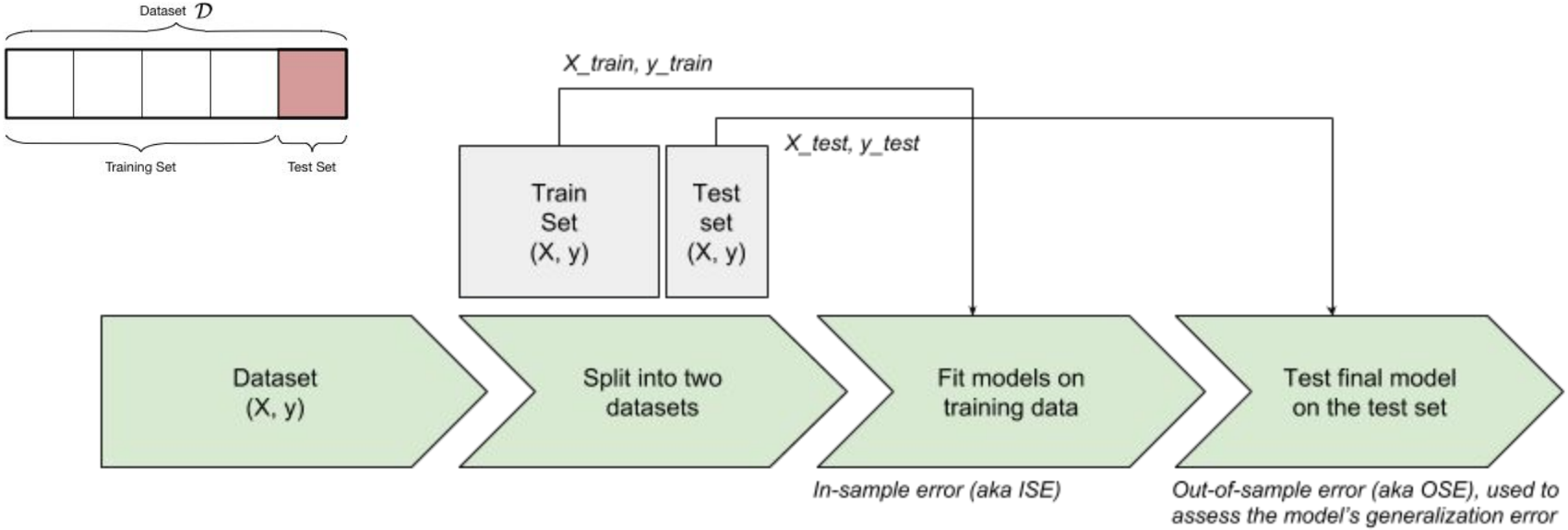

Low Bias
Low Variance
@NPCompleteTeens

Bias-variance Tradeoff

Model Selection

Learning Curves

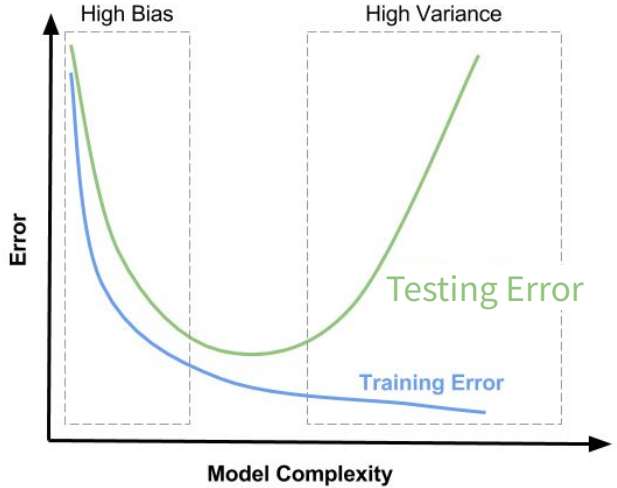
Train-Test Split



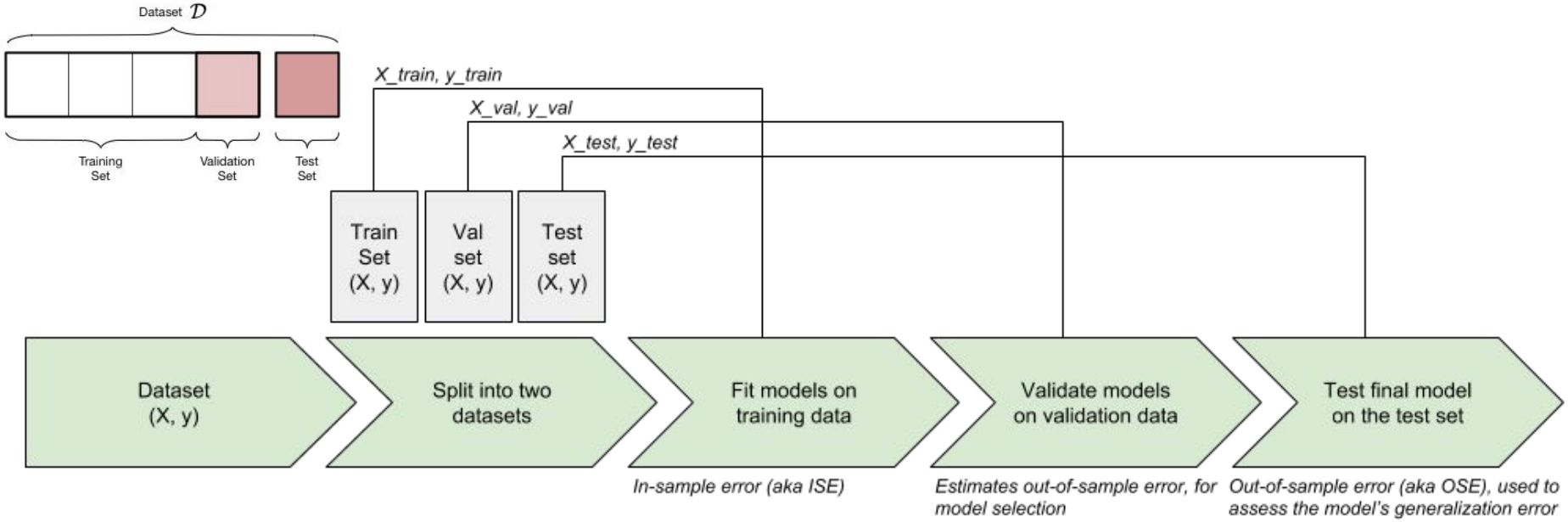
Train and Test Error

- *Training error* or in-sample error (ISE) measures how well our trained model performs on the training data

- *Testing error* or out-of-sample error (OSE) is how well our model performs on previously unseen data



Train-Validation-Test Split



Train-Validation-Test Split

Why should we have a validation set? Isn't it the same as the test set?

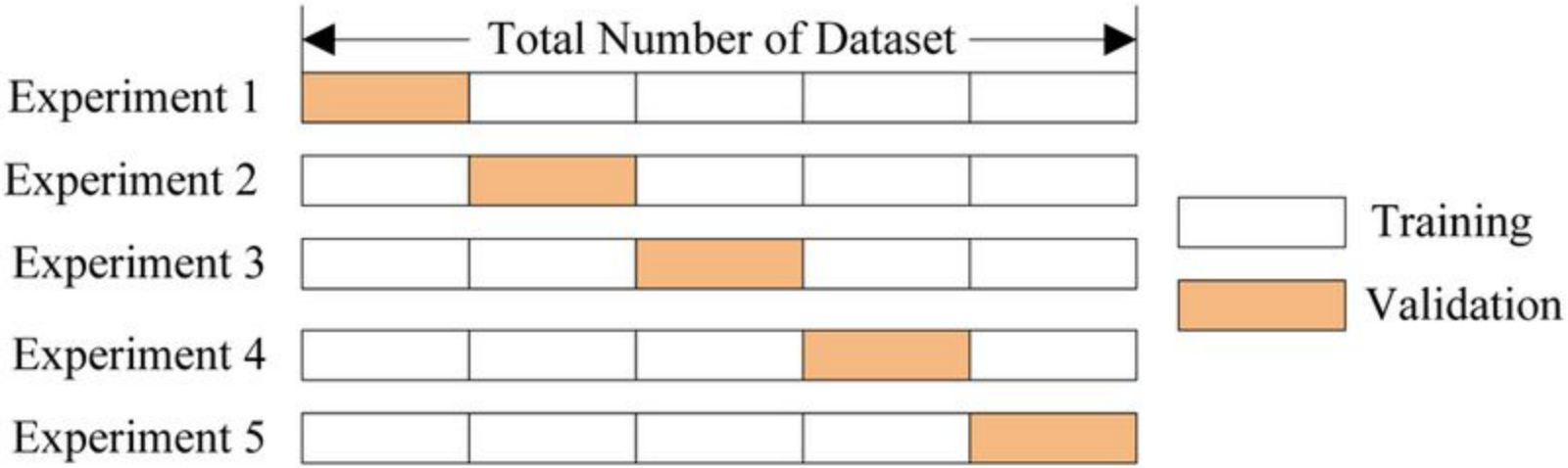
Validation set can help us deciding make decisions regarding hyper-parameters.

- Example: How many polynomial features to consider in our model?

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5$$

- Other examples:
  - Which learning rate should we use in iterative processes
  - How many iterations should we use to optimize our model
  - What should be the size of the Neural Network

Cross-Validation



- Measure the mean and variance of the **scores across each "fold"**
- Used to approximate the OSE **when we don't have a lot of data**

Summary of Evaluation Methods

- **Train-test split**
  - Most basic way to evaluate OSE
- **Train-validation-test split**
  - Needs sufficient data
  - Essential for hyperparameter tuning
- **Cross-validation**
  - Useful when you don’t have a lot of data to approximate OSE
  - Measure the mean and variance of the scores across each fold

Bias-variance Tradeoff

Model Selection

Learning Curves

How much data should we use?

- The obvious starting point about which data to use is... "why not just all of it?"

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
12	13	0	3	Saundercock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.0500	NaN	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750	NaN	S
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542	NaN	S

slow

Expensive

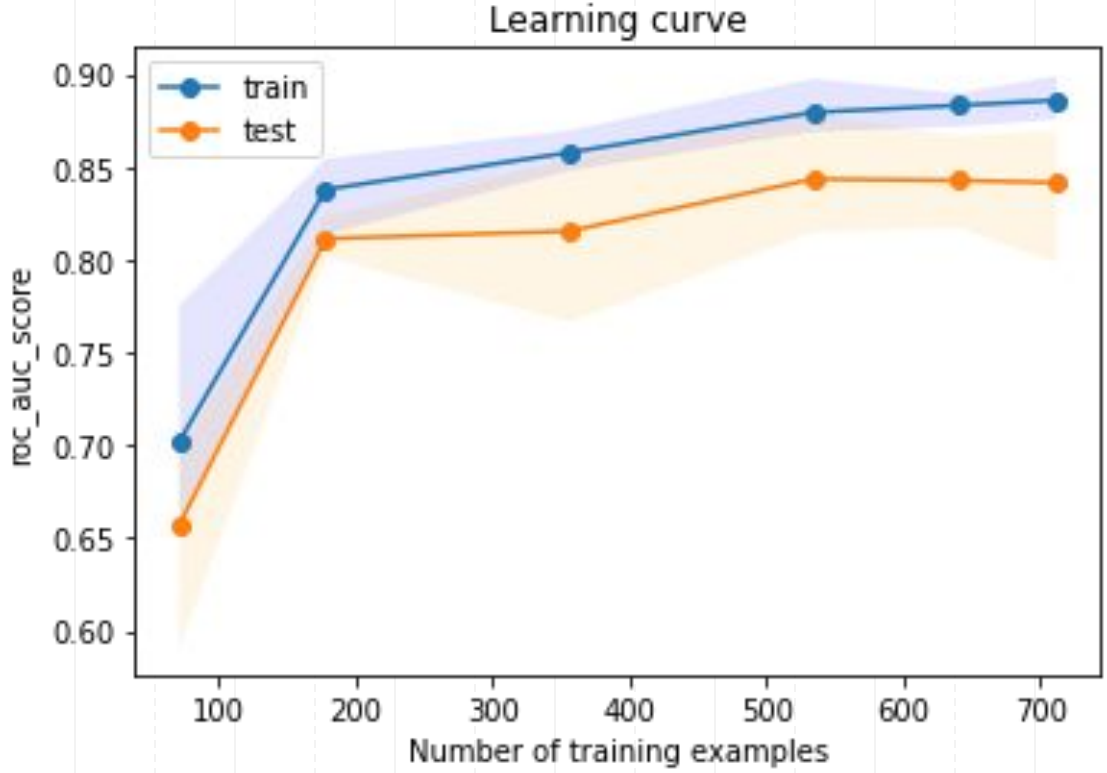
Insufficient

Biased

How much data should we use?

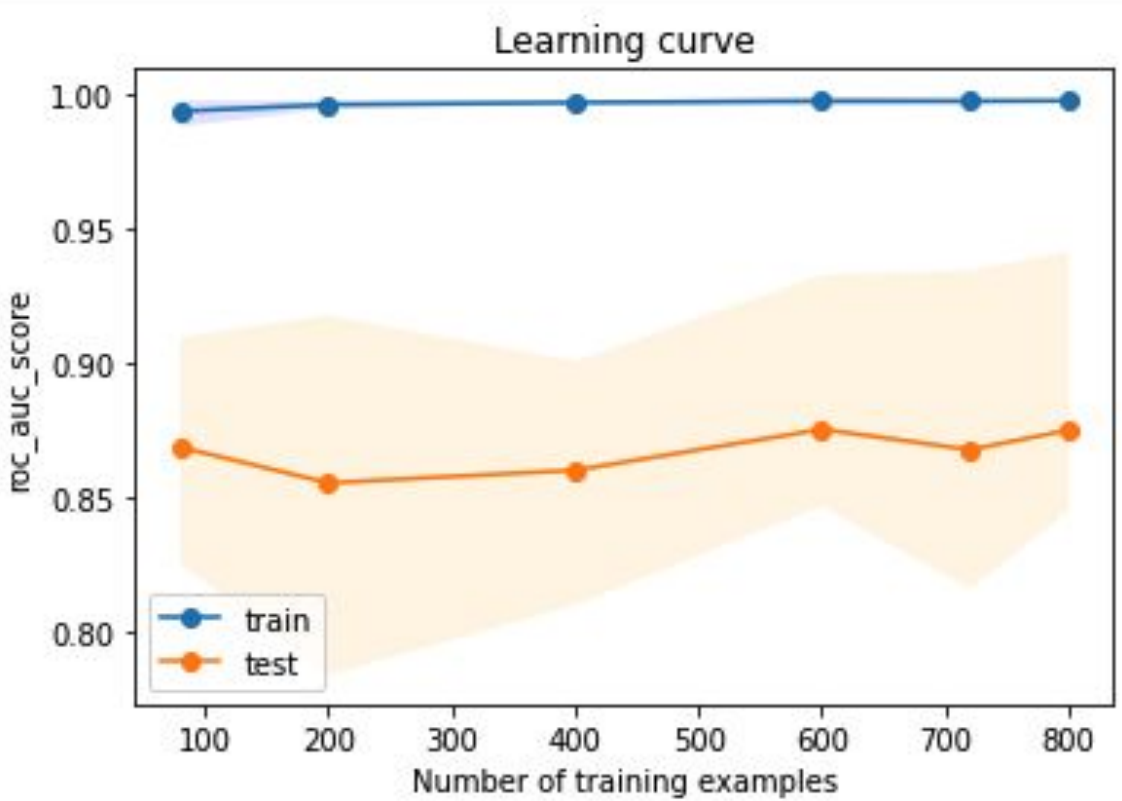
- The obvious starting point about which data to use is... "why not just all of it?"
    - Using too much data makes modelling unnecessarily **slow**
    - The data you have for training may have sample **bias**
    - Data can be **expensive** to come by - Would more data really help improving the model?
    - All the data may **not** be **enough**

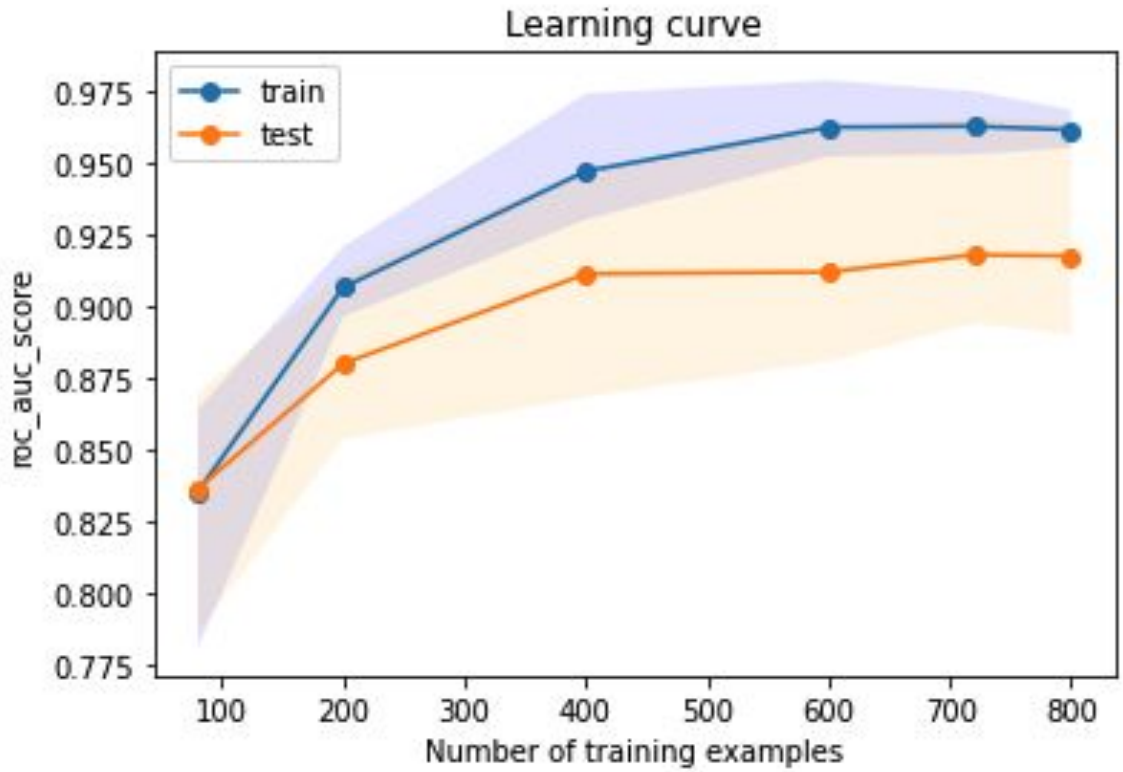
- To understand how our model is learning and how much data it needs:
    - *sklearn.model_selection.learning_curve*.

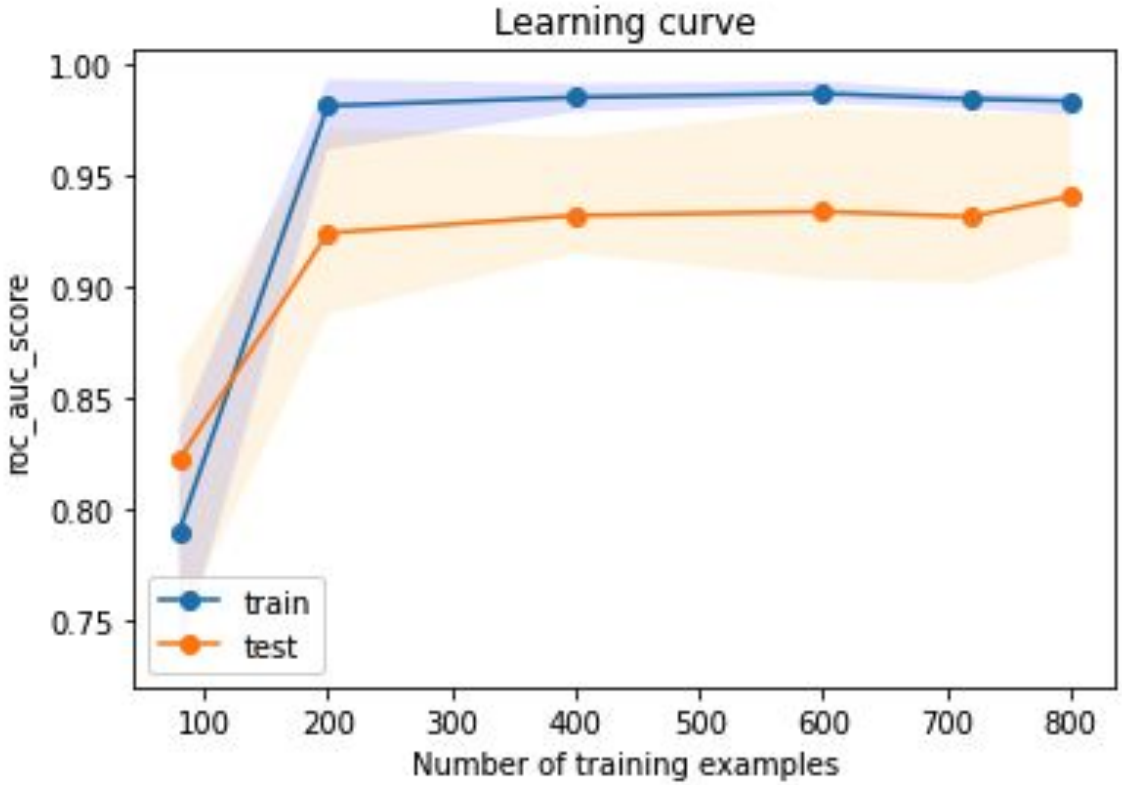

Learning curve
train
test
roc_auc_score
0.90
0.85
0.80
0.75
0.70
0.65
0.60
100
200
300
400
500
600
700
Number of training examples

How much data should we use?

- Learning curves show you how the performance changes with dataset size.
- Performs cross-validation on a metric you provide
- Model variance can also be seen in the learning curves


Learning curve
1.00
0.95
0.90
0.85
0.80
roc_auc_score
train
test
100
200
300
400
500
600
700
800
Number of training examples


Learning curve
train
test
roc_auc_score
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
0.775
100
200
300
400
500
600
700
800
Number of training examples


Learning curve
1.00
0.95
0.90
0.85
0.80
0.75
roc_auc_score
train
test
100
200
300
400
500
600
700
800
Number of training examples

3. Recap

Recap

- There is a tradeoff between **bias and variance**
  - Bias = underfitting = model is always wrong
  - Variance = overfitting = model “memorizes” training data and does not generalize
  - Need a balance between the two

- We can estimate the generalization power of our models by **splitting our data**
  - Train or fit models to training set, evaluate on test set
  - We want both train and test error to be low

- Plot learning curves
  - Don’t use data just because it exists
  - Your model needs your brain more than it needs your CPU

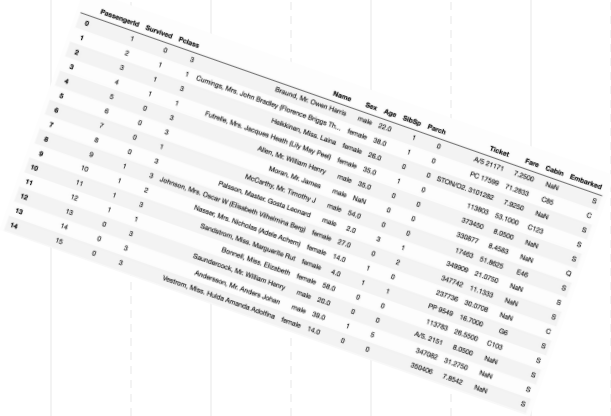
4. Q&A

SLU14 - Model complexity and Overfitting

1. Introduction

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
12	13	0	3	Saundercock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.0500	NaN	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750	NaN	S
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542	NaN	S

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	S
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	NaN	S
12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	G6	C
13	0	3	Saundercock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.0500	C103	S
14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750	NaN	S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542	NaN	S



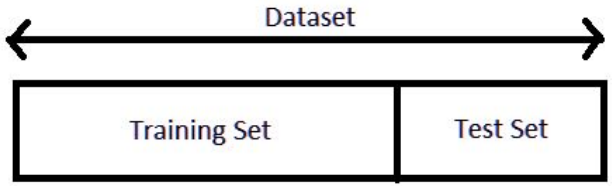
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	S
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	C
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
12	13	1	3	Saundercock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.0500	NaN	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750	NaN	S
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542	NaN	S

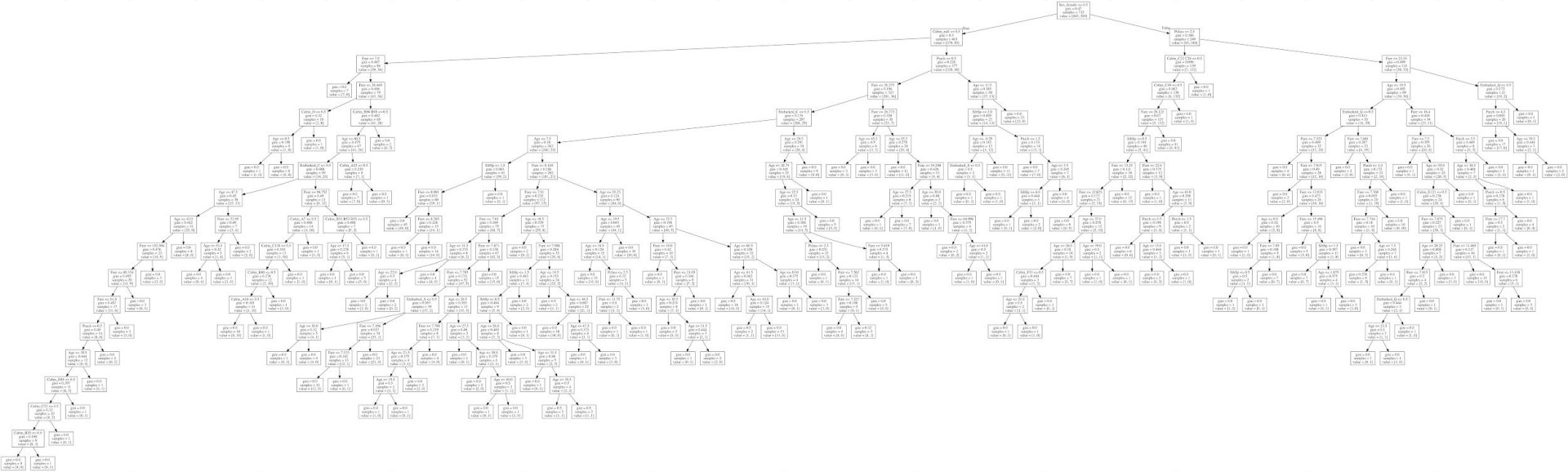

```
1 pd.get_dummies(df['Name']).head(3)
```

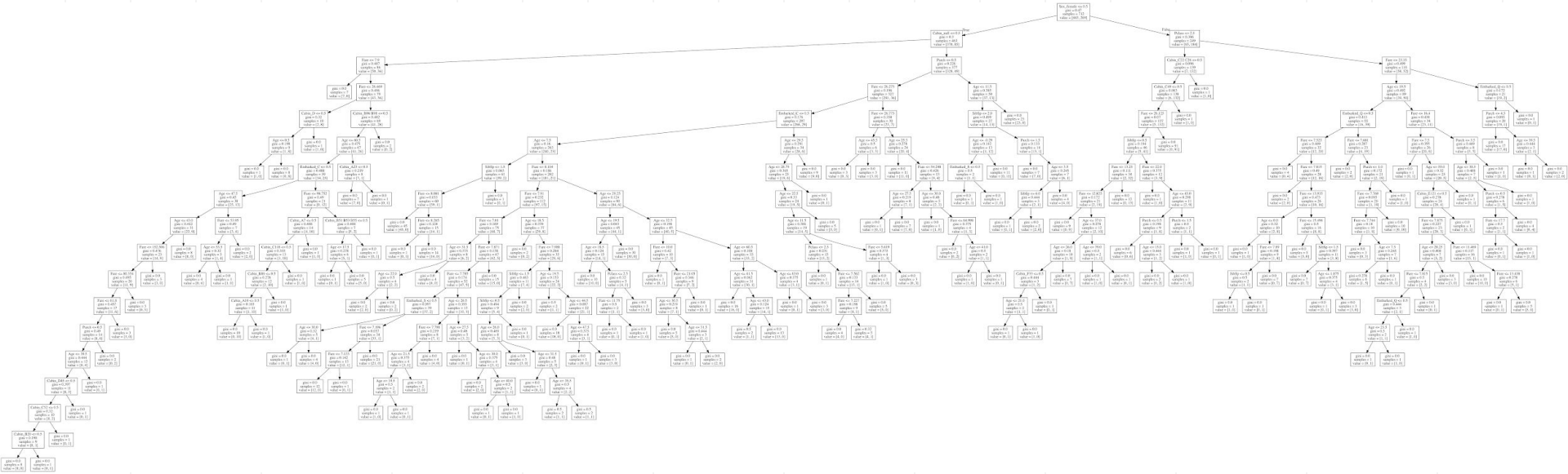
	Abbing, Mr. Anthony	Abbott, Mr. Rossmore Edward	Abbott, Mrs. Stanton (Rosa Hunt)	Abelson, Mr. Samuel	Abelson, Mrs. Samuel (Hannah Wizosky)	Adahl, Mr. Mauritz Nils Martin	Adams, Mr. John	Ahlin, Mrs. Johan (Johanna Persdotter Larsson)	Aks, Mrs. Sam (Leah Rosen)	Albimona, Mr. Nassef Cassem	...	Yrois, Miss. Henriette ("Mrs Harbeck")	Zabour, Miss. Hileni	Zabour, Miss. Thamine	Zimmerman, Mr. Leo	I
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

```
1 pd.get_dummies(df['Name']).head(3)
```

	Abbing, Mr. Anthony	Abbott, Mr. Rossmore Edward	Abbott, Mrs. Stanton (Rosa Hunt)	Abelson, Mr. Samuel	Abelson, Mrs. Samuel (Hannah Wizosky)	Adahl, Mr. Mauritz Nils Martin	Adams, Mr. John	Ahlin, Mrs. Johan (Johanna Persdotter Larsson)	Aks, Mrs. Sam (Leah Rosen)	Albimona, Mr. Nassef Cassem	...	Yrois, Miss. Henriette ("Mrs Harbeck")	Zabour, Miss. Hileni	Zabour, Miss. Thamine	Zimmerman, Mr. Leo	M
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	



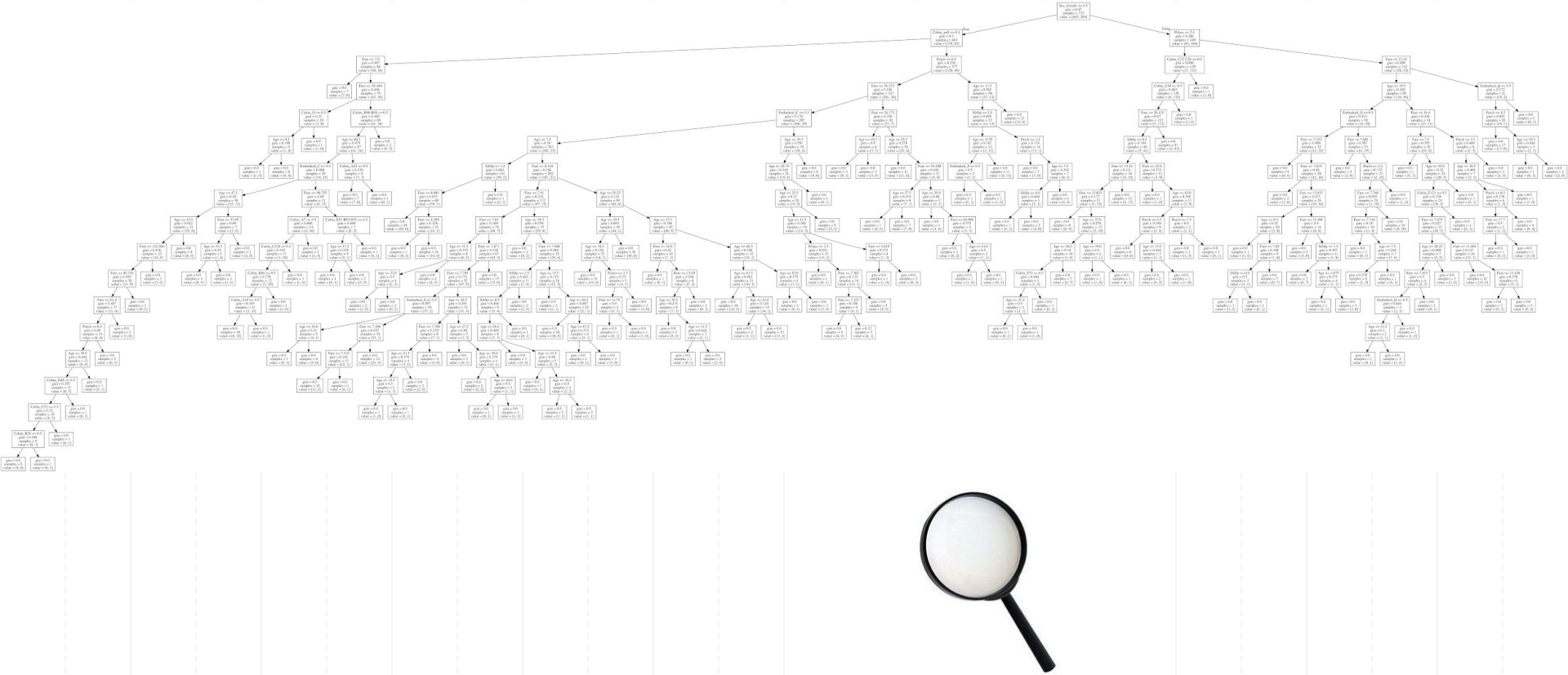


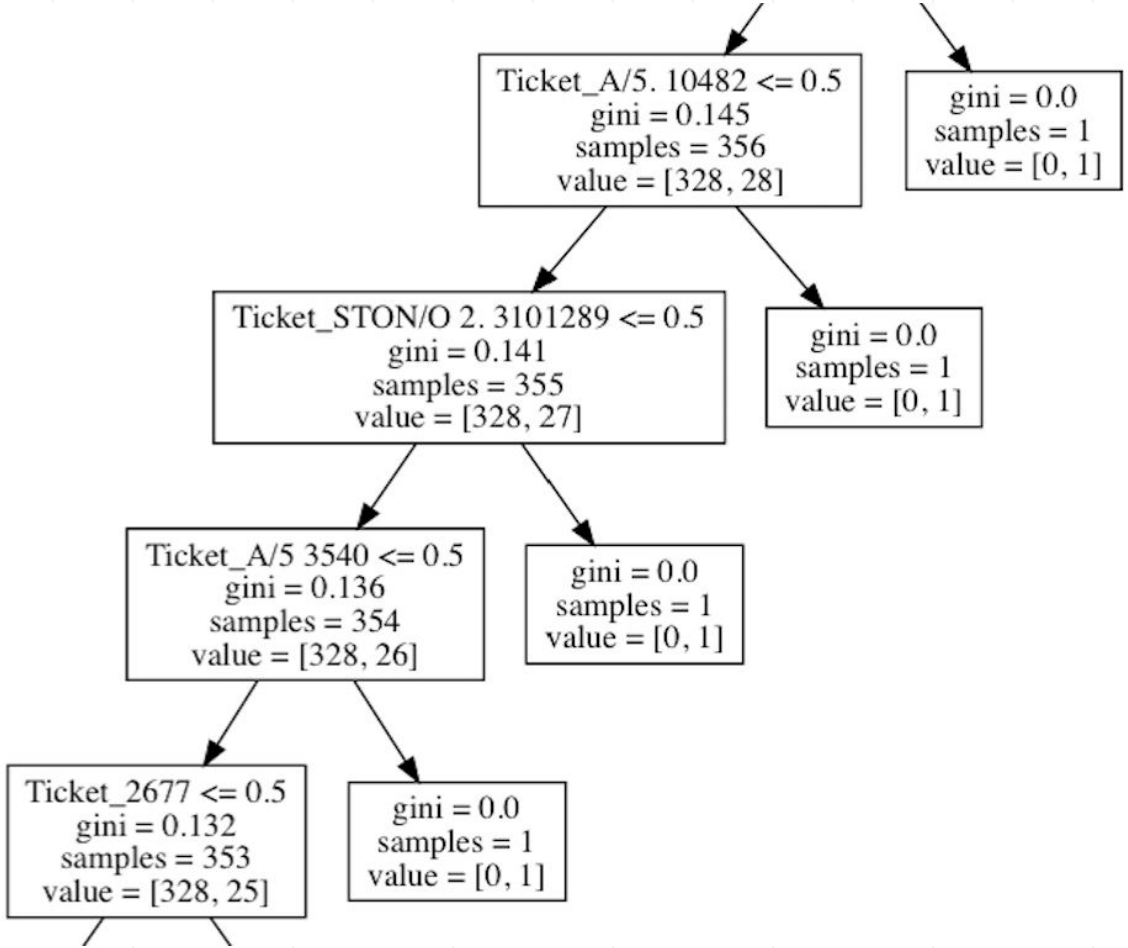


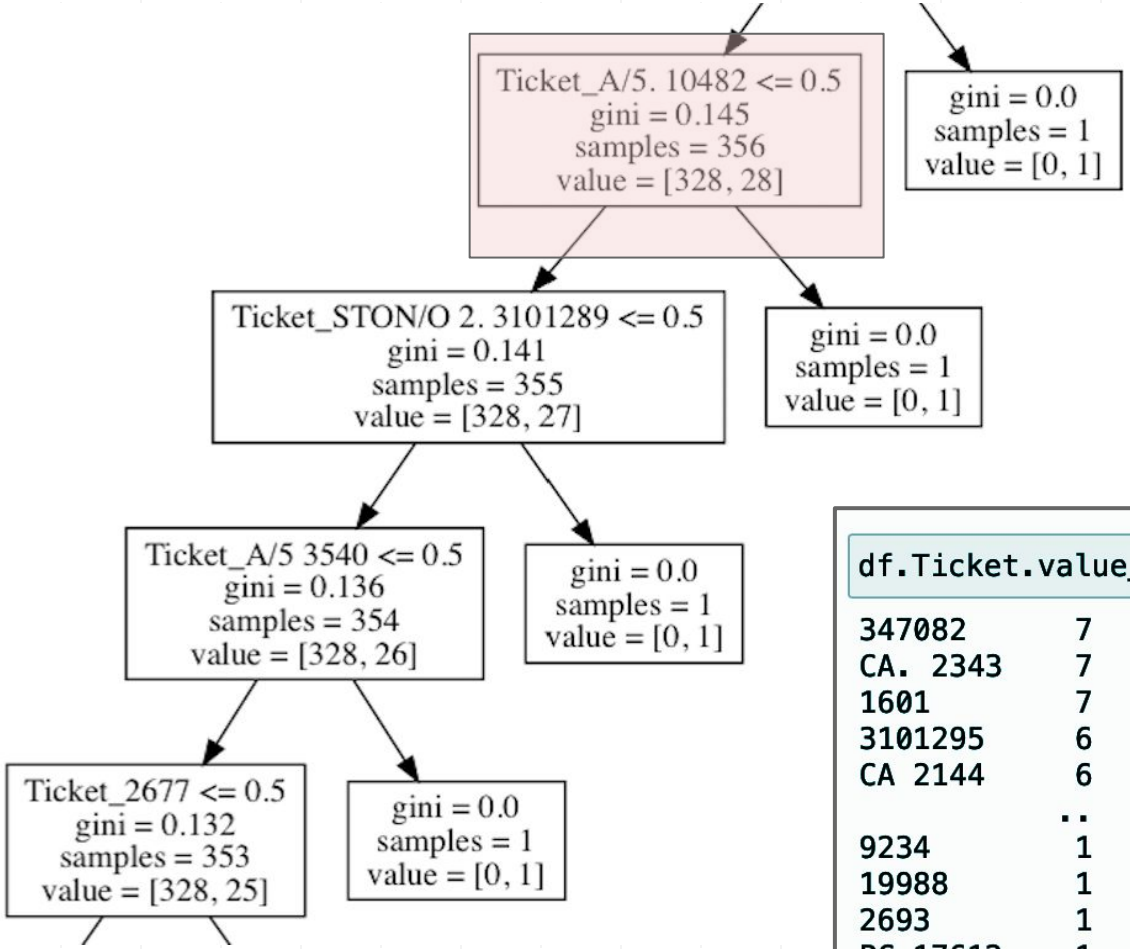
predictions-all-the-features.csv 0.77511

2 days ago by SamHopkins

add submission details


Sex_female <= 0.5
gini = 0.5
samples = 712
value = [445, 309]
True
False
Cabin_null <= 0.5
gini = 0.5
samples = 463
value = [239, 182]
Pclass <= 2.5
gini = 0.391
samples = 249
value = [206, 127]
Fare <= 7.5
gini = 0.482
samples = 294
value = [174, 120]
Fare <= 26.469
gini = 0.444
samples = 78
value = [51, 34]
Cabin_D <= 0.5
gini = 0.5
samples = 10
value = [5, 5]
Cabin_B96 B98 <= 0.5
gini = 0.482
samples = 68
value = [46, 29]
Age <= 6.5
gini = 0.198
samples = 5
value = [1, 8]
gini = 0.0
samples = 5
value = [0, 8]
Age <= 0.5
gini = 0.475
samples = 63
value = [46, 21]
gini = 0.0
samples = 5
value = [0, 5]
Embarked_C <= 0.5
gini = 0.219
samples = 8
value = [7, 1]
Cabin_A37 <= 0.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 47.5
gini = 0.45
samples = 39
value = [25, 14]
Fare <= 30.752
gini = 0.49
samples = 24
value = [13, 11]
Cabin_B51 B53 B55 <= 0.5
gini = 0.486
samples = 17
value = [7, 10]
gini = 0.0
samples = 7
value = [0, 7]
Age <= 0.5
gini = 0.432
samples = 32
value = [22, 10]
Cabin_C118 <= 0.5
gini = 0.444
samples = 3
value = [1, 2]
Age <= 17.5
gini = 0.346
samples = 14
value = [3, 11]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 192.946
gini = 0.408
samples = 28
value = [20, 8]
gini = 0.0
samples = 4
value = [0, 4]
Cabin_C <= 0.5
gini = 0.49
samples = 4
value = [2, 2]
gini = 0.0
samples = 2
value = [2, 0]
gini = 0.0
samples = 2
value = [0, 2]
Cabin_C148 <= 0.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 85.229
gini = 0.397
samples = 30
value = [21, 9]
gini = 0.0
samples = 1
value = [0, 1]
Cabin_B96 <= 0.5
gini = 0.278
samples = 25
value = [20, 5]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 8.2
gini = 0.245
samples = 24
value = [21, 3]
gini = 0.0
samples = 1
value = [0, 1]
Punch <= 0.5
gini = 0.227
samples = 23
value = [21, 2]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 30.5
gini = 0.444
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Cabin_T <= 0.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [0, 1]
Cabin_E <= 0.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [0, 1]
Cabin_A14 <= 0.5
gini = 0.444
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Cabin_B96 <= 0.5
gini = 0.278
samples = 25
value = [20, 5]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 8.05
gini = 0.219
samples = 24
value = [21, 3]
gini = 0.0
samples = 1
value = [0, 1]
Embarked_S <= 0.5
gini = 0.095
samples = 20
value = [19, 1]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 30.0
gini = 0.095
samples = 20
value = [19, 1]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 7.896
gini = 0.05
samples = 19
value = [18, 1]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 7.229
gini = 0.444
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Fare <= 7.775
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 7.896
gini = 0.105
samples = 18
value = [17, 1]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 7.788
gini = 0.059
samples = 17
value = [16, 1]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 27.5
gini = 0.117
samples = 8
value = [7, 1]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 27.5
gini = 0.133
samples = 7
value = [6, 1]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 30.5
gini = 0.245
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Age <= 30.5
gini = 0.355
samples = 4
value = [3, 1]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 30.0
gini = 0.444
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Age <= 27.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 30.5
gini = 0.278
samples = 5
value = [4, 1]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 30.5
gini = 0.375
samples = 4
value = [3, 1]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 30.5
gini = 0.444
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Age <= 30.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 30.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 30.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 26.275
gini = 0.5
samples = 321
value = [181, 140]
Age <= 11.5
gini = 0.494
samples = 58
value = [37, 53]
Embarked_C <= 0.5
gini = 0.175
samples = 297
value = [268, 29]
Fare <= 26.775
gini = 0.298
samples = 93
value = [72, 21]
SibSp <= 2.0
gini = 0.142
samples = 27
value = [2, 25]
Parch <= 0.5
gini = 0.153
samples = 31
value = [28, 3]
Age <= 7.2
gini = 0.128
samples = 283
value = [258, 25]
Age <= 29.2
gini = 0.291
samples = 28
value = [24, 4]
Age <= 25.5
gini = 0.5
samples = 4
value = [2, 2]
Fare <= 8.05
gini = 0.375
samples = 3
value = [1, 2]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 94.248
gini = 0.111
samples = 24
value = [22, 2]
Embarked_S <= 0.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 8.108
gini = 0.142
samples = 262
value = [240, 22]
SibSp <= 1.5
gini = 0.124
samples = 251
value = [232, 19]
Fare <= 8.05
gini = 0.444
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Age <= 20.75
gini = 0.095
samples = 20
value = [19, 1]
gini = 0.0
samples = 9
value = [9, 0]
Age <= 22.5
gini = 0.18
samples = 11
value = [10, 1]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 21.5
gini = 0.408
samples = 4
value = [3, 1]
gini = 0.0
samples = 2
value = [2, 0]
Age <= 30.5
gini = 0.48
samples = 5
value = [3, 2]
gini = 0.0
samples = 2
value = [0, 2]
gini = 0.0
samples = 3
value = [3, 0]
Age <= 1.5
gini = 0.219
samples = 7
value = [6, 1]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 8.05
gini = 0.133
samples = 14
value = [13, 1]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 8.05
gini = 0.111
samples = 18
value = [16, 2]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 7.91
gini = 0.105
samples = 17
value = [16, 1]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 7.854
gini = 0.153
samples = 12
value = [11, 1]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 22.5
gini = 0.444
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Age <= 22.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 22.5
gini = 0.48
samples = 5
value = [3, 2]
gini = 0.0
samples = 2
value = [0, 2]
gini = 0.0
samples = 3
value = [3, 0]
Age <= 11.5
gini = 0.086
samples = 22
value = [21, 1]
gini = 0.0
samples = 1
value = [0, 1]
Pclass <= 2.5
gini = 0.278
samples = 5
value = [4, 1]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 7.925
gini = 0.219
samples = 24
value = [21, 3]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 8.05
gini = 0.198
samples = 9
value = [8, 1]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 7.91
gini = 0.18
samples = 10
value = [9, 1]
gini = 0.0
samples = 1
value = [0, 1]
Fare <= 7.854
gini = 0.245
samples = 6
value = [5, 1]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 22.5
gini = 0.32
samples = 5
value = [4, 1]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 22.5
gini = 0.278
samples = 5
value = [4, 1]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 22.5
gini = 0.444
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Age <= 22.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 22.5
gini = 0.444
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Age <= 22.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 25.25
gini = 0.125
samples = 90
value = [84, 6]
Age <= 32.5
gini = 0.198
samples = 45
value = [40, 5]
Age <= 29.5
gini = 0.095
samples = 40
value = [38, 2]
Age <= 18.5
gini = 0.239
samples = 14
value = [12, 2]
Fare <= 7.741
gini = 0.049
samples = 40
value = [39, 1]
Fare <= 7.831
gini = 0.444
samples = 3
value = [2, 1]
Age <= 31.5
gini = 0.444
samples = 3
value = [2, 1]
Fare <= 7.788
gini = 0.5
samples = 2
value = [1, 1]
SibSp <= 0.5
gini = 0.444
samples = 3
value = [2, 1]
Age <= 30.5
gini = 0.355
samples = 4
value = [3, 1]
Age <= 27.5
gini = 0.5
samples = 2
value = [1, 1]
Age <= 30.5
gini = 0.444
samples = 3
value = [2, 1]
Age <= 30.5
gini = 0.375
samples = 4
value = [3, 1]
Age <= 30.5
gini = 0.444
samples = 3
value = [2, 1]
Age <= 30.5
gini = 0.5
samples = 2
value = [1, 1]
Age <= 30.5
gini = 0.444
samples = 3
value = [2, 1]
Age <= 30.5
gini = 0.5
samples = 2
value = [1, 1]
Age <= 30.5
gini = 0.444
samples = 3
value = [2, 1]
Age <= 30.5
gini = 0.5
samples = 2
value = [1, 1]
Fare <= 7.896
gini = 0.444
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Fare <= 7.788
gini = 0.444
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Age <= 27.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Age <= 30.5
gini = 0.444
samples = 3
value = [2, 1]
gini = 0.0
samples = 2
value = [2, 0]
Age <= 30


Ticket_A/5. 10482 <= 0.5
gini = 0.145
samples = 356
value = [328, 28]
gini = 0.0
samples = 1
value = [0, 1]
Ticket_STON/O 2. 3101289 <= 0.5
gini = 0.141
samples = 355
value = [328, 27]
gini = 0.0
samples = 1
value = [0, 1]
Ticket_A/5 3540 <= 0.5
gini = 0.136
samples = 354
value = [328, 26]
gini = 0.0
samples = 1
value = [0, 1]
Ticket_2677 <= 0.5
gini = 0.132
samples = 353
value = [328, 25]
gini = 0.0
samples = 1
value = [0, 1]




```
df.Ticket.value_counts()
```

```
347082      7
CA. 2343    7
1601        7
3101295     6
CA 2144     6
           ..
9234        1
19988       1
2693        1
PC 17612    1
370376      1
```


WHOOPSIES...

Motivation

Using data blindly is not a wise approach...

Motivation

What issues can arise?

- Unexpected behavior
- Overfitting (poor future performance)
- Lack of interpretability
- Unnecessary model complexity (long training time)

Solution!

There are many, here we are going to look into:

- Classic method of:

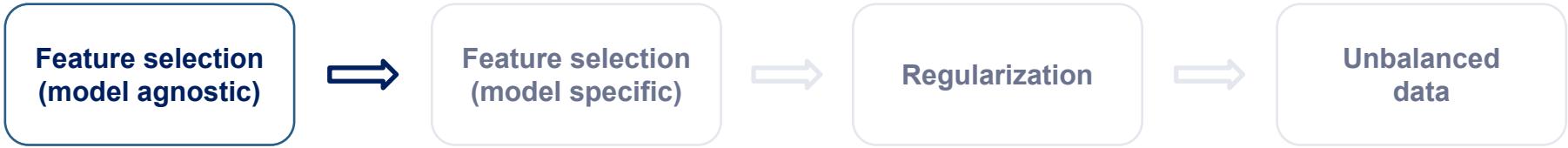
  'Let's look into the features and make sense of them.'

- **Regularisation** of Linear Models

Overview

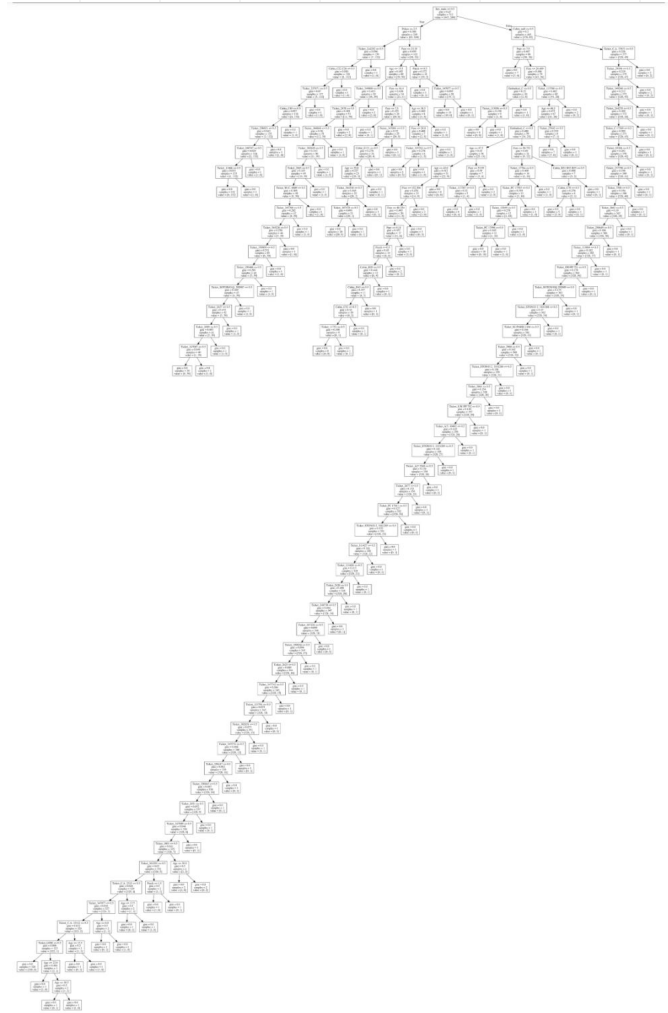
1. **Feature Selection**
    a. Model agnostic
    b. Model specific
2. **Regularization**
    a. Intuition and use-cases
    b. Types of regularization
3. **Unbalanced data**
    a. Recognizing it
    b. Avoiding some traps
    c. Oversampling and undersampling

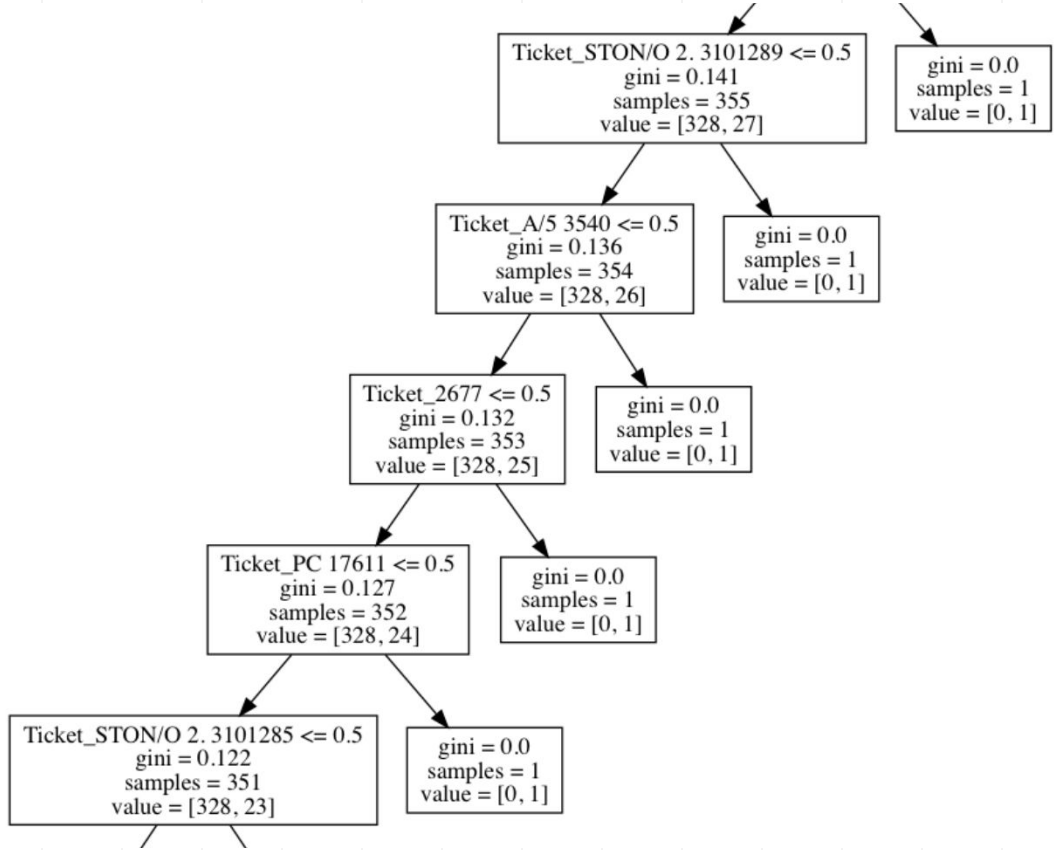
2. Topic Explanation


Feature selection (model agnostic)
Feature selection (model specific)
Regularization
Unbalanced data

Exploring feature information

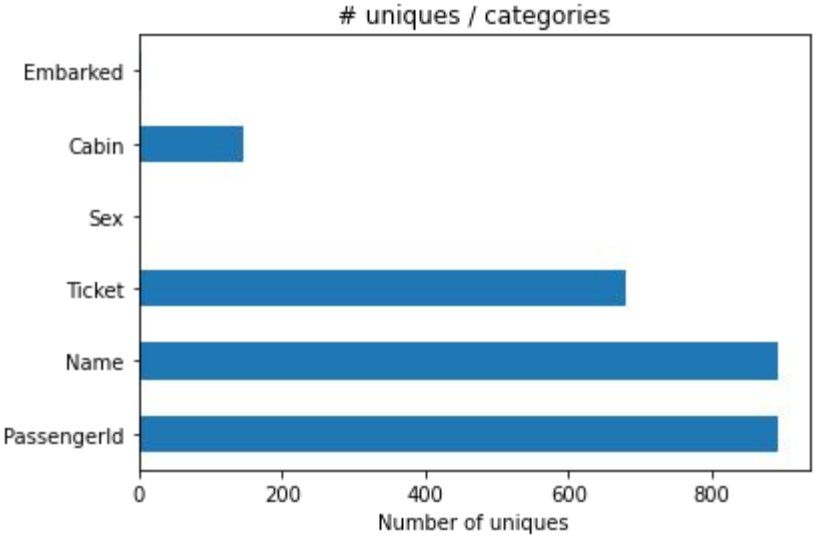
In Decision Trees, **looking at the first node** in the tree will most likely tell us what is one of the most important features to define the data




Ticket_STON/O 2. 3101289 <= 0.5
gini = 0.141
samples = 355
value = [328, 27]
gini = 0.0
samples = 1
value = [0, 1]
Ticket_A/5 3540 <= 0.5
gini = 0.136
samples = 354
value = [328, 26]
gini = 0.0
samples = 1
value = [0, 1]
Ticket_2677 <= 0.5
gini = 0.132
samples = 353
value = [328, 25]
gini = 0.0
samples = 1
value = [0, 1]
Ticket_PC 17611 <= 0.5
gini = 0.127
samples = 352
value = [328, 24]
gini = 0.0
samples = 1
value = [0, 1]
Ticket_STON/O 2. 3101285 <= 0.5
gini = 0.122
samples = 351
value = [328, 23]
gini = 0.0
samples = 1
value = [0, 1]

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
160	0	3	Sage, Master. Thomas Henry	male	NaN	8	2	CA. 2343	69.55	NaN	S
181	0	3	Sage, Miss. Constance Gladys	female	NaN	8	2	CA. 2343	69.55	NaN	S
202	0	3	Sage, Mr. Frederick	male	NaN	8	2	CA. 2343	69.55	NaN	S
325	0	3	Sage, Mr. George John Jr	male	NaN	8	2	CA. 2343	69.55	NaN	S
793	0	3	Sage, Miss. Stella Anna	female	NaN	8	2	CA. 2343	69.55	NaN	S
847	0	3	Sage, Mr. Douglas Bullen	male	NaN	8	2	CA. 2343	69.55	NaN	S
864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.55	NaN	S

Exploring feature information



- By doing some exploratory analysis we can see indeed that the *Name*, *Ticket* and *PassengerId* have too many values
- They won't add much information

Exploring feature information

You have a World model in your mind. **Use it.**

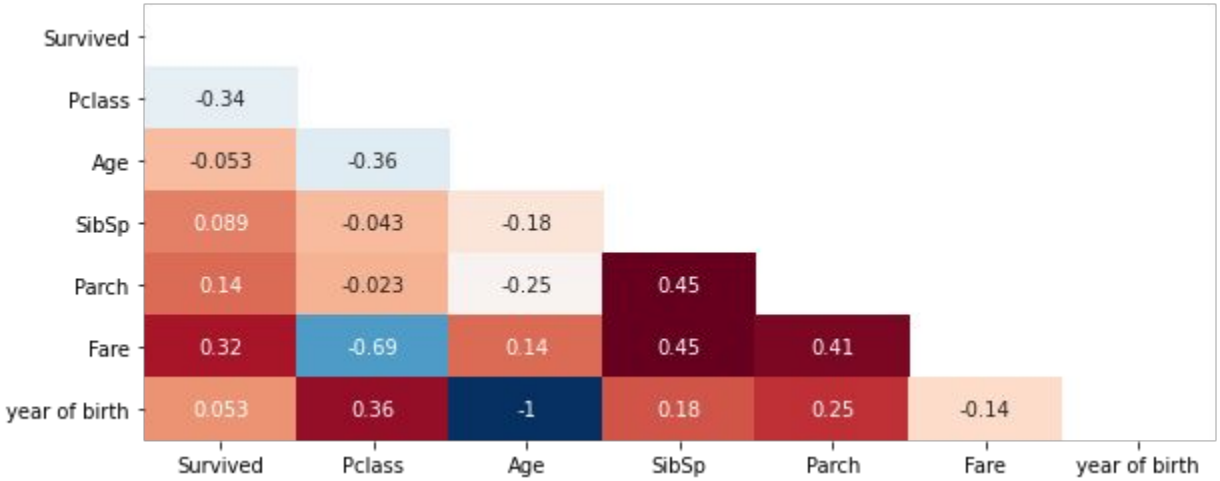
Common sense

Tree visualizations

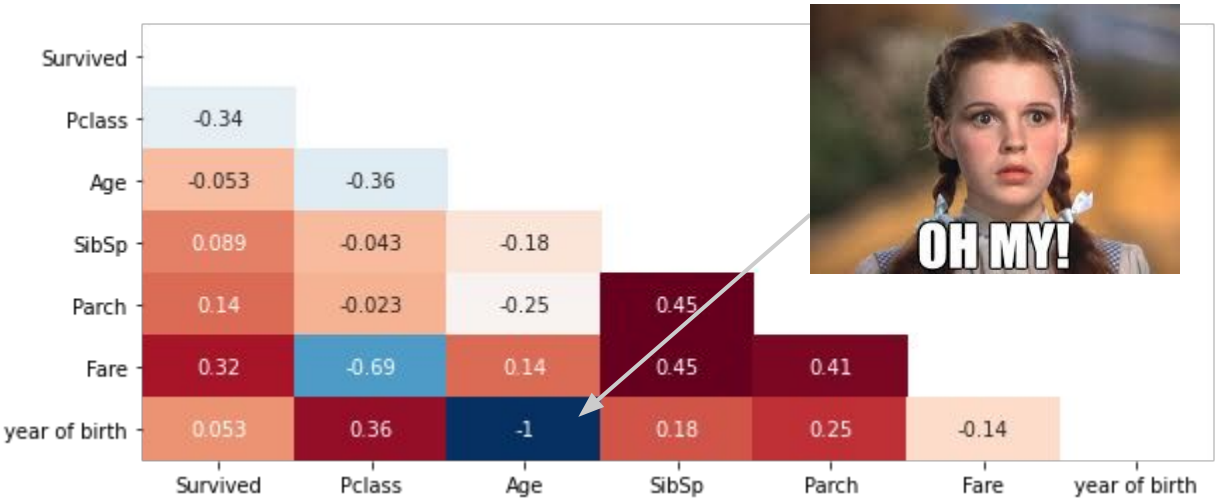
Correlation

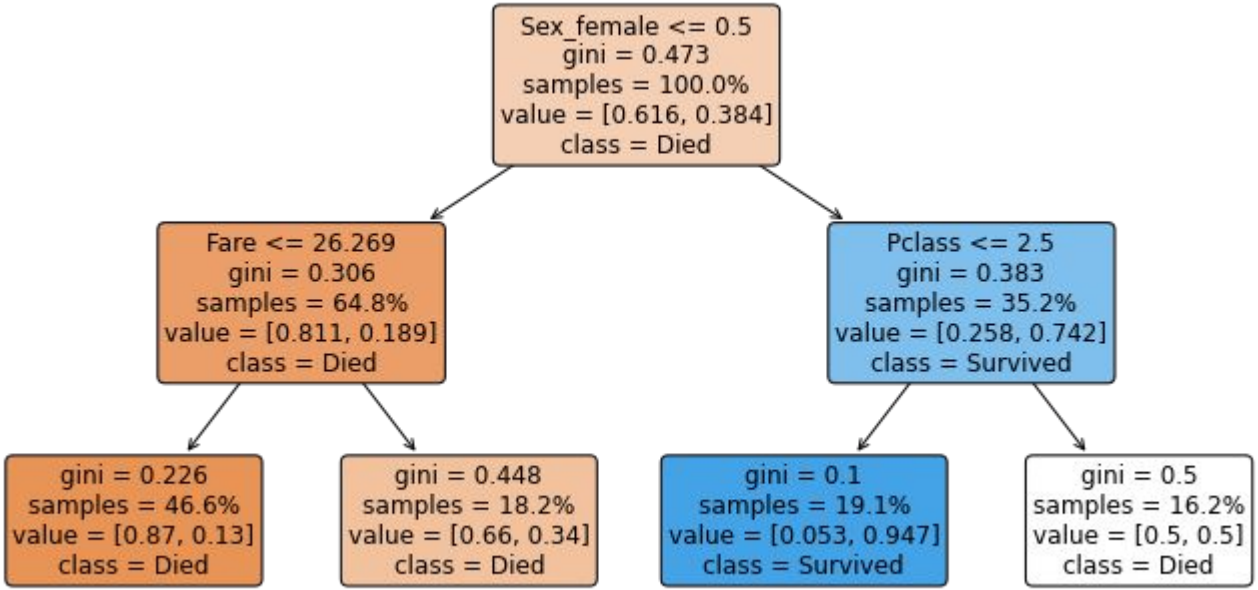
Mutual information

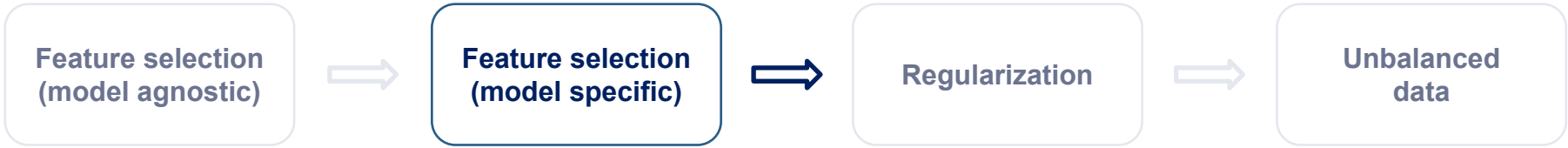
Correlation Matrix



Correlation Matrix

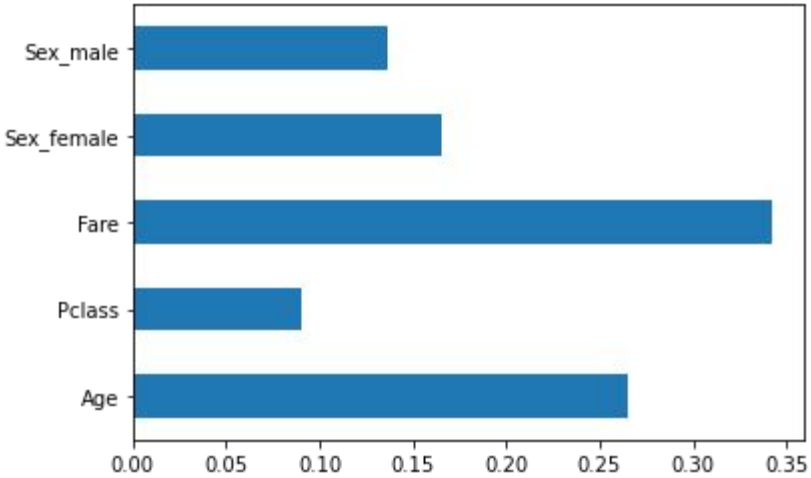



Sex_female <= 0.5
gini = 0.473
samples = 100.0%
value = [0.616, 0.384]
class = Died
Fare <= 26.269
gini = 0.306
samples = 64.8%
value = [0.811, 0.189]
class = Died
Pclass <= 2.5
gini = 0.383
samples = 35.2%
value = [0.258, 0.742]
class = Survived
gini = 0.226
samples = 46.6%
value = [0.87, 0.13]
class = Died
gini = 0.448
samples = 18.2%
value = [0.66, 0.34]
class = Died
gini = 0.1
samples = 19.1%
value = [0.053, 0.947]
class = Survived
gini = 0.5
samples = 16.2%
value = [0.5, 0.5]
class = Died

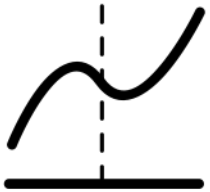

Feature selection (model agnostic)
Feature selection (model specific)
Regularization
Unbalanced data

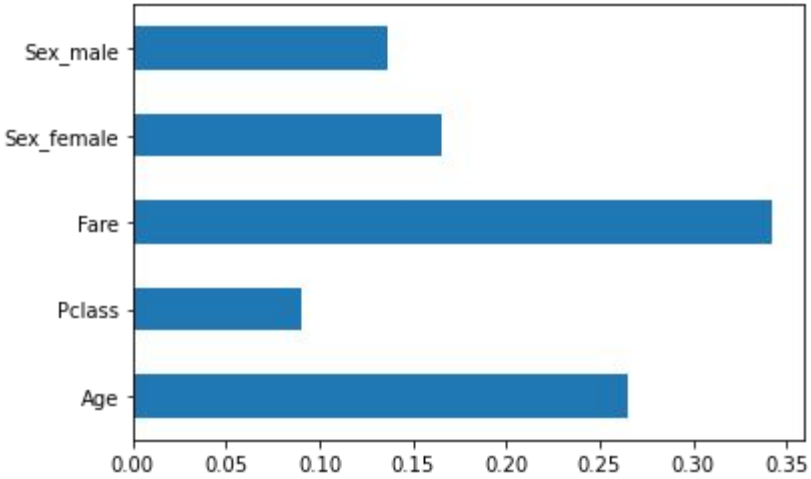
Feature importance in models

- One type of feature selection is that of looking at how important a feature is to a model that was already trained
- You can do this in models through the sets of weights it associates with each feature

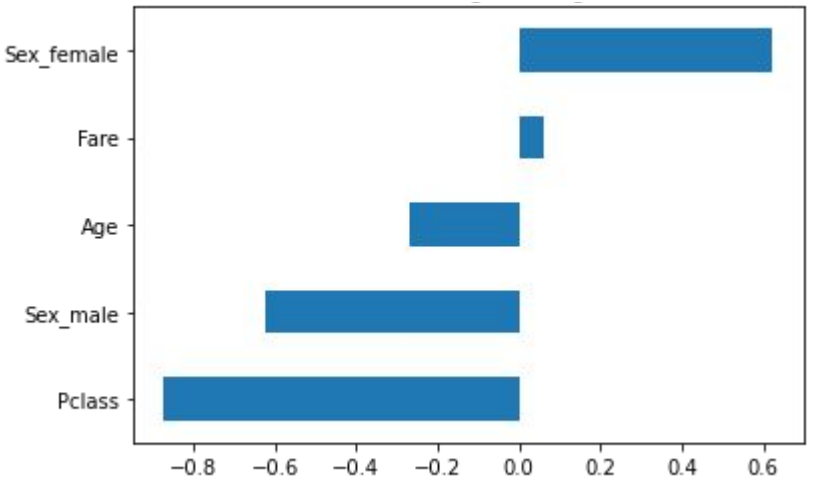


Importances

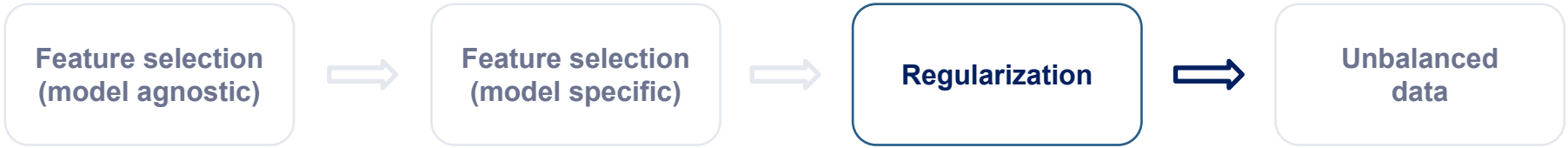




Importances



Coefficients

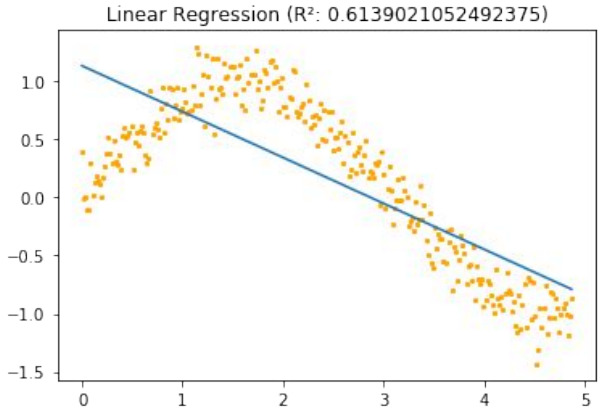

Feature selection (model agnostic)
Feature selection (model specific)
Regularization
Unbalanced data

Regularization

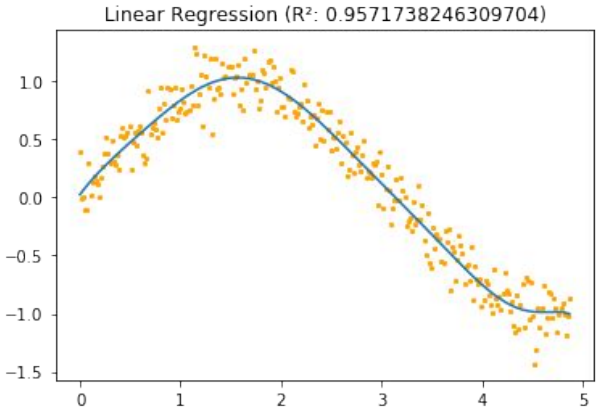
- When you have a dataset that is (a) small, (b) has way more features than observations, or (c) is sparse data, regularization can help a lot
- To force the model to use fewer features and still get similar results
- Regularization strength is a parameter you must set (i.e. not automatically optimized). It is a hyperparameter

Regularization Intuition

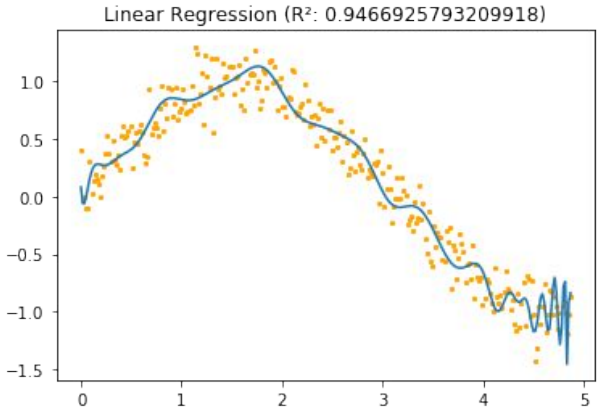
1-degree polynomial



9-degree polynomial



200-degree polynomial



Too many features??
Can we pick the useful ones automatically?

L_2 Loss Function

We need a new loss that penalizes the growth of useless feature coefficients...

$$L = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 + \lambda_2 \|\beta\|_2^2$$

The bigger this number is, the bigger the cost of growing the magnitude of coefficients

$$= \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 + \lambda_2 \sum_{k=1}^{K}\beta_k^2$$

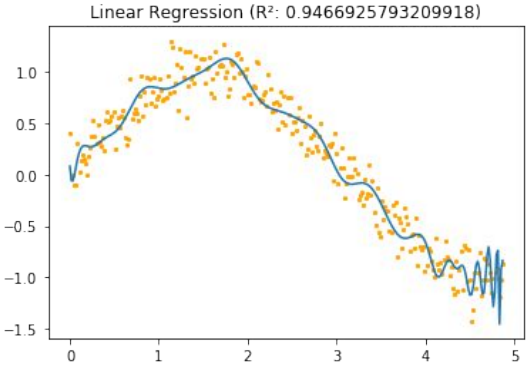
This is called L_2 loss

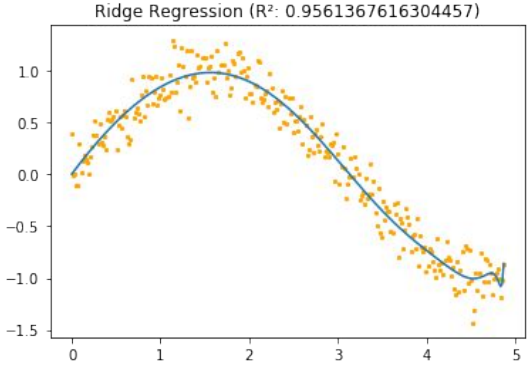
Low: good fit
High: bad fit

Low: simple model
High: complex model

Ridge Regression

Linear Regression: 200-degree polynomial

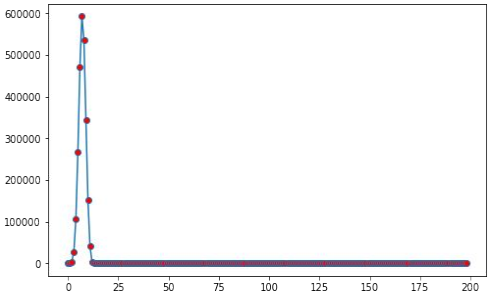


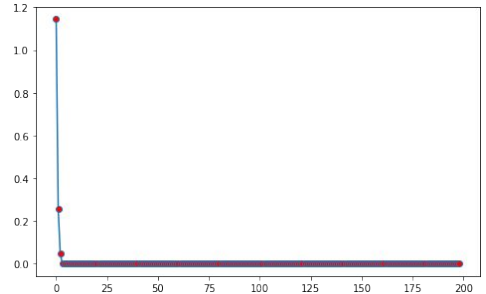


Ridge Regression: 200-degree polynomial

Uses L_2 loss

Plot of feature coefficients





Plot of feature coefficients

- **The model is less overfit**
- **Most features are zero**

Lasso and Elastic Net Regression

Lasso:

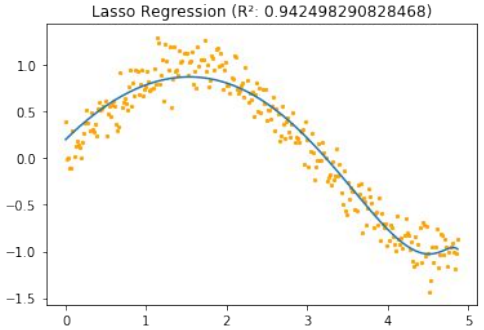
$$J = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 + \boxed{\lambda_1 \sum_{k=1}^{K} |\beta_k|}$$

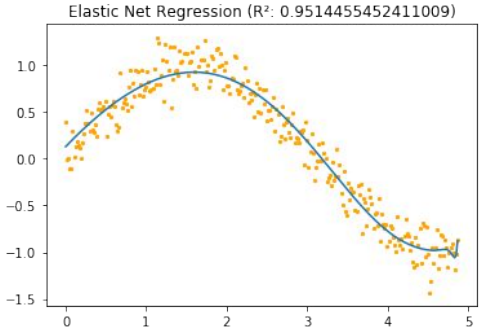
This is called L_1 loss

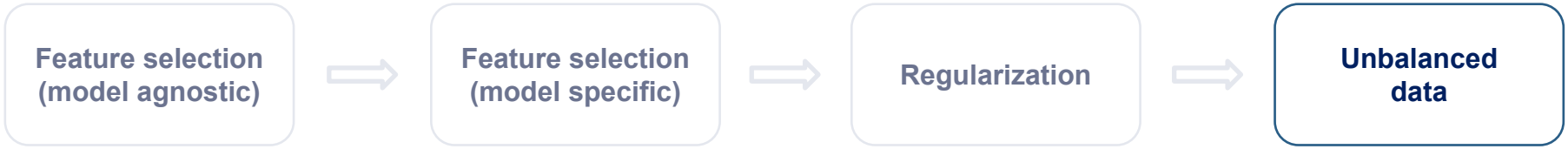
Elastic Net:

$$J = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 + \boxed{\lambda_1 \sum_{k=1}^{K} |\beta_k| + \lambda_2 \sum_{k=1}^{K} \beta_k^2}$$

$L_1 + L_2$ loss

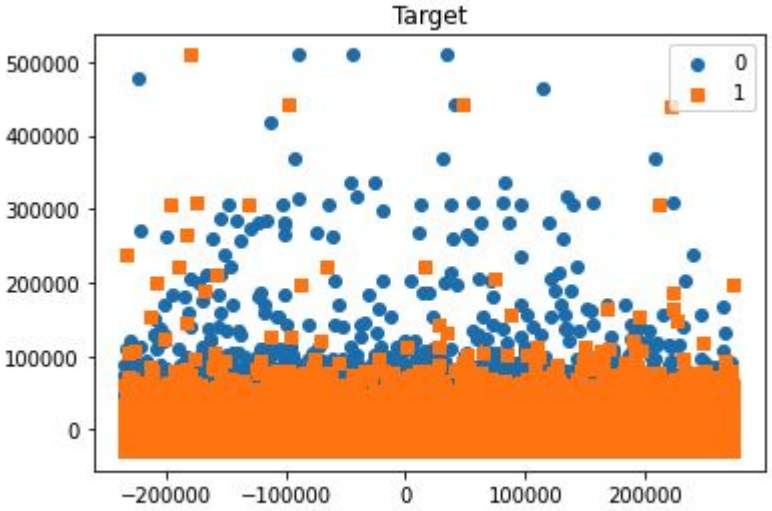





Feature selection (model agnostic)
Feature selection (model specific)
Regularization
Unbalanced data

Unbalanced data

- A lot of models work better when the data is balanced
- Traps:
  - Splits
  - Metrics
  - Training



Unbalanced data splits

- The first thing you should be careful with are splits

```
X_left, X_test, y_left, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

X_train, X_val, y_train, y_val = train_test_split(
    X_left, y_left, test_size=0.2/0.8, random_state=42
)
```

Unbalanced data splits

- The first thing you should be careful with are splits

```
X_left, X_test, y_left, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

X_train, X_val, y_train, y_val = train_test_split(
    X_left, y_left, test_size=0.2/0.8, random_state=42
)
```

```
Size of train: 120
Number of class 1 examples in train: 16 (13.33 %)
Size of validation: 40
Number of class 1 examples in validation: 6 (15.0 %)
Size of test: 40
Number of class 1 examples in test: 8 (20.0 %)
```

- It’s easy to get a poor split that impacts your full process if you are not paying attention

Unbalanced data

- The first thing you should be careful with are splits

```
X_left, X_test, y_left, y_test = train_test_split(
    X, y, test_size=0.2, random_state=1234, stratify=y
)

X_train, X_val, y_train, y_val = train_test_split(
    X_left, y_left, test_size=0.2/0.8, random_state=1234, stratify=y_left
)
```

Unbalanced data splits

- The first thing you should be careful with are splits

```
X_left, X_test, y_left, y_test = train_test_split(
    X, y, test_size=0.2, random_state=1234, stratify=y
)

X_train, X_val, y_train, y_val = train_test_split(
    X_left, y_left, test_size=0.2/0.8, random_state=1234, stratify=y_left
)
```

```
Size of train: 120
Number of class 1 examples in train: 18 (15.0 %)
Size of validation: 40
Number of class 1 examples in validation: 6 (15.0 %)
Size of test: 40
Number of class 1 examples in test: 6 (15.0 %)
```

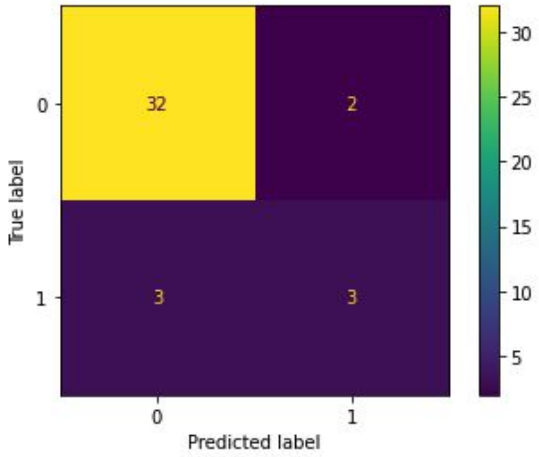
Metrics

Which one is better?

```
Accuracy: 87.50%
```

```
Majority class
Precision: 0.91
Rcall: 0.94
F1-Score: 0.93
==============
Minority class
Precision: 0.60
Recall: 0.50
F1-Score: 0.55
```

Validation

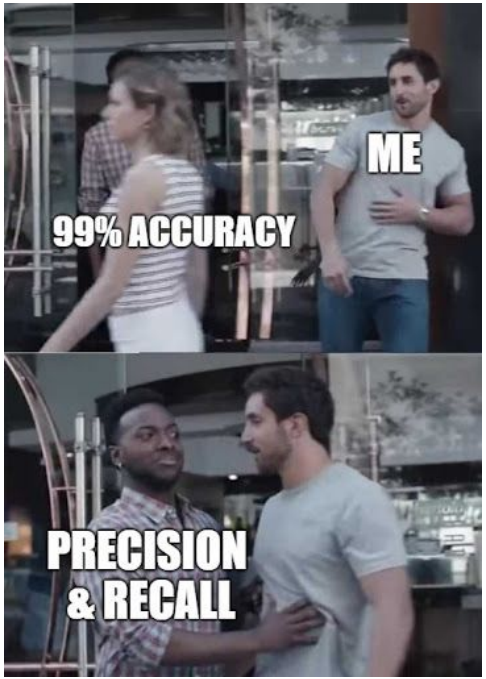


Metrics

Which one is better?

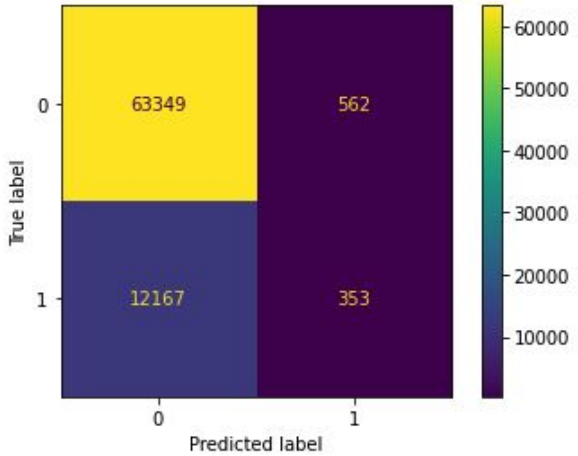
```
Accuracy: 87.50%
```

```
Majority class
Precision: 0.91
Rcall: 0.94
F1-Score: 0.93
==============
Minority class
Precision: 0.60
Recall: 0.50
F1-Score: 0.55
```

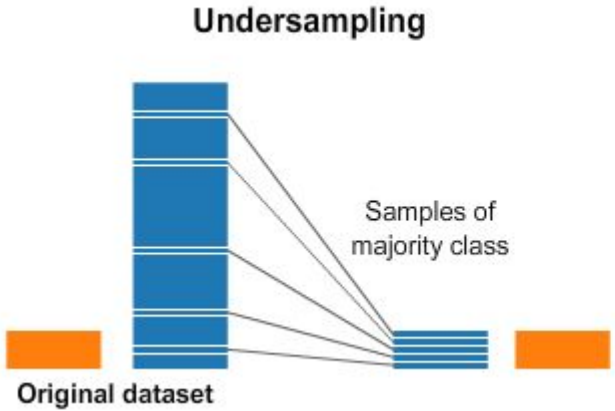


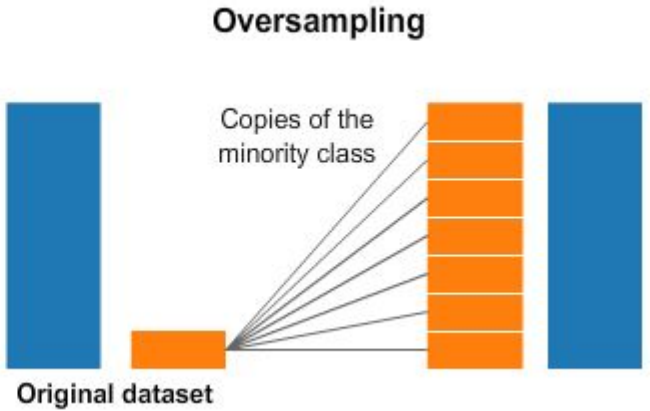
Training with unbalanced data

```
Majority class
Precision: 0.84
Rcall: 0.99
F1-Score: 0.91
==============
Minority class
Precision: 0.39
Recall: 0.03
F1-Score: 0.05
```

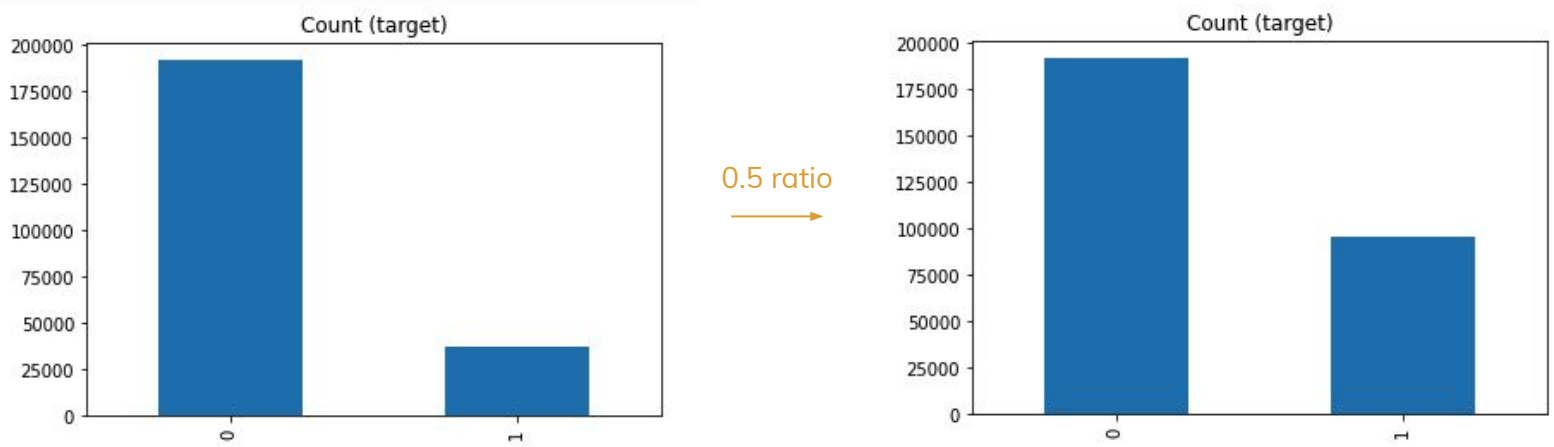


Resampling


Undersampling
Samples of
majority class
Original dataset

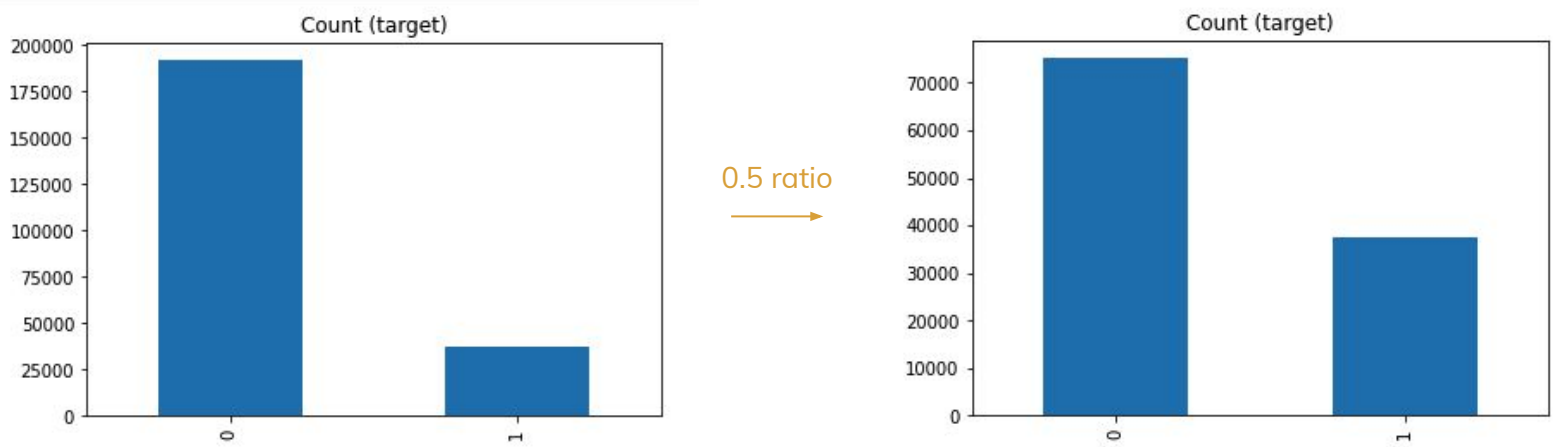

Oversampling
Copies of the
minority class
Original dataset

Resampling - oversampling




```
ros_sampler = RandomOverSampler(sampling_strategy=.5, random_state=42)
X_over, y_over = ros_sampler.fit_resample(X_train, y_train)
```

Resampling - undersampling

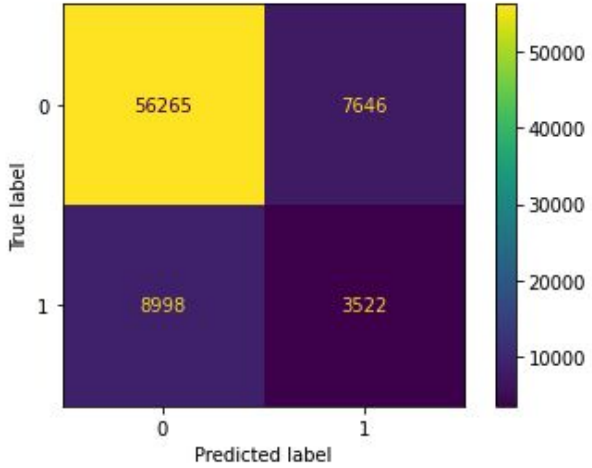



```
rus_sampler = RandomUnderSampler(sampling_strategy=.5, random_state=42)
X_under, y_under = rus_sampler.fit_resample(X_train, y_train)
```

Training with resampling

Trade-off

```
Majority class
Precision: 0.86
Recall: 0.88
F1-Score: 0.87
==============
Minority class
Precision: 0.32
Recall: 0.28
F1-Score: 0.30
```



3. Recap

Recap

- 🔍 Not all features are equally important
  - Understand each feature
  - Remove suspect features
  - Examine the features independently
  - Plot a few trees!
  - When modelling, check out the feature importances
- 🛡️ Regularization can help to combat overfitting
- ⚖️ Analyze your data for imbalance
  - Be careful with the data stratification and metric chosen
  - Use resampling techniques to train your model
  - Do not apply the **resampling** to the full dataset to avoid biasing your evaluation

4. Q&A

SLU 15 - HYPER Parameter Tuning

1. Introduction

Motivation

- Models can be *tweaked* and modified to better adapt to our specific problem.
- This “fine tuning” can be done in a systematic manner.

Overview

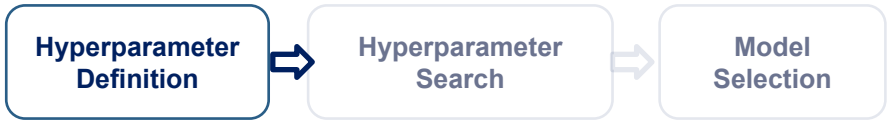
- Short recap
- Hyperparameter definition
- Hyperparameter search
- Model selection

Short Recap

1. Problem Definition
2. Data Preparation
3. Feature Extraction
4. Model training
5. Model testing

Can we do better?

2. Topic Explanation


Hyperparameter Definition
Hyperparameter Search
Model Selection

So… what are HYPER parameters?

Hyperparameters Definition

parameters

From the greek word hyper: over, above

Parameters that are above? Yes!

The parameters that define the structure of our estimators

…and set how the estimator learns.

Logistic Regression Example

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

parameters

HYPER parameters

`sklearn.linear_model.`LogisticRegression(***penalty=****'l2'*, ***dual=****False,*
tol=*0.0001*, ***C=****1.0*, ***fit_intercept=****True*, ***intercept_scaling=****1*,
class_weight=*None*, ***solver=****'warn'*, ***max_iter=****100*, ***warm_start=****False*)

Cheatsheet

Estimator	**Hyperparameters**
Logistic Regression	penalty type (l1 or l2), C (regularization parameter)
SVMs	kernel, C (regularization parameter)
Tree Ensembles	n_estimators, max_depth
KNN	n_neighbors, weights


So many choices
So little time!

Don’t despair yet!

HYPER parameter tuning can be automated!


Hyperparameter Definition
Hyperparameter Search
Model Selection

Hyperparameter search

General Hyper Parameter tuning algorithm

1. Split **train data** into **train** and **validation** subsets (K-folds);
2. Select hyperparameter values;
3. Instantiate estimator with selected values;
4. **Train estimator on training subset** ;
5. **Score estimator on the validation subset**;
6. Repeat 2-5 for **all desired hyperparameter combinations.**

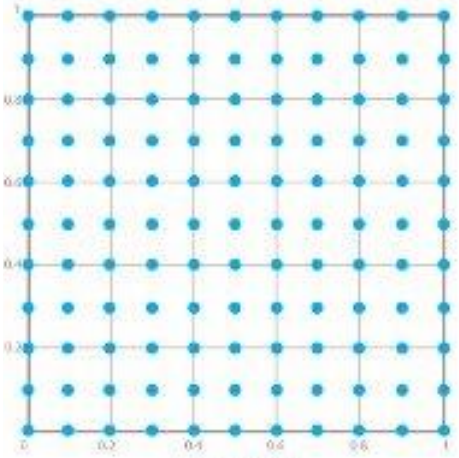
Goal: Select estimator with the best score on the validation metric.

Hyperparameter search (2)

Parameter Search Space- Defines the region of hyperparameters we want to test

Grid Search

(sklearn.model_selection.GridSearchCV)

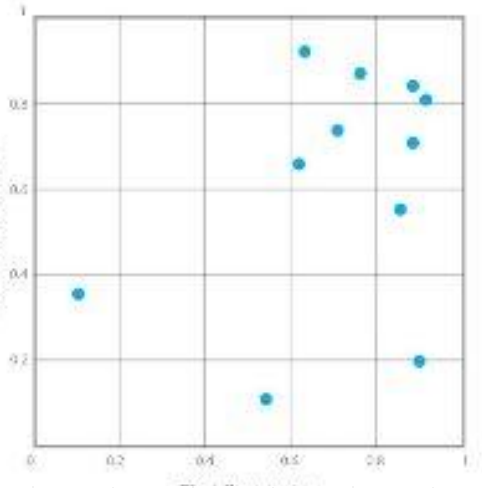


Searches the space linearly from start to end.

Should not be used when training takes considerable time.

Random Search

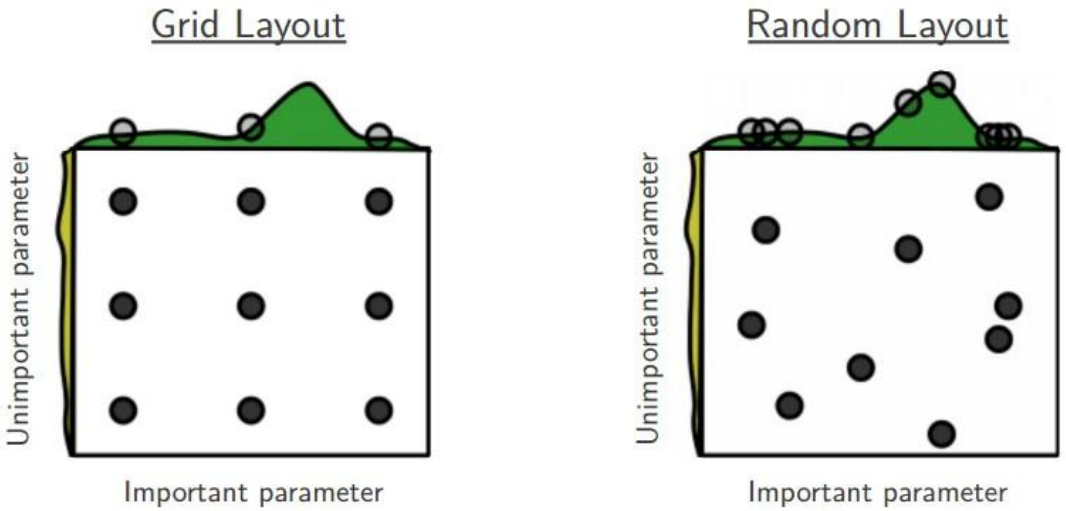
(sklearn.model_selection.RandomizedSearchCV)

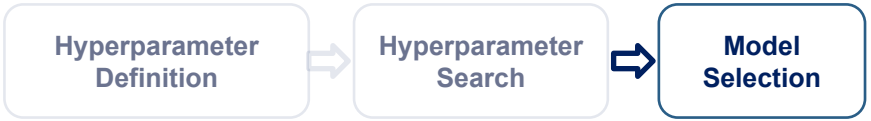


Searches the space randomly.

Covers a wide range of parameters with a low number of samples

Grid vs Random Search




Hyperparameter Definition
Hyperparameter Search
Model Selection

Model Selection

Now that we have several models assessed…

We just need to choose the best model!

… Right?!

Model Selection

There are other aspects besides the score that we must consider when selecting models:

- Training Time;
- Prediction Time;
- Interpretability.

3. Recap

Recap

- Hyperparameters define the structure of our estimators
  - Different from parameters, which are computed by the model
- Hyperparameter search to select best hyperparameters:
  - Grid search
  - Random search
- Model selection

Happy HYPER parameter tuning

4. Q&A