



ACM RecSys Challenge 2019

Two-stage Model for Automatic Hotel Recommendation at Scale

XIANZHE ZHANG, XIAO WANG, JIAOKUN LIU
MENTOR: ROBERT PALOVICS

CS341 Project in Mining Massive Data Sets
Stanford University

Task Introduction – ACM RecSys Challenge

Find your ideal hotel on **Trivago**

The screenshot displays the Trivago search interface. At the top, the Trivago logo is centered, with currency options for USD and EN, and a 'Sign up' link on the right. The search bar contains 'San Francisco'. Below the search bar, there are fields for 'Check-in' (Thu, 05/09/19) and 'Check-out' (Sat, 05/11/19), and a 'Room' selection (One-person room). A blue 'Search' button is to the right. Below these fields, there are filters for 'Price / night' (set to \$600+), 'Accommodation' (All types), 'Guest rating' (All), 'Hotel location' (City center), and 'More filters' (Select). A 'View Map' button is visible on the left. On the right, there is a 'Sort by' dropdown set to 'Our recommendations' and a link for 'How payments to us affect ranking'. The main search results area shows 'Hotel The Club Donatello' with a 4-star rating, 8.6 score, and 4322 reviews. The hotel is located in San Francisco, 0.2 miles to Union Square. A price comparison table shows the hotel's website at \$358, Booking.com at \$359, and Expedia at \$359 for 2 nights for \$717. A 'View Deal' button is prominently displayed.

USD EN Sign up

San Francisco

Check-in Thu, 05/09/19 < > Check-out Sat, 05/11/19 < > Room One-person room Search

Price / night \$600+ Accommodation All types Guest rating All Hotel location City center More filters Select

View Map

Sort by Our recommendations

How payments to us affect ranking

Hotel The Club Donatello
★★★★ Hotel
San Francisco, 0.2 miles to Union Square
8.6 Excellent (4322 reviews)
Excellent location · Extremely clean

Hotel Website \$358
Booking.com \$359
Cancelon.com \$475
More deals from \$358

Hotelwiz \$592
\$359
Expedia
2 nights for \$717
View Deal

Dataset and Evaluation Metrics

user_id	session_id	timestamp	step	action_type	reference	platform	city	device	current_filters	impressions
93F7WGHBP03A	569f5ea70df51	1541543231	1	search for destination	Barcelona, Spain	US	Barcelona, Spain	desktop		
93F7WGHBP03A	569f5ea70df51	1541543269	2	filter selection	Focus on Distance	US	Barcelona, Spain	desktop	Focus on Distance	
93F7WGHBP03A	569f5ea70df51	1541543269	3	search for poi	Port de Barcelona	US	Barcelona, Spain	desktop	Focus on Distance	
93F7WGHBP03A	569f5ea70df51	1541543371	4	interaction item deals	40255	US	Barcelona, Spain	desktop		
93F7WGHBP03A	569f5ea70df51	1541543425	5	clickout item	40255	US	Barcelona, Spain	desktop		6744 40181 40630 84610 2282416 1258693 974937 147509 128238 7998246 40255 3058538 1637385 40285 147502 921707 40849 6757 12770 893733 685091 147522 40708 860451 6819
93F7WGHBP03A	569f5ea70df51	1541543741	6	search for item	81770	US	Barcelona, Spain	desktop		
93F7WGHBP03A	569f5ea70df51	1541543770	7	interaction item info	81770	US	Barcelona, Spain	desktop		
93F7WGHBP03A	569f5ea70df51	1541543813	8	clickout item	81770	US	Barcelona, Spain	desktop		6832 40396 6621784 40197 6743 147488 40635 6177052 6742 1319782 40763 945255 83855 39937 1870125 1354432 6812 82400 40181 6834 81770 5056102 40797 923935 40284

- **Mean Reciprocal Rank (MRR)**

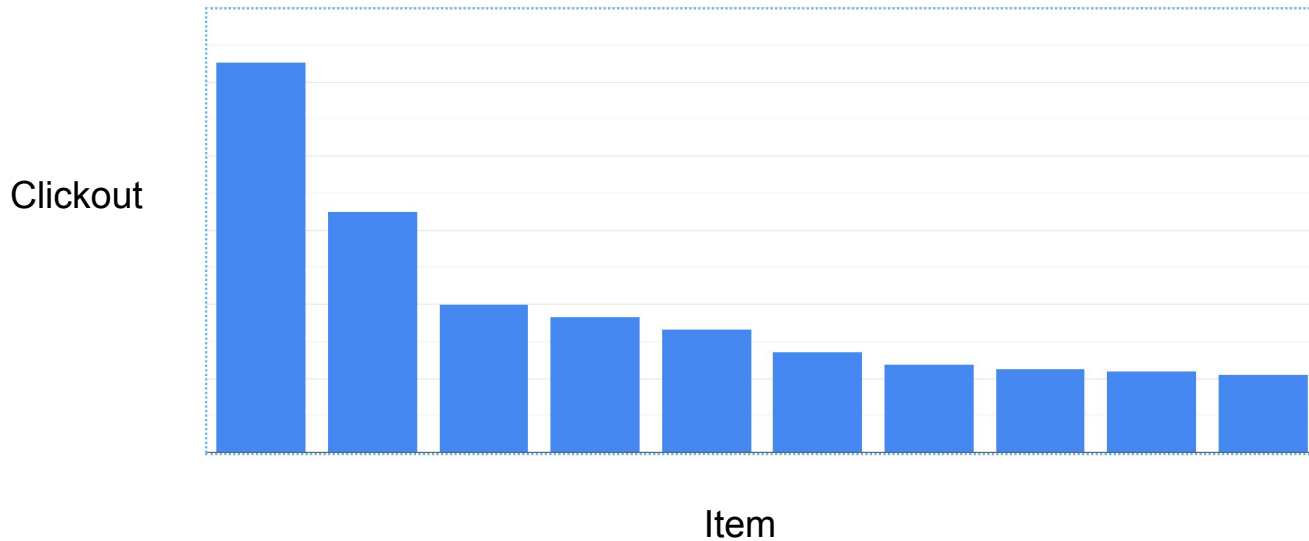
$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

Query	Proposed Results	Correct response	Rank	Reciprocal rank
cat	catten, cati, cats	cats	3	1/3
tori	torii, tori , toruses	tori	2	1/2
virus	viruses , virii, viri	viruses	1	1

$$(1/3 + 1/2 + 1) / 3 = 0.6111$$

Baseline Model - Based on Popularity

- Get the number of clickout that each item received
- The final submission will have an impression list sorted according to the number of clickout per item



Results

Leaderboard



MRRScore

Baseline 0.28

Results

Leaderboard



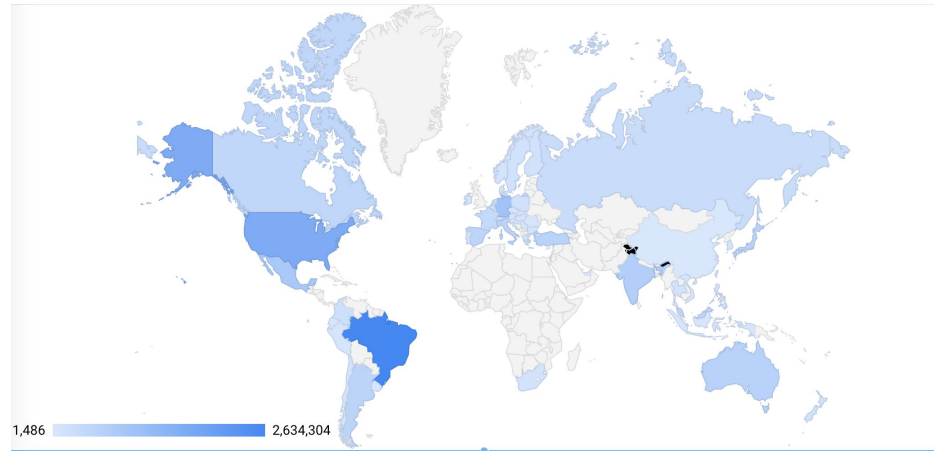
MRRScore

Transition Matrix 0.50

Baseline 0.28

Feature Engineering - Statistics

- Users : 730, 803
- Sessions: 826, 842
- Clickout: 910, 683
- Records: 15, 932, 992
- Time range: 6 days
- User Distribution: Asian, Europe, North and South America



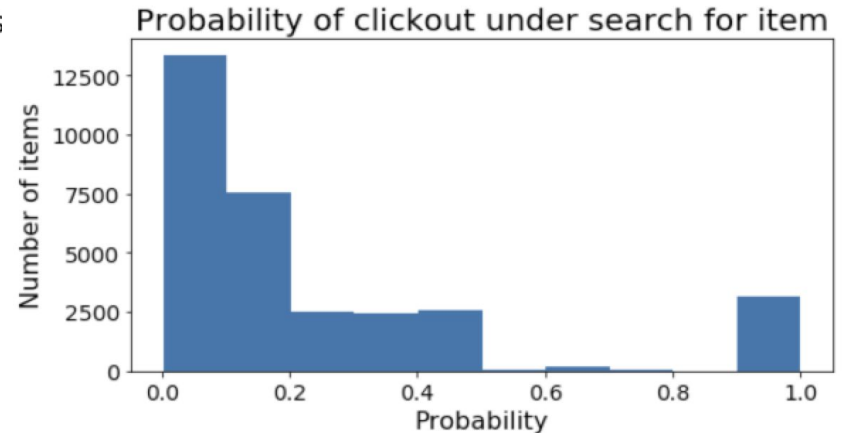
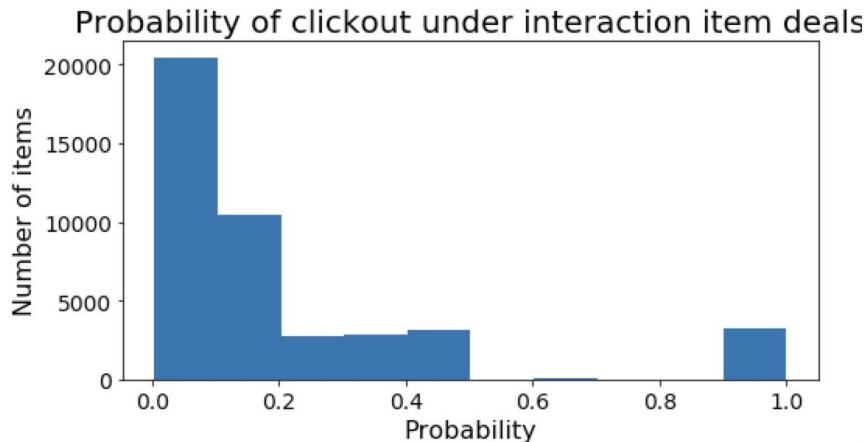
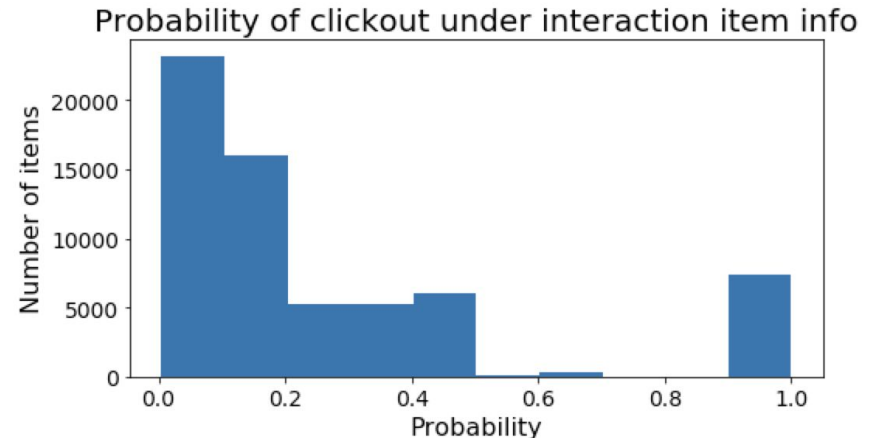
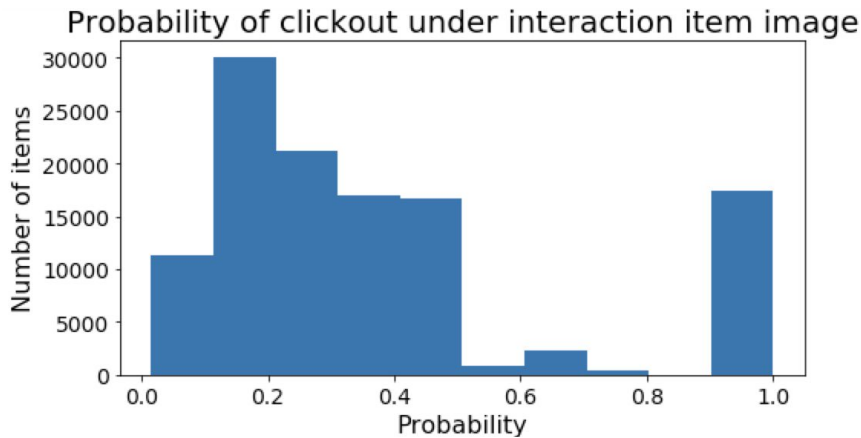
Feature Engineering - Session Features

- **Interact before:** The user interacted with the item before in another session
- **Position in the list:** The front positions are more likely to be chosen
- **First interact:** The item first interaction in one session period

session_id	step	reference	item id
f7c78f27	1	interaction item info	7818446
f7c78f27	2	interaction item image	7818446
f7c78f27	3	interaction item image	7818446
f7c78f27	4	interaction item deals	7818446
f7c78f27	5	interaction item deals	7818446
f7c78f27	6	interaction item deals	7818446
f7c78f27	7	interaction item deals	7818446
f7c78f27	8	search for item	2681512
f7c78f27	9	interaction item image	2681512
f7c78f27	10	interaction item image	2681512
f7c78f27	11	clickout item	2681512
f7c78f27	12	interaction item image	2099360
f7c78f27	13	interaction item image	2099360
f7c78f27	14	interaction item image	929533
f7c78f27	15	interaction item image	929533
f7c78f27	16	interaction item image	929533
f7c78f27	17	interaction item image	929533

Feature Engineering - Interaction Features

interaction item image, interaction item info,
interaction item deals, search for item



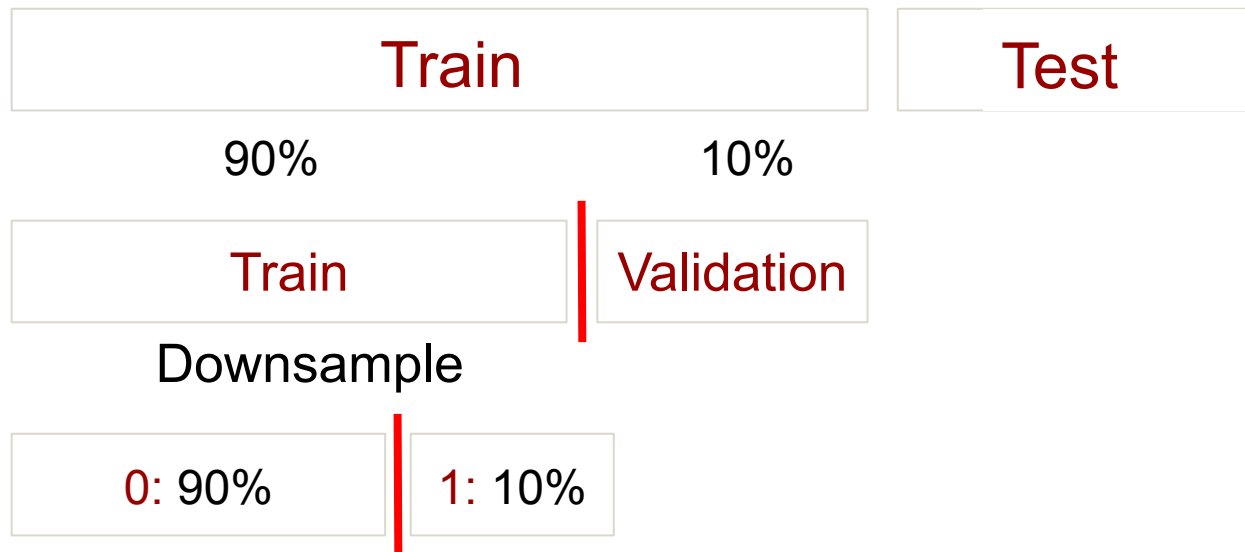
Methodology - Binary Classification

Label

Positive examples 1: clicked out item

Negative examples 0: unclicked out items

Pipeline



Logistic Regression (LR) as baseline: AUC 0.78

Decision Tree (DT): AUC 0.80

XGBoost with hyper - parameters tuning: AUC 0.83

Results

Leaderboard



MRRScore

XGBoost 0.58

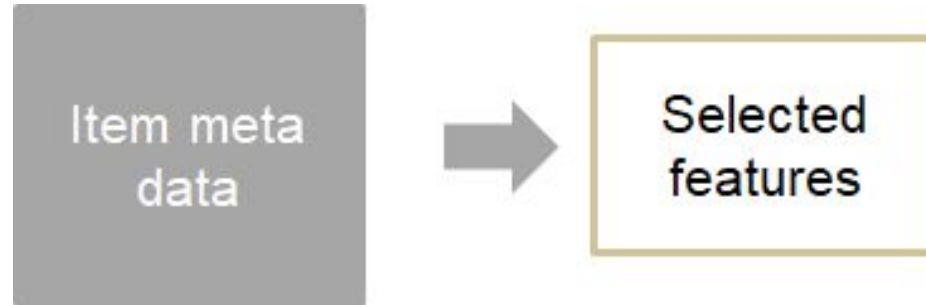
LR, DT: 0.57

Transition Matrix 0.50

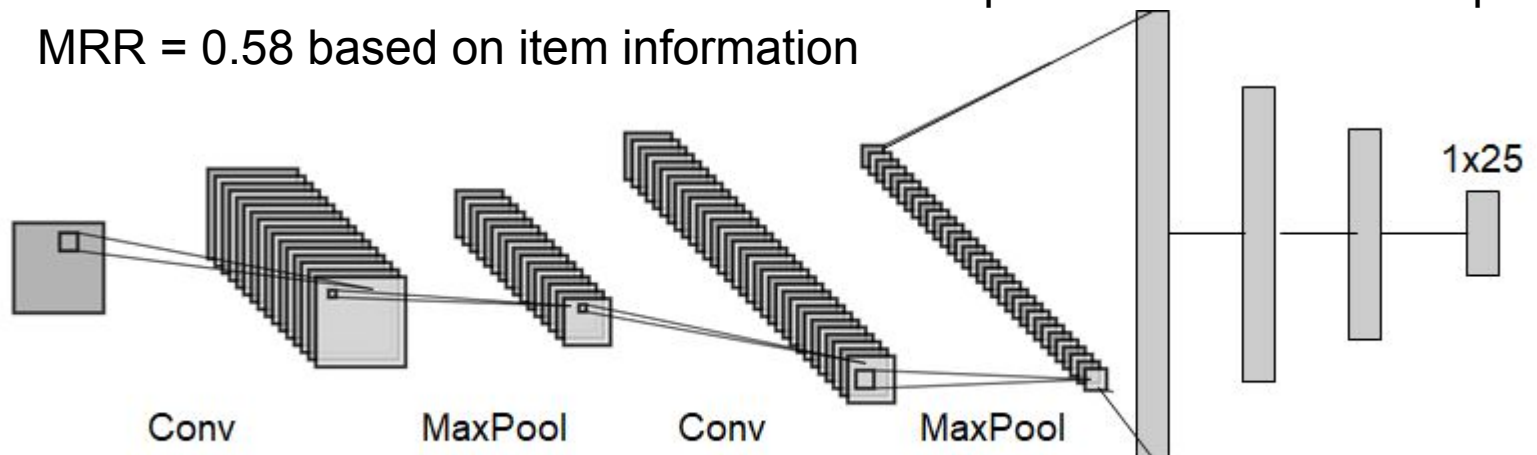
Baseline 0.28

More useful features – item metadata

- Over 150 items properties can be derived from data given
- Directly input into model decreases performance
- SVD can reduce redundant information
 - Five-star hotels always have wifi

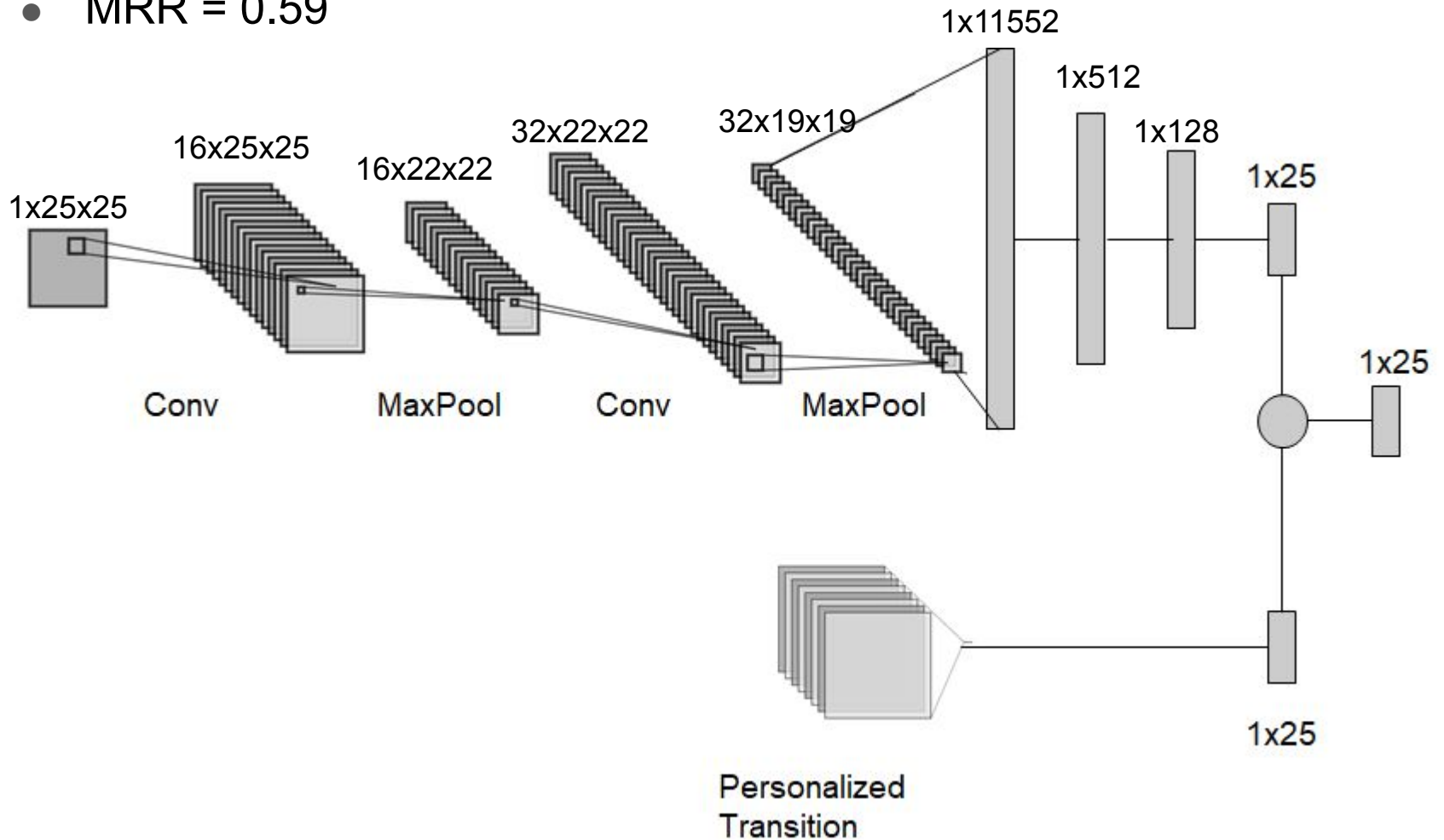


- We need a model to find features' relationship and items' relationship
- MRR = 0.58 based on item information

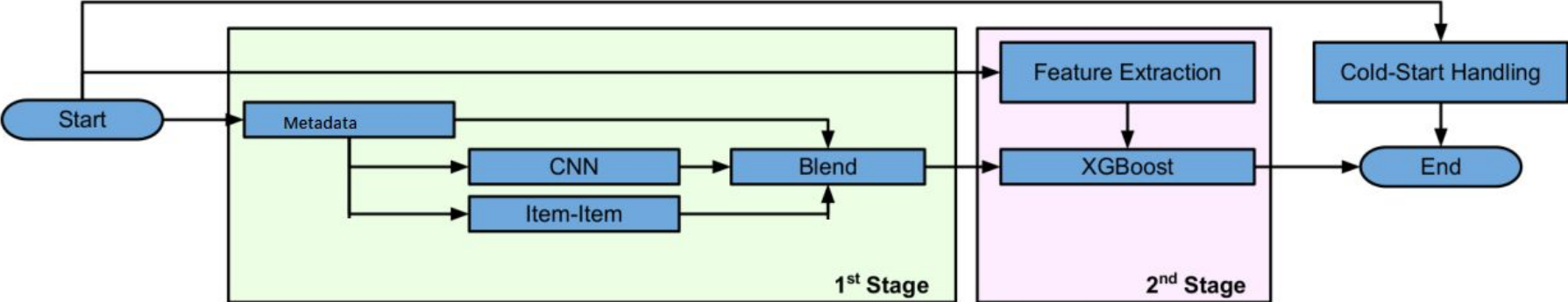


CNN based Model

- Input user-item interaction info by concatenating transition prob
- MRR = 0.59



Full pipeline



MRR

Leaderboard



MRRScore

Ensemble 0.60

CNN 0.59

XGBoost 0.58

Transition Matrix 0.50

Baseline 0.28

Lessons Learned and future work

- Importance of feature engineering
- Large intermedia result
- Test more models