Evaluation of Recommender Systems

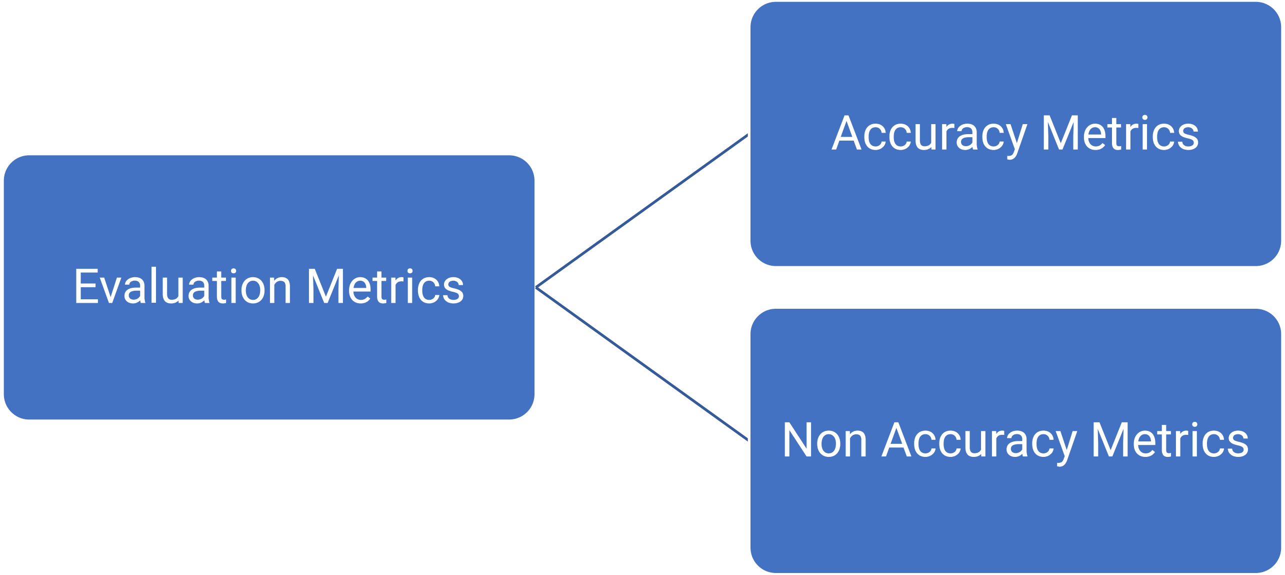
Analytics Vidhya

Why is evaluation difficult?

- Variability in datasets for different domains
  - Movie Recommender (n_users >> n_items) vs Resarch Paper Recommender (n_items >> n_users)

- Goals for evaluation may differ
  - Traditionally accuracy has been considered important
  - User Satisfaction is important – Difficult to measure
  - Coverage
  - Novelty
  - No of Purchases

Types of Evaluation Metrics



Evaluation Metrics

Accuracy Metrics

Non Accuracy Metrics

Accuracy Metrics: Types of Output

- Output types
  - **Prediction:** A (numerical) prediction indicating to what degree the current user will like or dislike a certain item
  - **Recommendation:** A top-N list of recommended items

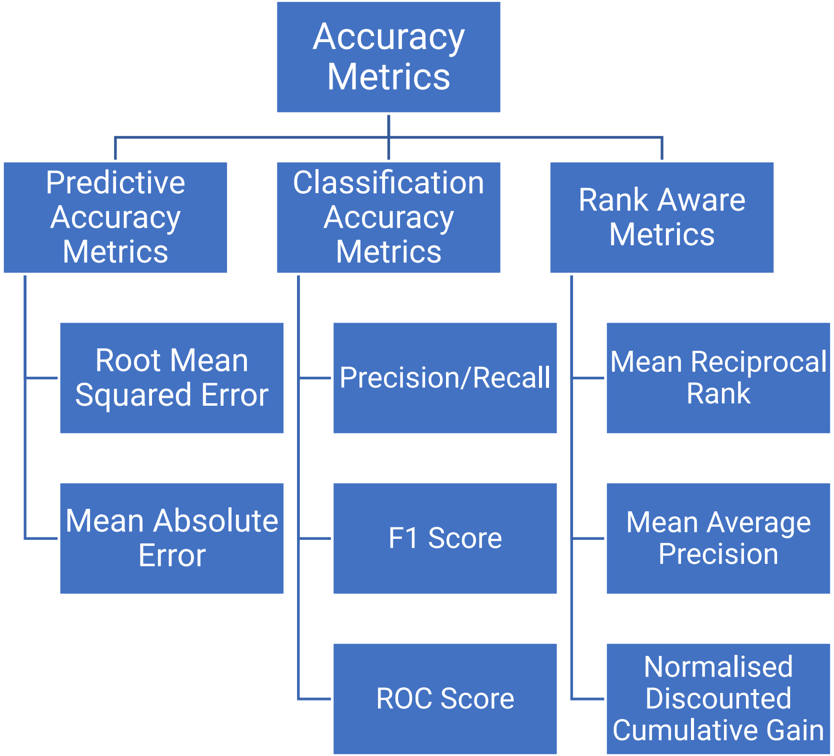
Prediction

User	Movie	Predicted Rating
Tom	Argo	5
Tom	Seven	4
Tom	Righteous Kill	3

Top n Recommendations

$$Tom\colon \{Argo, Seven, Righteous\ Kill\}$$

Accuracy Metrics Taxonomy



Accuracy Metrics

Predictive Accuracy Metrics

- Root Mean Squared Error
- Mean Absolute Error

Classification Accuracy Metrics

- Precision/Recall
- F1 Score
- ROC Score

Rank Aware Metrics

- Mean Reciprocal Rank
- Mean Average Precision
- Normalised Discounted Cumulative Gain

Analytics Vidhya

Predictive & Classification Accuracy Metrics

Predictive Accuracy Metrics

- Datasets with items rated by users
  - MovieLens datasets 100K-10M ratings
  - Netflix 100M ratings
- Historic user ratings constitute ground truth
- Metrics measure error rate
  - Mean Absolute Error (*MAE*) computes the deviation between predicted ratings and actual ratings

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | p_i - r_i |$$

  - Root Mean Square Error (*RMSE*) is like *MAE*, but places more emphasis on larger deviation

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (p_i - r_i)^2}$$

Predictive Accuracy Metrics – Pros & Cons

- **Advantages**
    - ○ Mechanics of the computation are simple and easy to understand
    - ○ Mean absolute error has well studied statistical properties that provide for testing the significance of a difference between the mean absolute errors of two systems.
- **Disadvantages**
    - ○ Mean absolute error may be less appropriate for tasks such as Find Good Items where a ranked result is returned to the user, who then only views items at the top of the ranking.
    - ○ Mean absolute error may be less appropriate when the granularity of true preference (a domain feature) is small

Classification Accuracy Metrics: Precision and Recall

- Measure the frequency with which recommender system makes correct or incorrect decisions about whether the item is good
- **Precision:** a measure of exactness, determines the fraction of relevant items retrieved out of all items retrieved
    - E.g. the proportion of recommended movies that are actually good

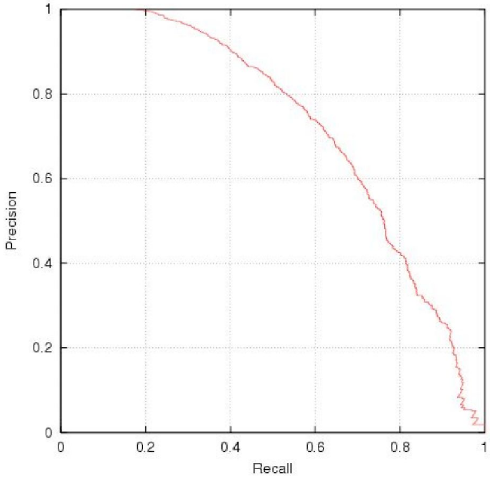
$$Precision = \frac{tp}{tp + fp} = \frac{|good\ movies\ recommended|}{|\text{all recommendations}|}$$

- **Recall:** a measure of completeness, determines the fraction of relevant items retrieved out of all relevant items
    - E.g. the proportion of all good movies recommended

$$Recall = \frac{tp}{tp + fn} = \frac{|good\ movies\ recommended|}{|all\ good\ movies|}$$

Precision & Recall

- E.g. typically when a recommender system is tuned to increase precision, recall decreases as a result (or vice versa)



F1 Metric

- The **F$_1$ Metric** attempts to combine Precision and Recall into a single value for comparison purposes.
  - May be used to gain a more balanced view of performance

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Metrics: Precision@K

Evaluates the list of recommendations

- Precision at k is the proportion of recommended items in top-k set that are relevant

$$P@k(u) = \frac{\#\{\textit{relevant content in the top k postitions}\}}{k}$$

- Suppose that my precision at 10 in a top-10 recommendation problem is 80%, this means that 80% of recommendations I make are relevant to the user

Metrics: Recall@K

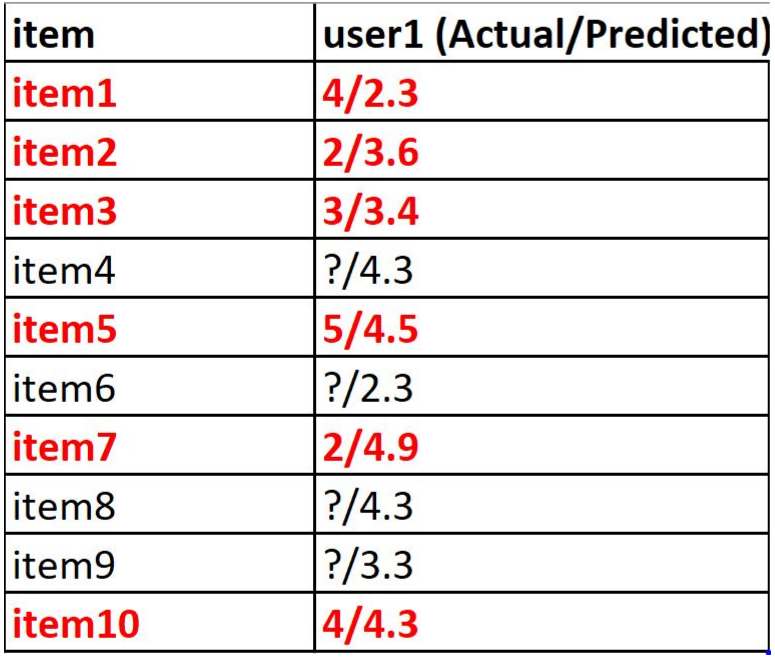
Evaluates the list of recommendations

- Recall at k is the proportion of relevant items found in the top-k recommendations

$$Recall@k = \frac{\#\ of\ recommended\ items\ @k\ that\ are\ relevant}{total\ \#\ of\ relevant\ items}$$

- Suppose that we computed recall at 10 and found it is 40% in our top-10 recommendation system. This means that 40% of the total number of the relevant items appear in the top-k result

Example for Precison@K & Recall@K



item	user1 (Actual/Predicted)
item1	**4/2.3**
item2	**2/3.6**
item3	**3/3.4**
item4	?/4.3
item5	**5/4.5**
item6	?/2.3
item7	**2/4.9**
item8	?/4.3
item9	?/3.3
item10	**4/4.3**

Let's say relevant items are ones with rating greater than 3.5

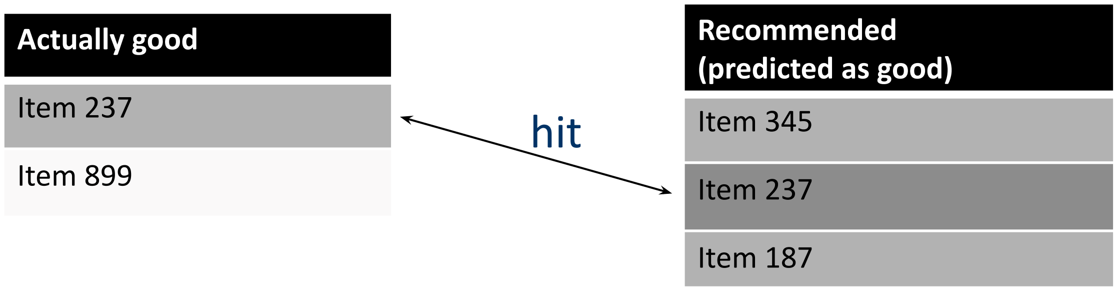
- Relevant items: item5, item10 and item1
  total # of relevant items = 3
- Recommended items @ 3: item7, item5 and item10
  # of recommended items at 3 = 3

- Precision@3 $= \frac{2}{3}$
- Recall@3 $= \frac{2}{3}$

Analytics Vidhya

Rank Aware Metrics: Rank Position Matters

For a user:



Actually good
Item 237
Item 899

hit

Recommended (predicted as good)
Item 345
Item 237
Item 187

- **Rank metrics** extend recall and precision to take the positions of correct items in a ranked list into account
    - Relevant items are more useful when they appear earlier in the recommendation list
    - Particularly important in recommender systems as lower ranked items may be overlooked by users

Mean Reciprocal Rank

Evaluates the list of recommendations

$$\mathrm{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{\mathrm{rank}_i}$$

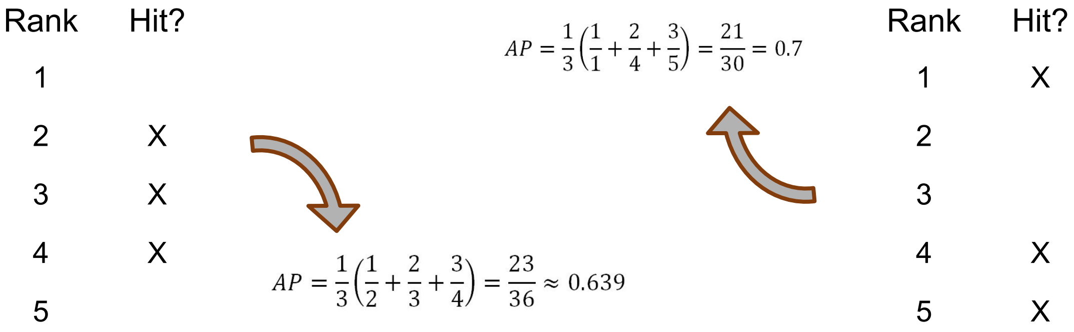
- Suppose we have recommended 3 movies to a user, say A, B, C in the given order, but the user only liked movie C. As the rank of movie C is 3, the reciprocal rank will be 1/3

- For multiple recommendations, the Mean Reciprocal Rank is the mean of all reciprocal ranks.

- Larger the mean reciprocal rank, better the recommendations

Mean Average Precision

- Average Precision (*AP*) is a ranked precision metric that places emphasis on highly ranked correct predictions (hits)
- Essentially it is the average of precision values determined after each successful prediction, i.e.
- If a relevant document never gets retrieved, we assume the precision to be 0



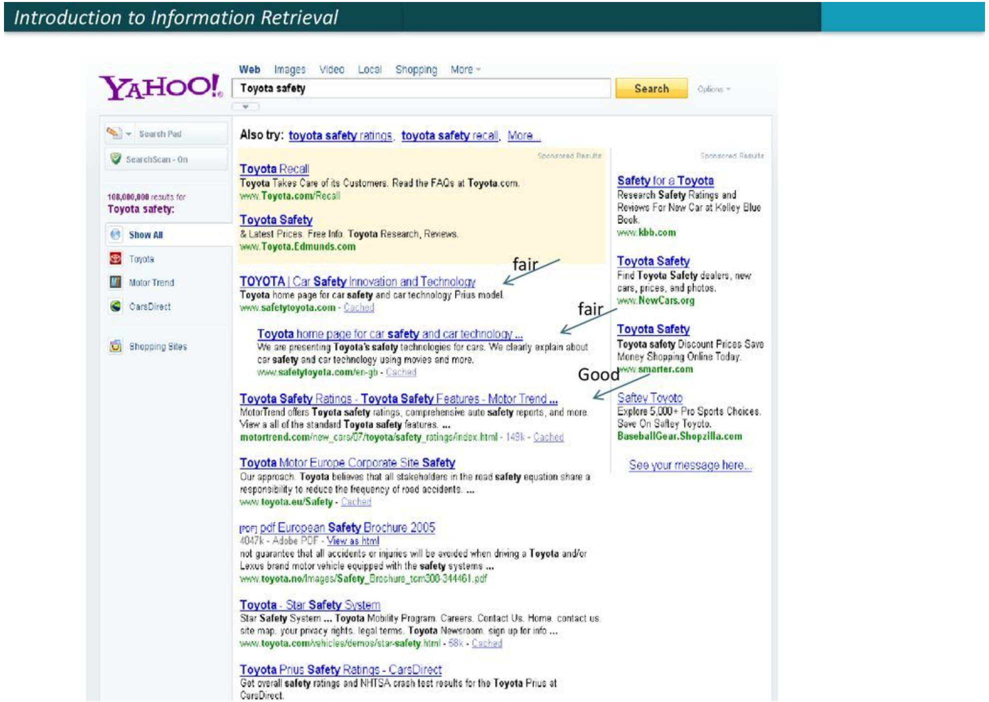
Rank	Hit?
1	
2	X
3	X
4	X
5	

$$AP = \frac{1}{3}\left(\frac{1}{2}+\frac{2}{3}+\frac{3}{4}\right) = \frac{23}{36} \approx 0.639$$

$$AP = \frac{1}{3}\left(\frac{1}{1}+\frac{2}{4}+\frac{3}{5}\right) = \frac{21}{30} = 0.7$$

Rank	Hit?
1	X
2	
3	
4	X
5	X

Beyond Binary Relevance



Introduction to Information Retrieval

Toyota safety

108,000,000 results for **Toyota safety:**

Also try: toyota safety ratings, toyota safety recall, More...

Sponsored Results

Toyota Recall
Toyota Takes Care of its Customers. Read the FAQs at Toyota.com.
www.Toyota.com/Recall

Toyota Safety
& Latest Prices. Free Info. Toyota Research, Reviews.
www.Toyota.Edmunds.com

TOYOTA | Car Safety Innovation and Technology
Toyota home page for car safety and car technology Prius model.
www.safetytoyota.com - Cached

Toyota home page for car safety and car technology ...
We are presenting Toyota's safety technologies for cars. We clearly explain about car safety and car technology using movies and more.
www.safetytoyota.com/en-gb - Cached

Toyota Safety Ratings - Toyota Safety Features - Motor Trend ...
MotorTrend offers Toyota safety ratings, comprehensive auto safety reports, and more. View a all of the standard Toyota safety features. ...
motortrend.com/new_cars/07/toyota/safety_ratings/index.html - 149k - Cached

Toyota Motor Europe Corporate Site Safety
Our approach. Toyota believes that all stakeholders in the road safety equation share a responsibility to reduce the frequency of road accidents. ...
www.toyota.eu/Safety - Cached

[PDF] pdf European Safety Brochure 2005
4047k - Adobe PDF - View as html
not guarantee that all accidents or injuries will be avoided when driving a Toyota and/or Lexus brand motor vehicle equipped with the safety systems ...
www.toyota.no/Images/Safety_Brochure_tcm306-344461.pdf

Toyota - Star Safety System
Star Safety System ... Toyota Mobility Program. Careers. Contact Us. Home. contact us. site map. your privacy rights. legal terms. Toyota Newsroom. sign up for info ...
www.toyota.com/vehicles/demos/star-safety.html - 58k - Cached

Toyota Prius Safety Ratings - CarsDirect
Get overall safety ratings and NHTSA crash test results for the Toyota Prius at CarsDirect.

Sponsored Results

Safety for a Toyota
Research Safety Ratings and Reviews For New Car at Kelley Blue Book.
www.kbb.com

Toyota Safety
Find Toyota Safety dealers, new cars, prices, and photos.
www.NewCars.org

Toyota Safety
Toyota safety Discount Prices Save Money Shopping Online Today.
www.smarter.com

Saftey Toyoto
Explore 5,000+ Pro Sports Choices. Save On Saftey Toyoto.
BaseballGear.Shopzilla.com

See your message here...

Normalised Discounted Cumulative Gain

- Discounted cumulative gain (DCG)
  - Logarithmic reduction factor

$$DCG_{pos} = rel_1 + \sum_{i=2}^{pos} \frac{rel_i}{\log_2 i}$$

Where:

- *pos* denotes the position up to which relevance is accumulated
- rel_i returns the relevance of recommendation at position *i*

- Idealized discounted cumulative gain (IDCG)
  - Assumption that items are ordered by decreasing relevance

$$IDCG_{pos} = rel_1 + \sum_{i=2}^{|h|-1} \frac{rel_i}{\log_2 i}$$

- Normalized discounted cumulative gain (nDCG)
  - Normalized to the interval [0..1]

$$nDCG_{pos} \frac{DCG_{pos}}{IDCG_{pos}}$$

Analytics Vidhya

- **Let’s say there are 10 ranked movies on 0-3 relevance scale:**
  - 3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- **Discounted Gain**
  - 3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0

  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
- **Discounted Cumulative Gain**
  - 3, 5, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61
- **Inverse Discounted Cumulative Gain**
  - 3, 3, 3, 2, 2, 2, 1, 0, 0, 0

  = 3, 3/1, 3/1.59, 2/2, 2/2.32, 2/2.59, 1/2.81, 0, 0, 0

  = 3, 3, 1.89, 1, 0.86, 0.77, 0.36, 0, 0, 0

  Cumulative □ 3, 6, 7.89, 8.75, 9.52, 9.88, 9.88, 9.88

NDCG = DCG/IDCG

1, 0.83, 0.87, 0.832, 0.808, 0.88, 0.97, 0.97

NDCG: Example 2

Rank	Hit?
1	
2	X
3	X
4	X
5	

$$DCG_5 = \frac{1}{\log_2 2} + \frac{1}{\log_2 3} + \frac{1}{\log_2 4} = 2.13$$

$$IDCG_5 = 1 + \frac{1}{\log_2 2} + \frac{1}{\log_2 3} = 2.63$$

$$nDCG_5 \frac{DCG_5}{IDCG_5} \approx 0.81$$

Online vs Offline Experimentation

Offline experimentation	Online experimentation
Ratings, transactions	Ratings, feedback
Historic session (not all recommended items are rated)	Live interaction (all recommended items are rated)
Ratings of unrated items unknown, but interpreted as “bad” (default assumption, user tend to rate only good items)	“Good/bad” ratings of not recommended items are unknown
If default assumption does not hold: True positives may be too small False negatives may be too small	False/true negatives cannot be determined

Beyond Accuracy

- Understanding that good recommendation accuracy alone does not give users of recommender systems an effective and satisfying experience
- For instance, a recommender might achieve high accuracy by only computing predictions for easy-to-predict items—but those are the very items for which users are least likely to need predictions
- Coverage:
  - The coverage of a recommender system is a measure of the domain of items in the system over which the system can form predictions or make recommendations
  - Coverage can be most directly defined on predictions by asking “What percentage of items can this recommender form predictions for?
  - What percentage of available items does this recommender ever recommend to users? – more popular with ecommerce
  - Must be measured in combination with accuracy to prevent bogus predictions

Beyond Accuracy

- Obvious recommendations have two disadvantages:
  - Customers who are interested in those products have already purchased them
  - We do not need recommender systems to tell them which products are popular overall.
- Novelty
  - recommendation system that simply recommends movies that were directed by the user’s favorite director
  - If the system recommends a movie that the user wasn’t aware of, the movie will be novel, but probably not serendipitous.
  - A simple modification is to create a list of “obvious” recommendations, and remove the obvious ones from each recommendation list before presenting it to users.
- Serendipity
  - a recommender that recommends a movie by a new director is more likely to provide serendipitous recommendations

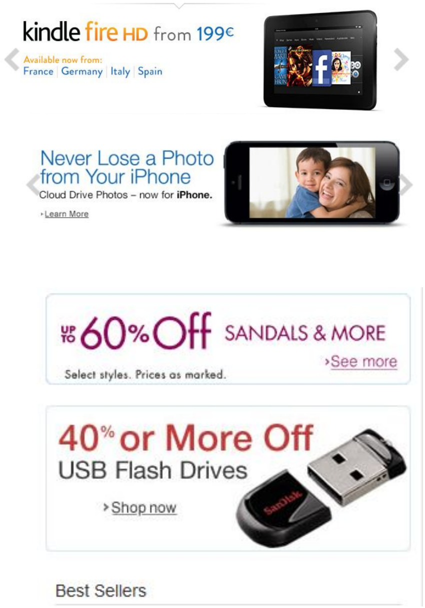
Discussion & Summary

- Focus on how to perform empirical evaluations on historical datasets
- Discussion about different methodologies and metrics for measuring the accuracy or coverage of recommendations.
- Overview of which research designs are commonly used in practice.
- From a technical point of view, measuring the accuracy of predictions is a well accepted evaluation goal
  - but other aspects that may potentially impact the overall effectiveness of a recommendation system remain largely underdeveloped.

What is a good recommendation?

What are the measures in practice?

- Total sales numbers
- Promotion of certain items
- …
- Click-through-rates
- Interactivity on platform
- …
- Customer return rates
- Customer satisfaction and loyalty


kindle fire HD from 199€
Available now from:
France | Germany | Italy | Spain
Never Lose a Photo from Your iPhone
Cloud Drive Photos – now for iPhone.
Learn More
UP TO 60% Off SANDALS & MORE
See more
Select styles. Prices as marked.
40% or More Off
USB Flash Drives
Shop now
SanDisk
Best Sellers