Baseline Modeling and Cross-Validation Approaches

kaggle DAYS MEETUP

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

INTERNSHALA kaggle LOGICNI Devfolio

Who am I?

- **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

What is Baseline Modeling?

kaggle DAYS MEETUP

INTERNSHALA kaggle LOGICNI Devfolio

What is Baseline Modeling?

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

kaggle DAYS MEETUP

INTERNSHALA kaggle LOGICNI Devfolio

What is Baseline Modeling?

- 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.

What is Baseline Modeling?

kaggle DAYS MEETUP

Examples:

- Binary classification task (Cat Vs Dog):

What is Baseline Modeling?

kaggle DAYS MEETUP

Examples:

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

What is Baseline Modeling?

Examples:

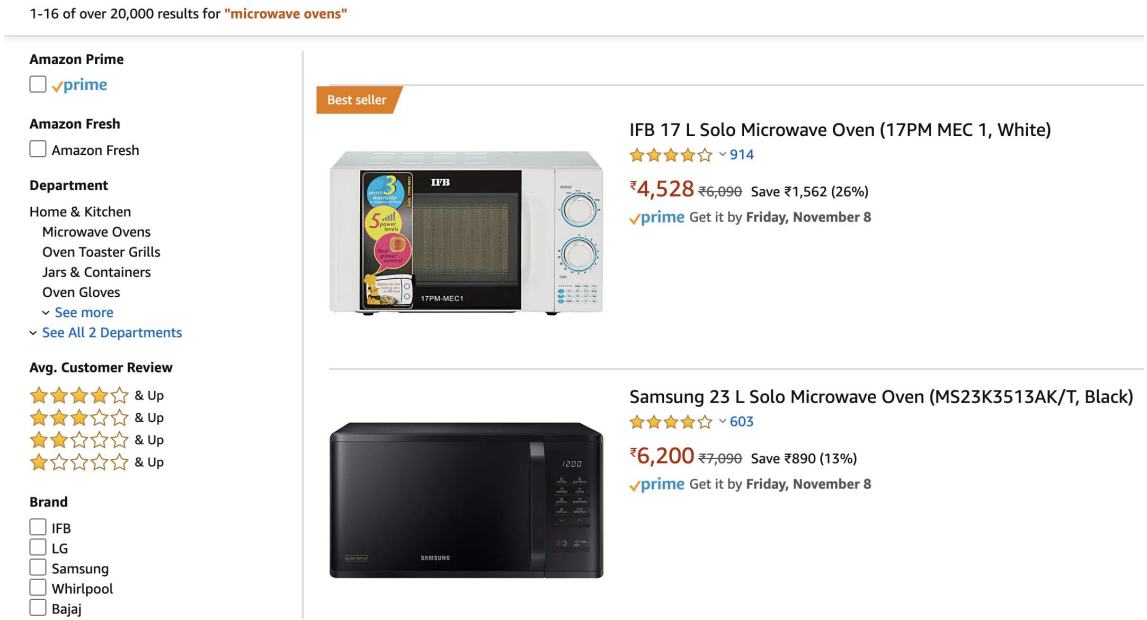
- Recommendation system:

What is Baseline Modeling?

kaggle DAYS MEETUP

Examples:

- Recommendation system:



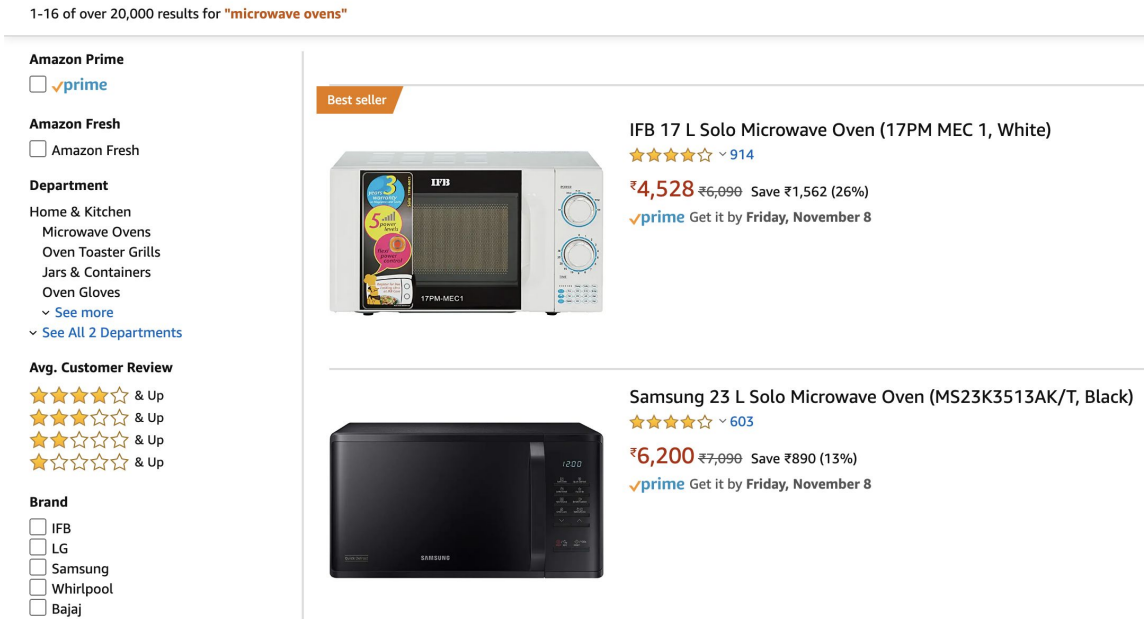
Hint

What is Baseline Modeling?

kaggle DAYS MEETUP

Examples:

- Recommendation system: Popularity based Algorithm.



Hint

What is Cross-Validation?

kaggle DAYS MEETUP

INTERNSHALA kaggle LOGICNI Devfolio

What is Cross-Validation?

- Procedure to evaluate how good or bad is our model

What is Cross-Validation?

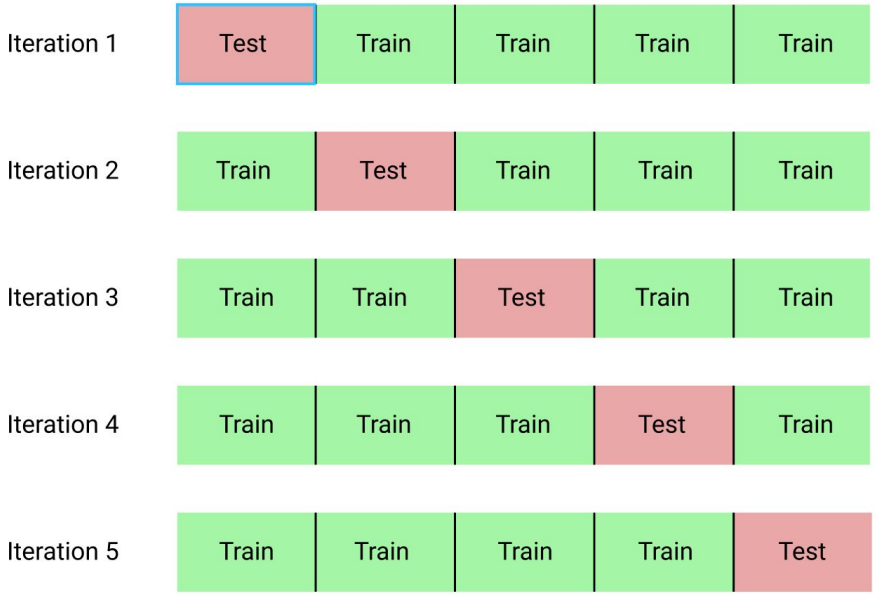
kaggle DAYS MEETUP

- 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

What is Cross-Validation?

kaggle DAYS MEETUP

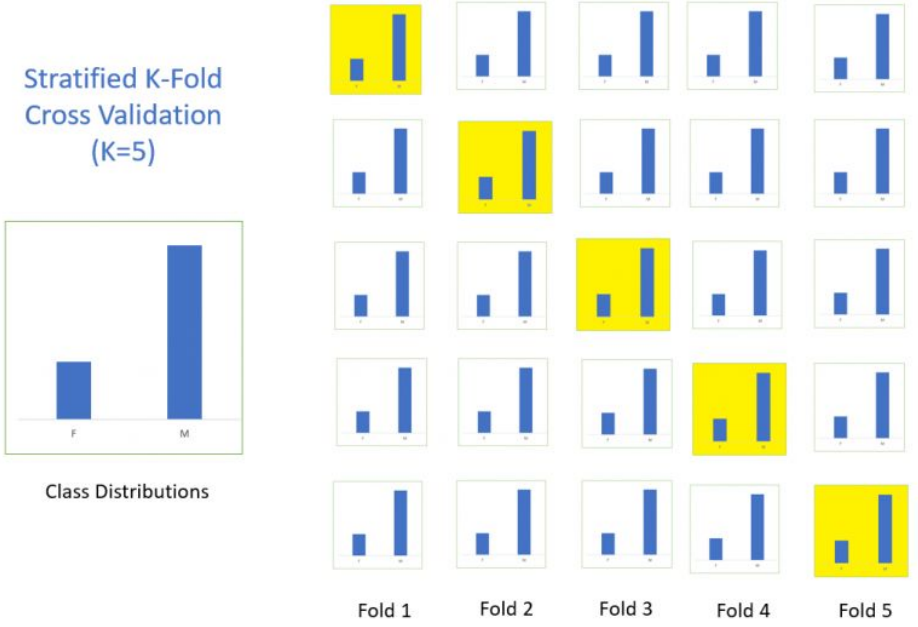
- **K-Fold Cross validation:**
  - Split the data into k folds
  - Choose any (k-1) folds as training set.
  - Choose remaining k^{th} fold as test set.
  - Repeat above selection k times covering each fold as test set once.



What is Cross-Validation?

kaggle DAYS MEETUP

- **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.

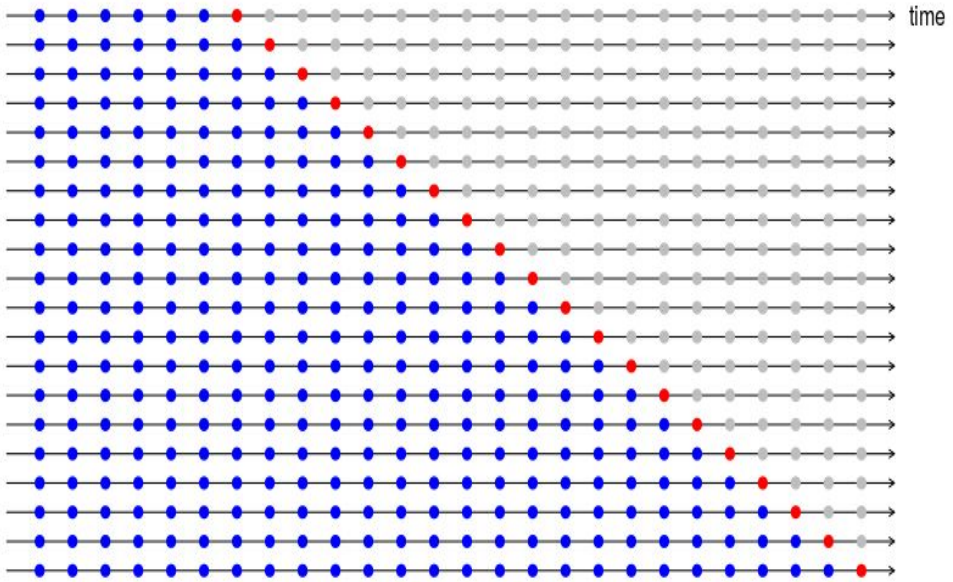


What is Cross-Validation?

kaggle DAYS MEETUP

- **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)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/



INTERNSHALA kaggle LOGICNI Devfolio

kaggle DAYS MEETUP

Fire up your Notebooks

Questions & Answers

kaggle DAYS MEETUP

INTERNSHALA kaggle LOGICNI Devfolio

References/ Credits

- 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/

kaggle DAYS MEETUP

INTERNSHALA kaggle LOGICNI Devfolio

THANK YOU