Imputing Missing Data

TLDR: It’s Hard and mostly “It Depends”

Kenneth Farr - 2nd Watch Engineering Manager
ken@farr.ai
github.com/kfarr3/presentation-data-imputation

IMPUTATION

- Finance

  assign (a value) to something by inference from the value of the products or processes to which it contributes.

- Theology

  ascribe (righteousness, guilt, etc.) to someone by virtue of a similar quality in another.

Kinds of Missing Data

Kinds of Missing Data

Kinds of Missing Data

- **MAR** - Missing At Random

Kinds of Missing Data

- **MAR** - Missing At Random

- **MCAR** - Missing Completely At Random

Kinds of Missing Data

- **MAR** - Missing At Random
- **MCAR** - Missing Completely At Random
- **MNAR** - Missing Not At Random

MAR - Missing At Random

- Missing data is systematically related to the observed data but not the unobserved data

MAR - Missing At Random

- Missing data is systematically related to the observed data but not the unobserved data

- Whether a value is missing can typically be predicted from observed data.

MAR - Missing At Random

- Missing data is systematically related to the observed data but not the unobserved data

- Whether a value is missing can typically be predicted from observed data.

- Temperature Sensors in the kitchen malfunction whenever the microwave runs, around noon M-F.

MCAR - Missing Completely At Random

- Missing data is systematically unrelated to the observed and unobserved data

MCAR - Missing Completely At Random

- Missing data is systematically unrelated to the observed and unobserved data

- Whether a value is missing can not be predicted from observed data.

MCAR - Missing Completely At Random

- Missing data is systematically unrelated to the observed and unobserved data

- Whether a value is missing can not be predicted from observed data.

- Temperature Sensors transmit readings over an unreliable UDP network, some packets are dropped and result in missing data.

MNAR - Missing Not At Random

- Missing data is systematically related to the unobserved data

MNAR - Missing Not At Random

- Missing data is systematically related to the unobserved data

- Whether a value is missing can not be predicted from observed data.

MNAR - Missing Not At Random

- Missing data is systematically related to the unobserved data

- Whether a value is missing can not be predicted from observed data.

- Temperature Sensors fail to transmit readings when the value is < 20*C. Missing data is systematically related to the temperature.

What To Do?

Data Scientists Spend

80%

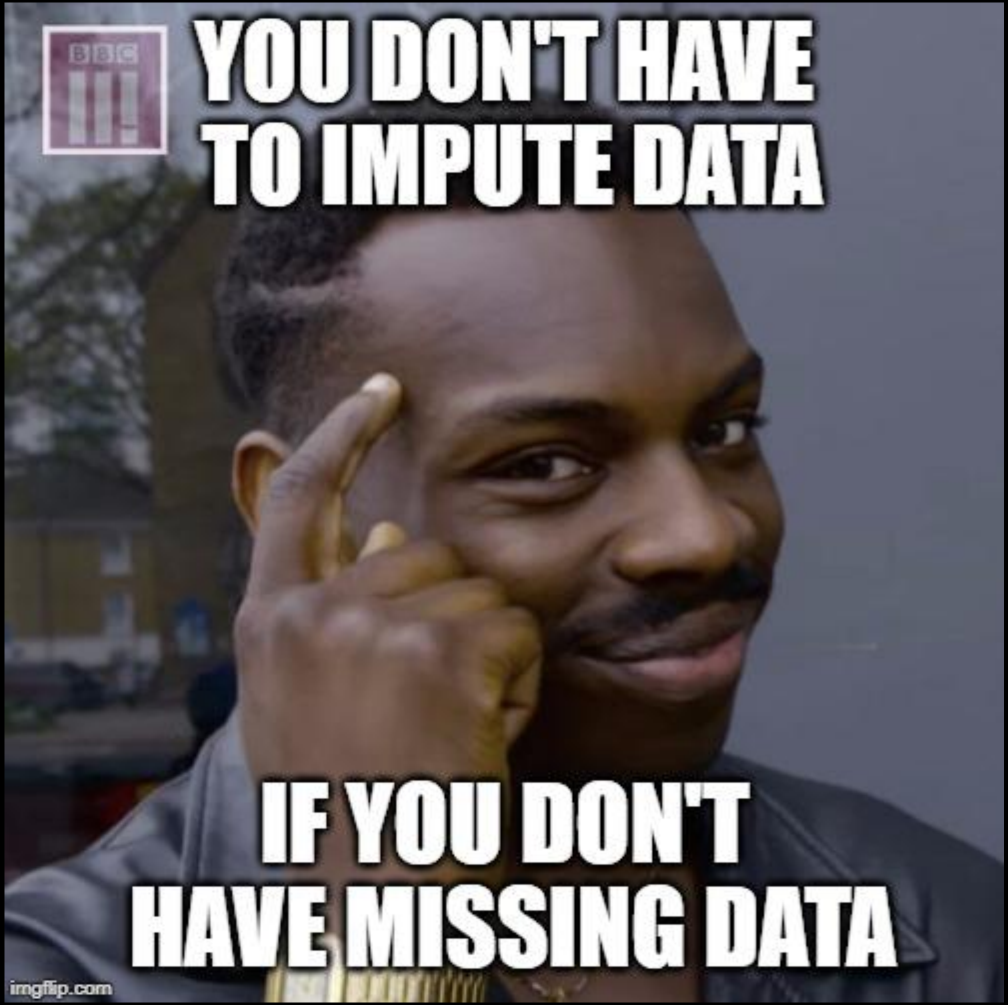
of time spent cleaning data

That's How Intuition Is Gained

Imputation of Missing Data

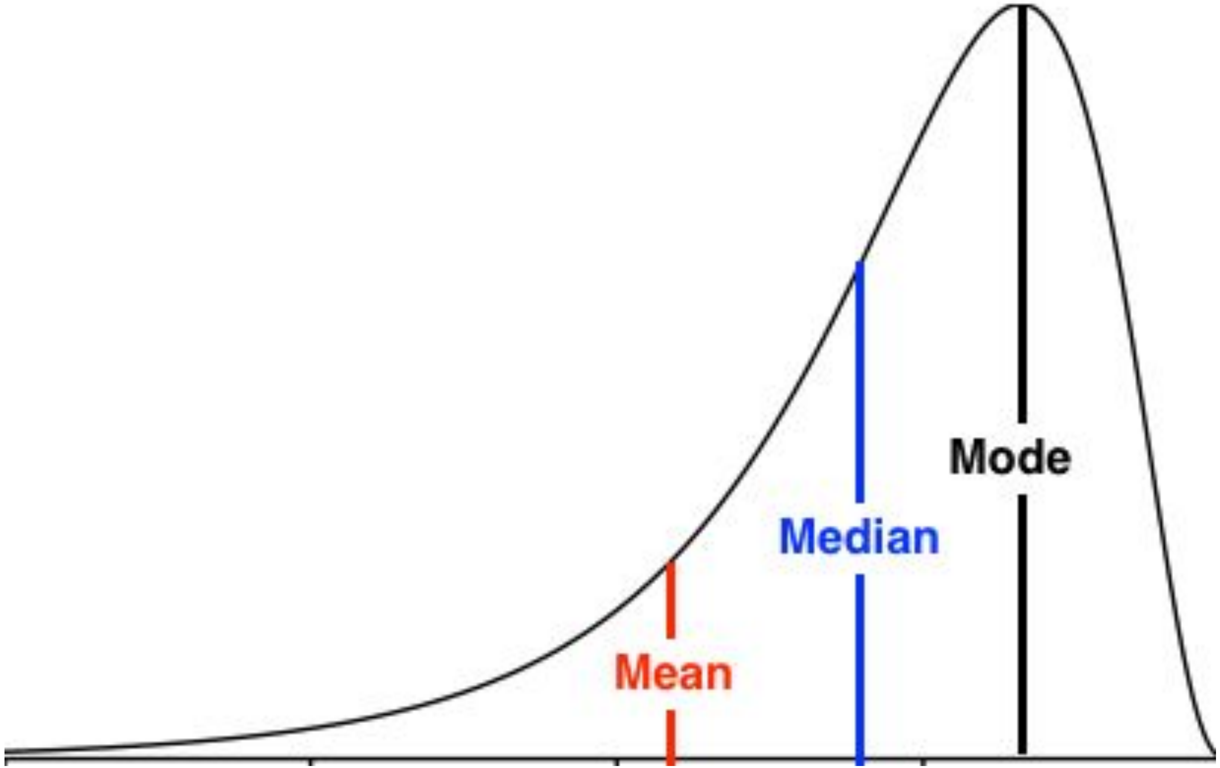
- **MAR** - Missing At Random
  - **Good Candidate**
- **MCAR** - Missing Completely At Random
  - **Best Candidate**
- **MNAR** - Missing Not At Random
  - **Best To Understand Data Better**

How to Impute Simply

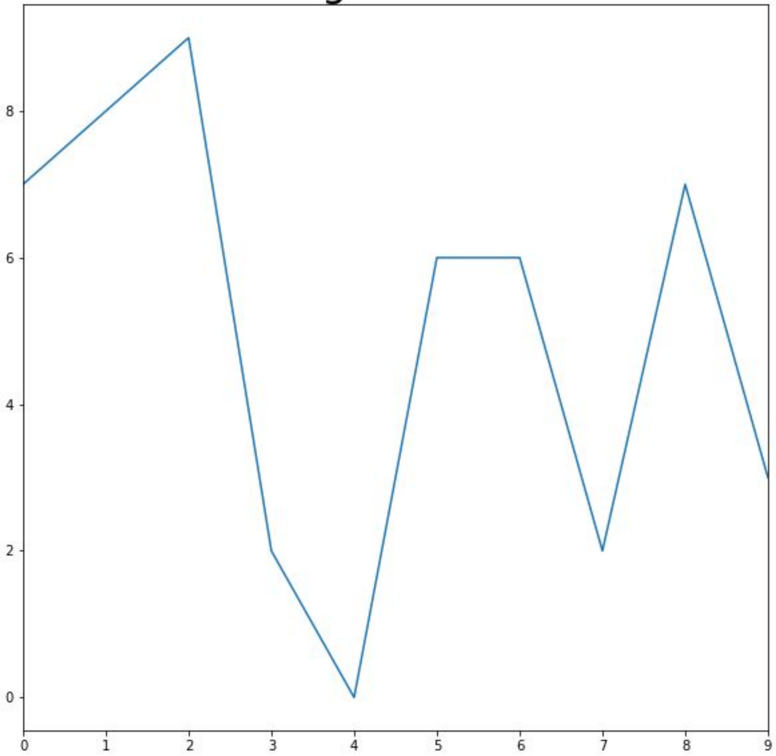

YOU DON'T HAVE
TO IMPUTE DATA
IF YOU DON'T
HAVE MISSING DATA
imgflip.com

Simple! Fill Average

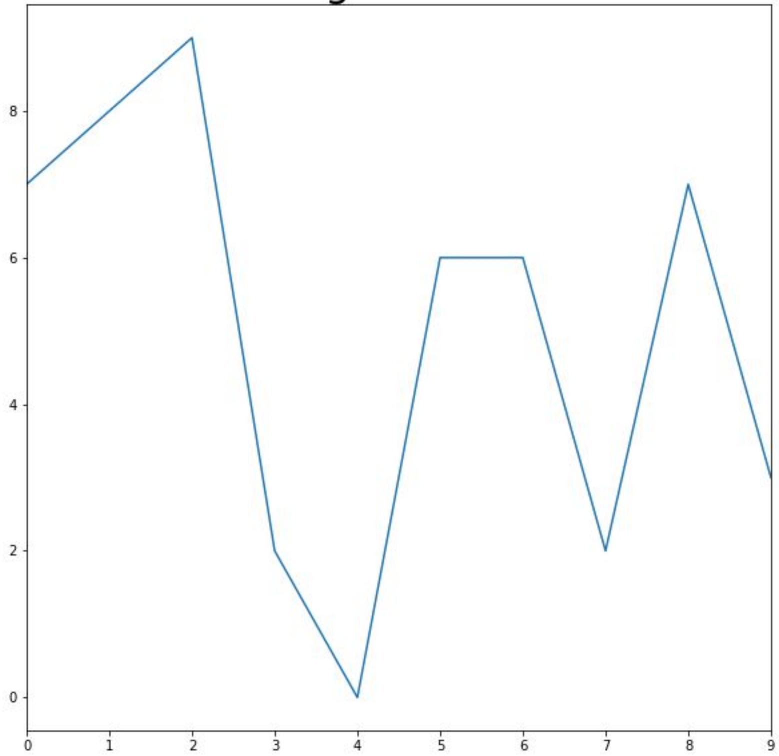
Simple! Fill Average - um...which average?



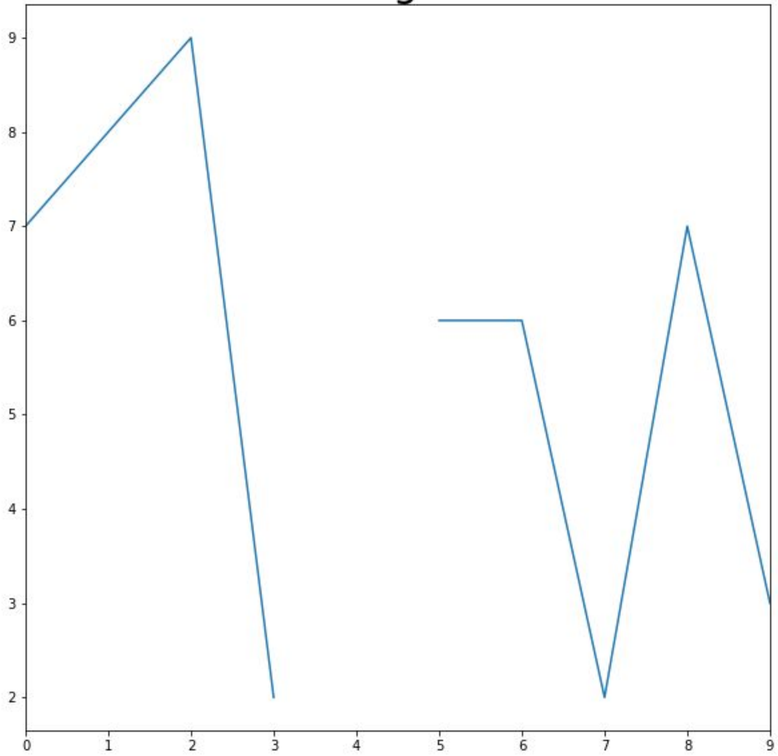
Original Data



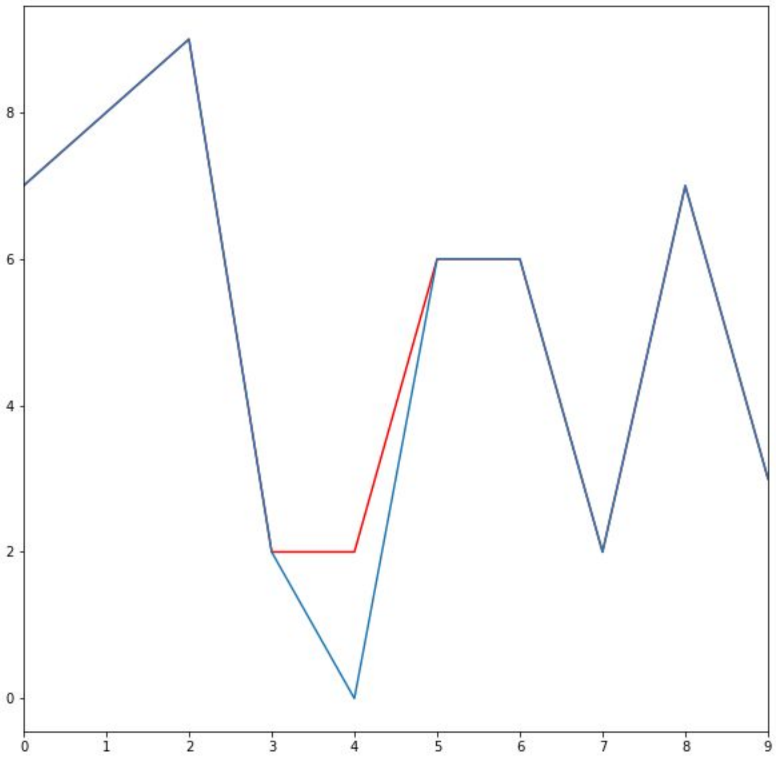
Original Data



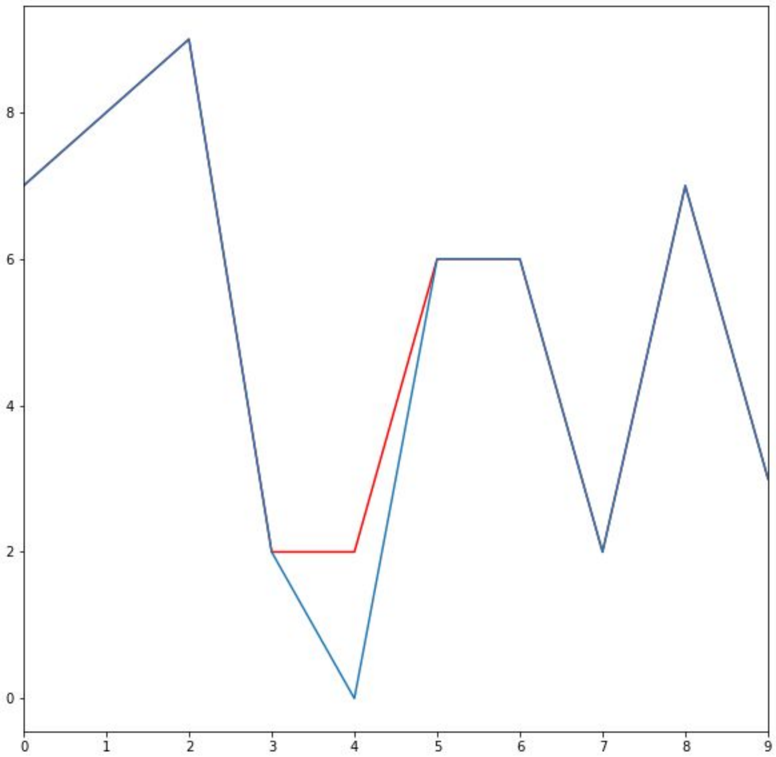
Missing Data



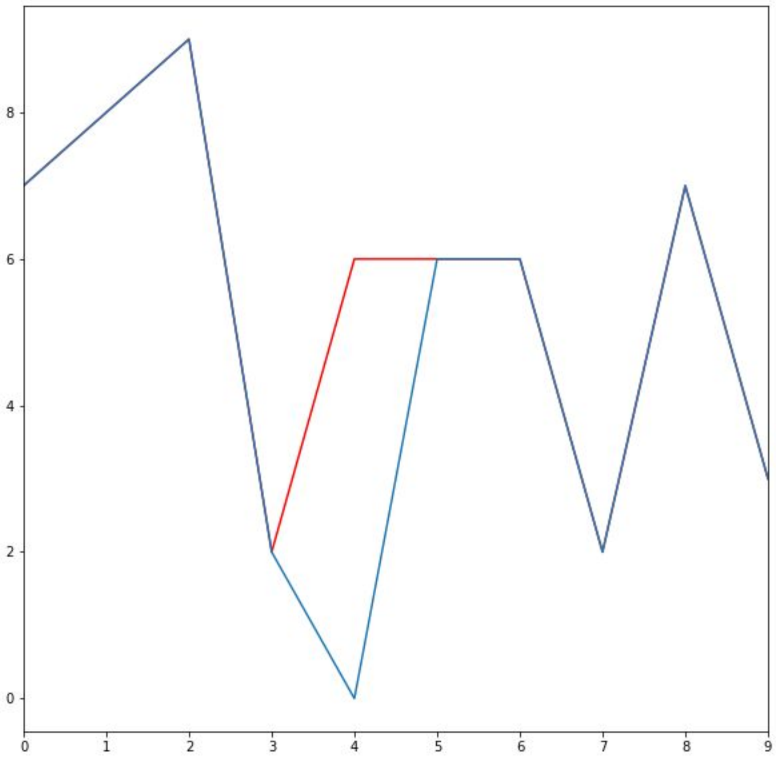
Forward Fill



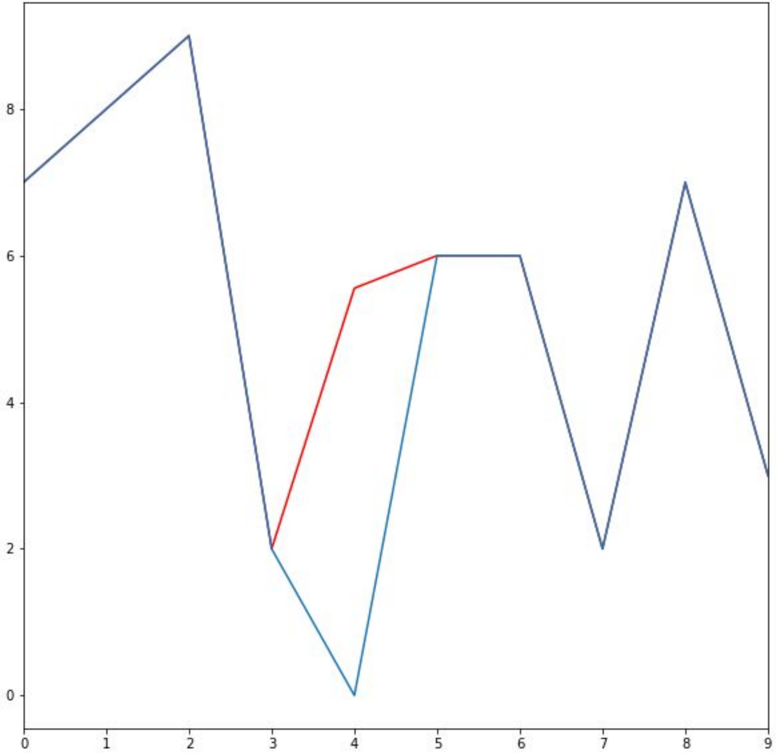
Forward Fill

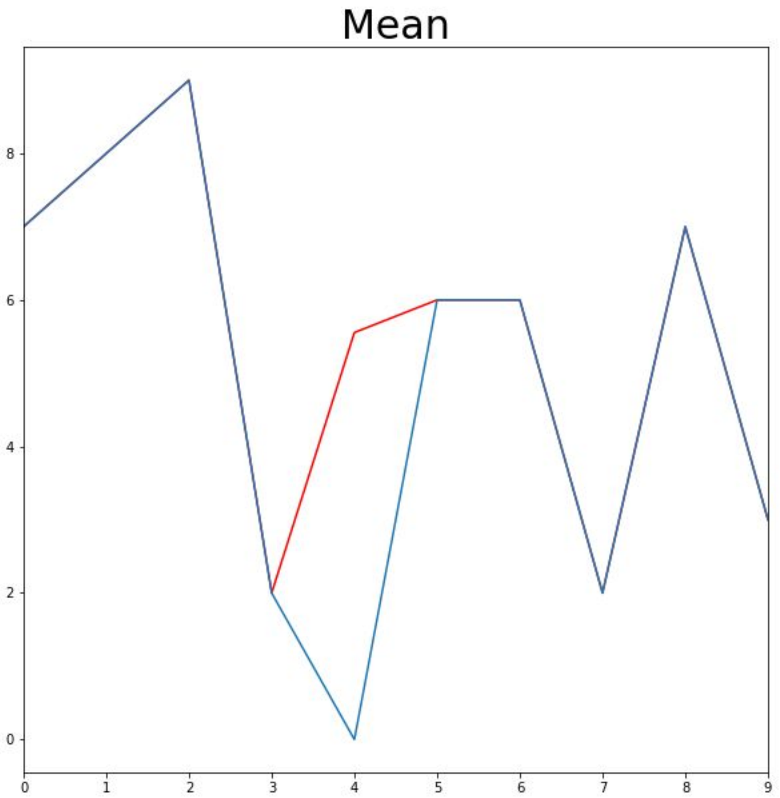


Back Fill

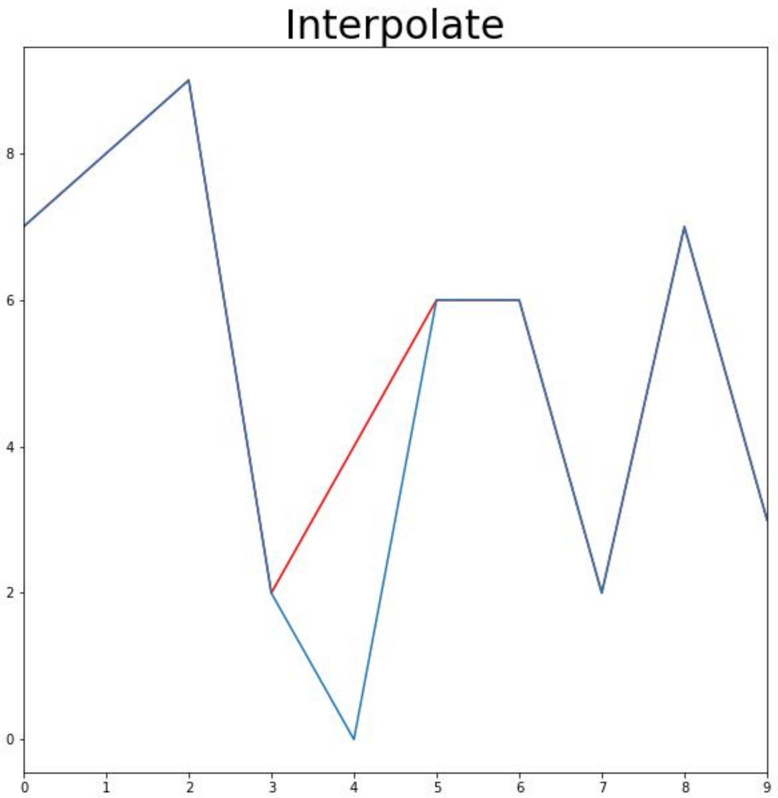


Mean





Mean
0
2
4
6
8
0 1 2 3 4 5 6 7 8 9



Interpolate
0
2
4
6
8
0 1 2 3 4 5 6 7 8 9

Titanic Data

Titanic Data Set

	pclass	sex	age	sib_sp	parch	fare	cabin	embarked	survived	cabin_floor
0	3	male	22.0	1	0	7.2500	NaN	S	0	NaN
1	1	female	38.0	1	0	71.2833	C85	C	1	C
2	3	female	26.0	0	0	7.9250	NaN	S	1	NaN
3	1	female	35.0	1	0	53.1000	C123	S	1	C
4	3	male	35.0	0	0	8.0500	NaN	S	0	NaN

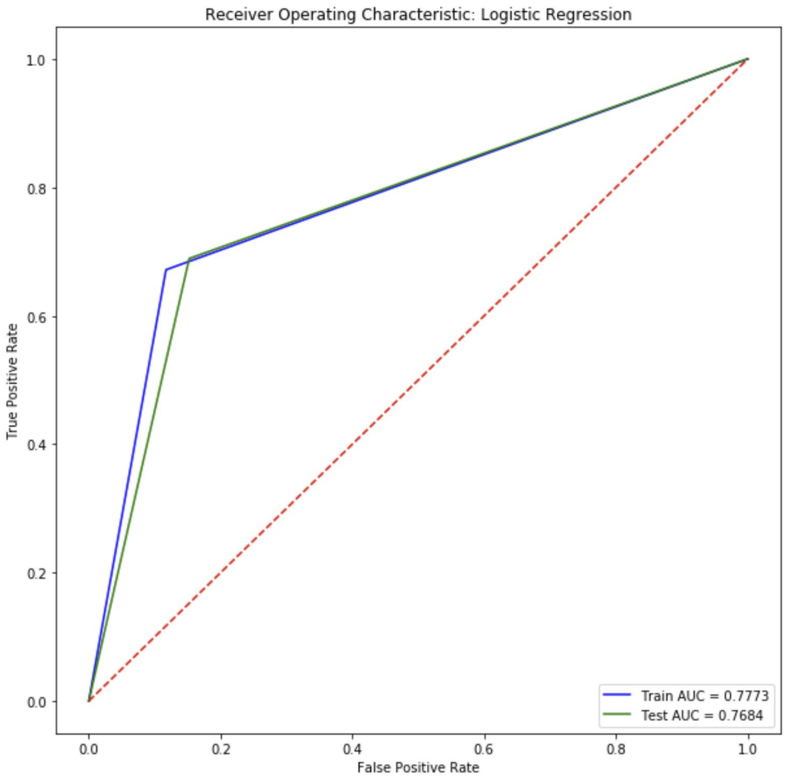
What's Missing?

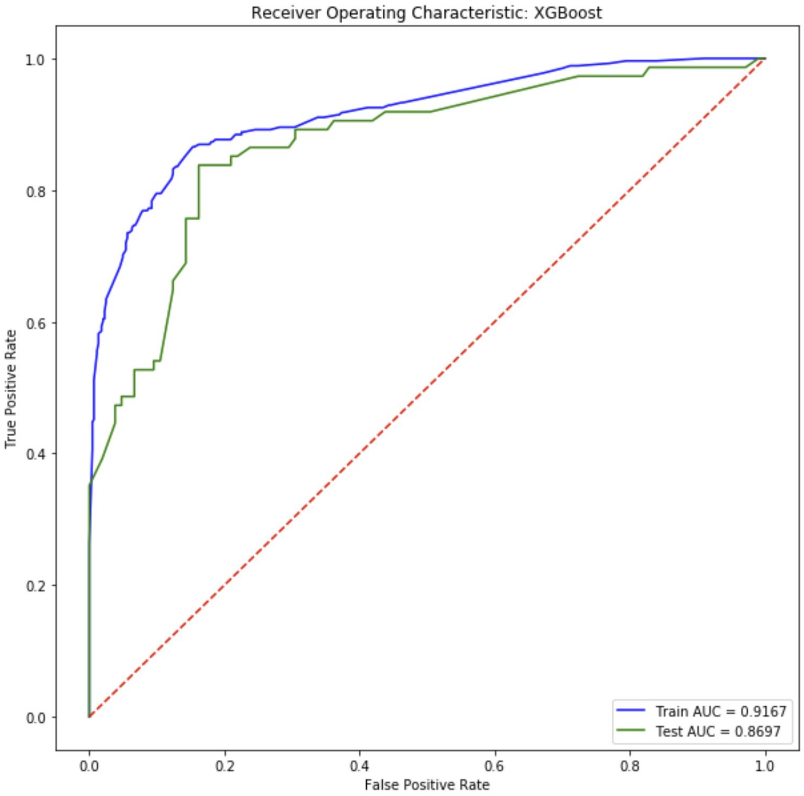
[4]:

```
for col in df.columns:
    missing_count = df[df[col].isna()].shape[0]
    total_count = df.shape[0]
    print(f"{col:20} {missing_count:8} {(100*missing_count/total_count):0.2f}%")
```

```
pclass                      0 0.00%
sex                         0 0.00%
age                       177 19.87%
sib_sp                      0 0.00%
parch                       0 0.00%
fare                        0 0.00%
cabin                     687 77.10%
embarked                    2 0.22%
survived                    0 0.00%
cabin_floor               687 77.10%
```

Baseline: Logistic Regression v XGBoost, Drop NA





AUC
LR: 0.77
XG: 0.87

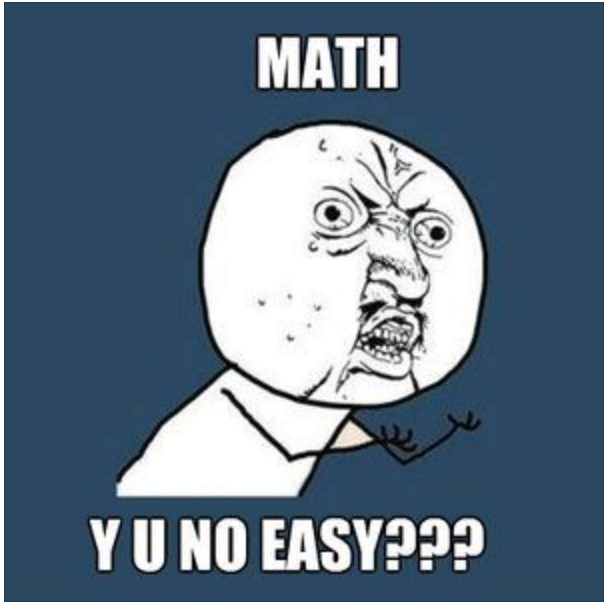
XG Acc
0.80

Add Age where NA=Mean

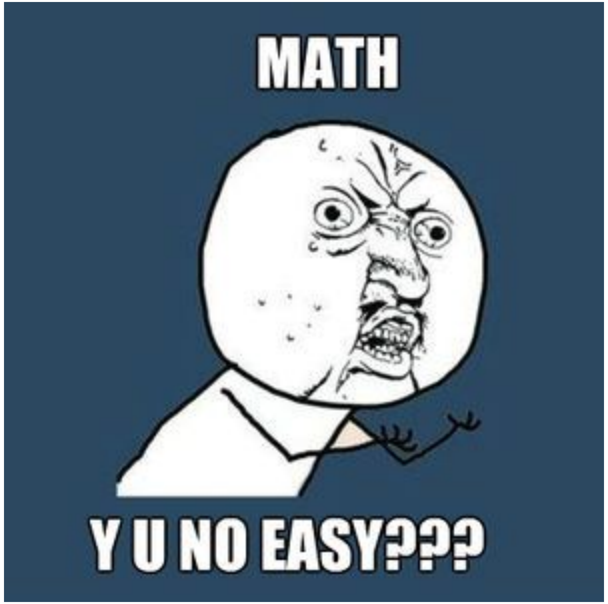
	Drop NA	Age.NA = Mean	Increase
LR AUC	0.7684	0.7894	0.021
XG AUC	0.8697	0.8667	-0.003
XG Accuracy	0.7989	0.7877	-0.0112


I DONT UNDERSTAND

Statistics is Hard

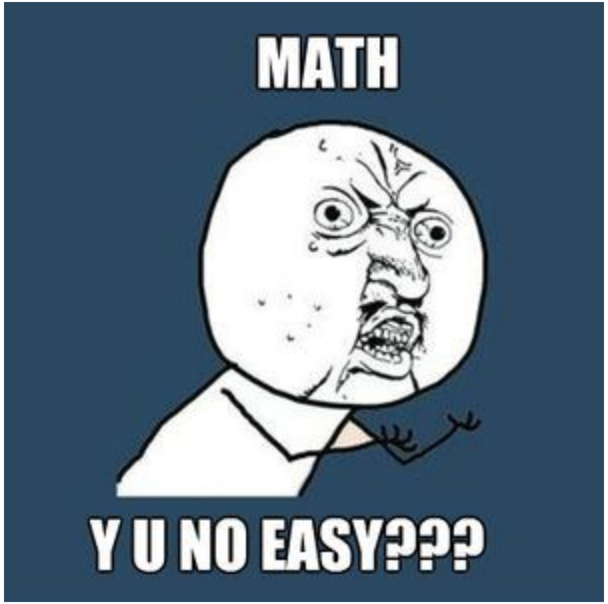


Statistics is Hard



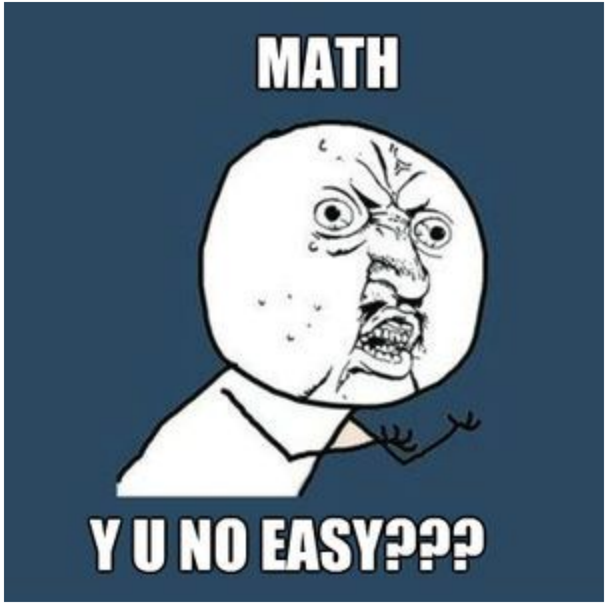
- Age is not a great predictor?

Statistics is Hard



- Age is not a great predictor?
- Mean Age is a worse predictor

Statistics is Hard



- Age is not a great predictor?
- Mean Age is a worse predictor
- Other types may produce better results

Cabin Floor to Mode

	Age.NA = Mean	Cabin.NA = Mode	Increase
LR AUC	0.7894	0.8050	0.0156
XG AUC	0.8667	0.8625	-0.0042
XG Accuracy	0.7877	0.8156	0.0279

Predict Missing Values

XG Boost Predictors for each Missing Value

	Error
Age_Model	11.37 (RMSE)
Cabin_Model	0.63 (Accuracy)

Minor Improvements w/o Hyper Parameter Tuning

	Means	Predictors	Increase
LR AUC	0.8050	0.8069	0.0019
XG AUC	0.8625	0.8879	0.0254
XG Accuracy	0.8156	0.8212	0.0056

XG Accuracy dropping NA: 0.7989
XG Accuracy predicting: 0.8212
Increase in Accuracy of 2.23%

What’s Next?

- Hyperparameter Tuning, General Model Improvements

What’s Next?

- Hyperparameter Tuning, General Model Improvements

- Additional papers: http://jmlr.org/papers/volume18/17-073/17-073.pdf

What’s Next?

- Hyperparameter Tuning, General Model Improvements

- Additional papers: http://jmlr.org/papers/volume18/17-073/17-073.pdf

- Learning Never Stops

FIN

ken@farr.ai
github.com/kfarr3