

Baseline Modeling and Cross-Validation Approaches



Date: 09-11-2019 | Venue: Internshala, Gurugram | Speaker: Mayank Kumar Jha

INTERNSHALA

kaggle™



Devfolio

Who am I?

kaggle
DAYS

MEETUP

- **Mayank Kumar Jha**
 - Data Scientist at Srijan Technologies
 - Kaggle Competitions Expert
 - Won few competitions in the past:
 - AIM Identify the author Challenge by Machine Hack (rank 2)
 - ZS Young data scientist challenge 2018 by Hackerearth (rank 3)
 - World data science challenge by Bitgrit (rank 4)
 - Also a competitive programmer:
 - Won 2 silver and 3 bronze medals at Hackerrank
 - B.Tech Gold Medalist in Academics
- Hobbies : Competitive ML, Algorithms design, Anime series lover
- Can be reached at
 - LinkedIn (<https://www.linkedin.com/in/mk9440/>)
 - Kaggle (<https://www.kaggle.com/mk9440>)
 - Other profiles can be accessed at **mk9440** as well

INTERNSHALA

kaggle

LOGICA

Devfolio

What is Baseline Modeling?

kaggle
DAYS

MEETUP

INTERNSHALA

kaggle™

LOGICA

Devfolio

What is Baseline Modeling?

- Any starting point/solution for a given problem.
- “Baseline” means basic model/solution.

kaggle
DAYS

MEETUP

INTERNSHALA

kaggle™

LOGICA

Devfolio

What is Baseline Modeling?

kaggle
DAYS

MEETUP

- Any starting point/solution for a given problem.
- “Baseline” means basic model/solution.
- Can be any basic or even high-level solution depends upon problem solver.

INTERNSHALA

kaggle™

LOGICA

Devfolio

What is Baseline Modeling?

Examples:

- Binary classification task (Cat Vs Dog):



What is Baseline Modeling?

Examples:

- Binary classification task (Cat Vs Dog): Random predictions



What is Baseline Modeling?

kaggle
DAYS

MEETUP

Examples:

- Recommendation system:

INTERNSHALA

kaggle™

LOGICA

Devfolio

What is Baseline Modeling?

Examples:

- Recommendation system:

1-16 of over 20,000 results for "microwave ovens"

Amazon Prime

☐ [prime](#)

Amazon Fresh

☐ Amazon Fresh

Department

Home & Kitchen

Microwave Ovens

Oven Toaster Grills

Jars & Containers

Oven Gloves

[See more](#)

[See All 2 Departments](#)

Avg. Customer Review

★★★★★ & Up

★★★★☆ & Up

★★★☆☆ & Up

★★☆☆☆ & Up

Brand

☐ IFB

☐ LG

☐ Samsung

☐ Whirlpool

☐ Bajaj

Best seller



IFB 17 L Solo Microwave Oven (17PM MEC 1, White)

★★★★★ < 914

₹4,528 ₹6,090 Save ₹1,562 (26%)

[prime](#) Get it by Friday, November 8



Samsung 23 L Solo Microwave Oven (MS23K3513AK/T, Black)

★★★★★ < 603

₹6,200 ₹7,090 Save ₹890 (13%)

[prime](#) Get it by Friday, November 8

Hint

What is Baseline Modeling?

Examples:

- Recommendation system: Popularity based Algorithm.

1-16 of over 20,000 results for "microwave ovens"

Amazon Prime

☐ [prime](#)

Amazon Fresh

☐ Amazon Fresh

Department

Home & Kitchen
Microwave Ovens
Oven Toaster Grills
Jars & Containers
Oven Gloves

[See more](#)

[See All 2 Departments](#)

Avg. Customer Review

★★★★☆ & Up

★★★★☆ & Up

★★★★☆ & Up

★★★★☆ & Up

Brand

☐ IFB

☐ LG

☐ Samsung

☐ Whirlpool

☐ Bajaj

Best seller



IFB 17 L Solo Microwave Oven (17PM MEC 1, White)

★★★★☆ ~ 914

₹4,528 ₹6,090 Save ₹1,562 (26%)

[prime](#) Get it by Friday, November 8



Samsung 23 L Solo Microwave Oven (MS23K3513AK/T, Black)

★★★★☆ ~ 603

₹6,200 ₹7,090 Save ₹890 (13%)

[prime](#) Get it by Friday, November 8

Hint

What is Cross-Validation?

kaggle
DAYS

MEETUP

INTERNSHALA

kaggle™

LOGICA

Devfolio

What is Cross-Validation?

kaggle
DAYS

MEETUP

- Procedure to evaluate how good or bad is our model

INTERNSHALA

kaggle™

LOGICA

Devfolio

What is Cross-Validation?

kaggle
DAYS

MEETUP

- Procedure to evaluate how good or bad is our model
- Most useful variants are:
 - K-Fold cross validation
 - Stratified K-Fold Cross Validation
 - Time Series Cross Validation

INTERNSHALA

kaggle

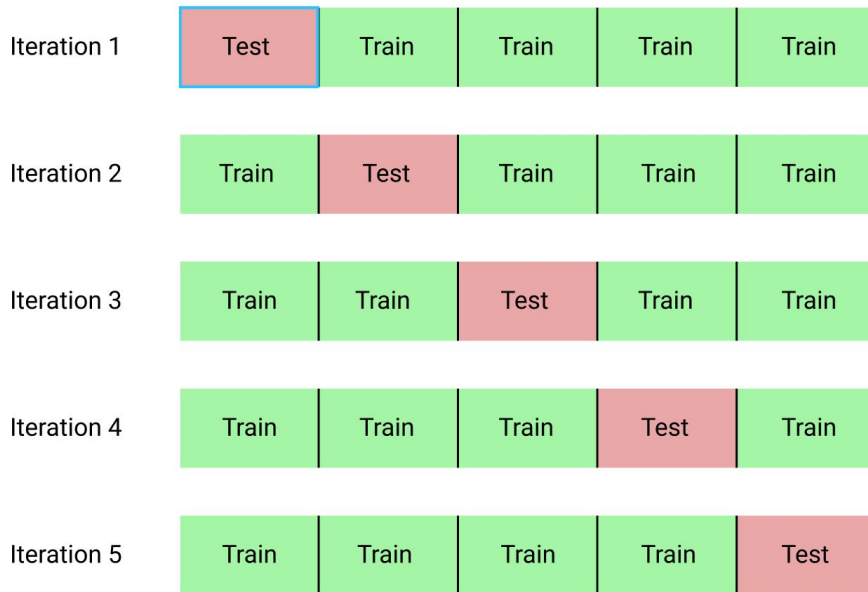
LOGICA

Devfolio

What is Cross-Validation?

- **K-Fold Cross validation:**

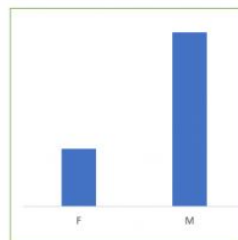
- Split the data into k folds
- Choose any (k-1) folds as training set.
- Choose remaining k^{th} fold as test set.
- Repeat above selection k times covering each fold as test set once.



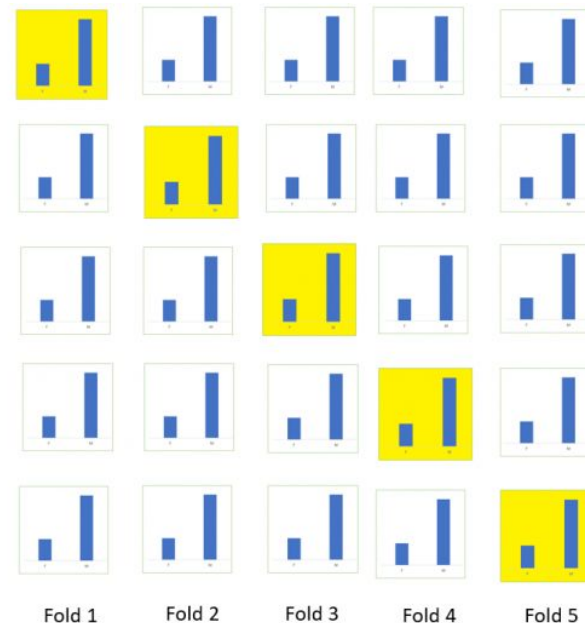
What is Cross-Validation?

- **Stratified K-Fold Cross validation:**
 - Split the data into k folds such that each fold has same class distribution as of whole data.
 - Choose any (k-1) folds as training set.
 - Choose remaining k^{th} fold as test set.
 - Repeat above selection k times covering each fold as test set once.

Stratified K-Fold
Cross Validation
(K=5)



Class Distributions

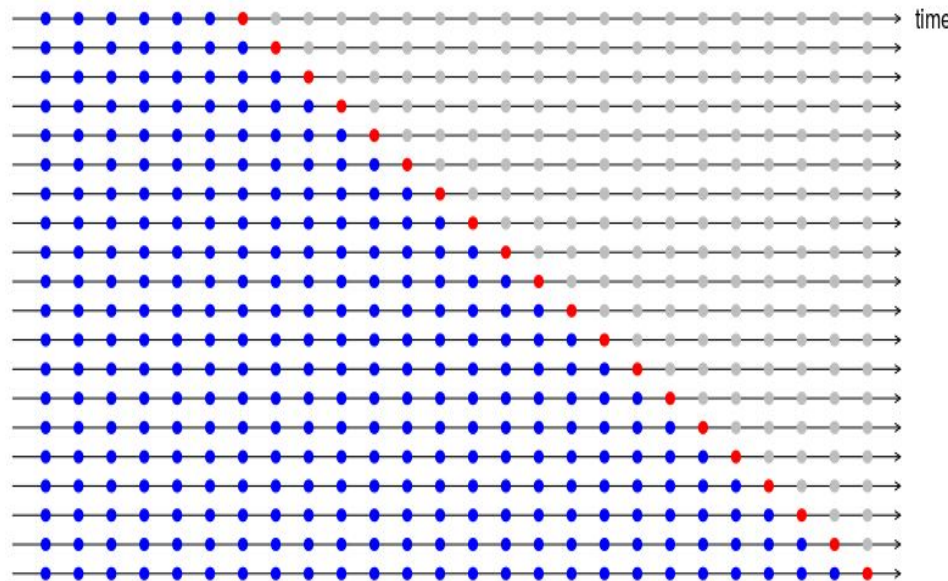


What is Cross-Validation?

- **Time Series Cross validation:**

- Also called forward chaining or rolling.
- Sort the data over timestamp feature in ascending order.
- Split the data into k folds.
- Choose 1st fold as training set.
- Choose remaining $(k-1)^{\text{th}}$ folds as test set.
- Repeat above selection k times by sequentially appending each Fold from test data into train data and removing initial fold from test data.

Image reference: <https://robjhyndman.com/hyndsight/tscv/>



Let's be practical.....

kaggle
DAYS

MEETUP

Fire up your Notebooks



INTERNSHALA

kaggle™

LOGICA

Devfolio



References/ Credits

kaggle
DAYS

MEETUP

- <http://viralslacker.com/cat-vs-dog-memes-funny/>
- <https://www.amazon.in/>
- <https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85>
- <https://robjhyndman.com/hyndsight/tscv/>
- https://en.wikipedia.org/wiki/Project_Jupyter
- <https://www.memesmonkey.com/topic/question>
- <https://image.slidesharecdn.com/>

INTERNSHALA

kaggle™

LOGICA

Devfolio

THANK YOU

