


ACM RecSys Challenge 2019

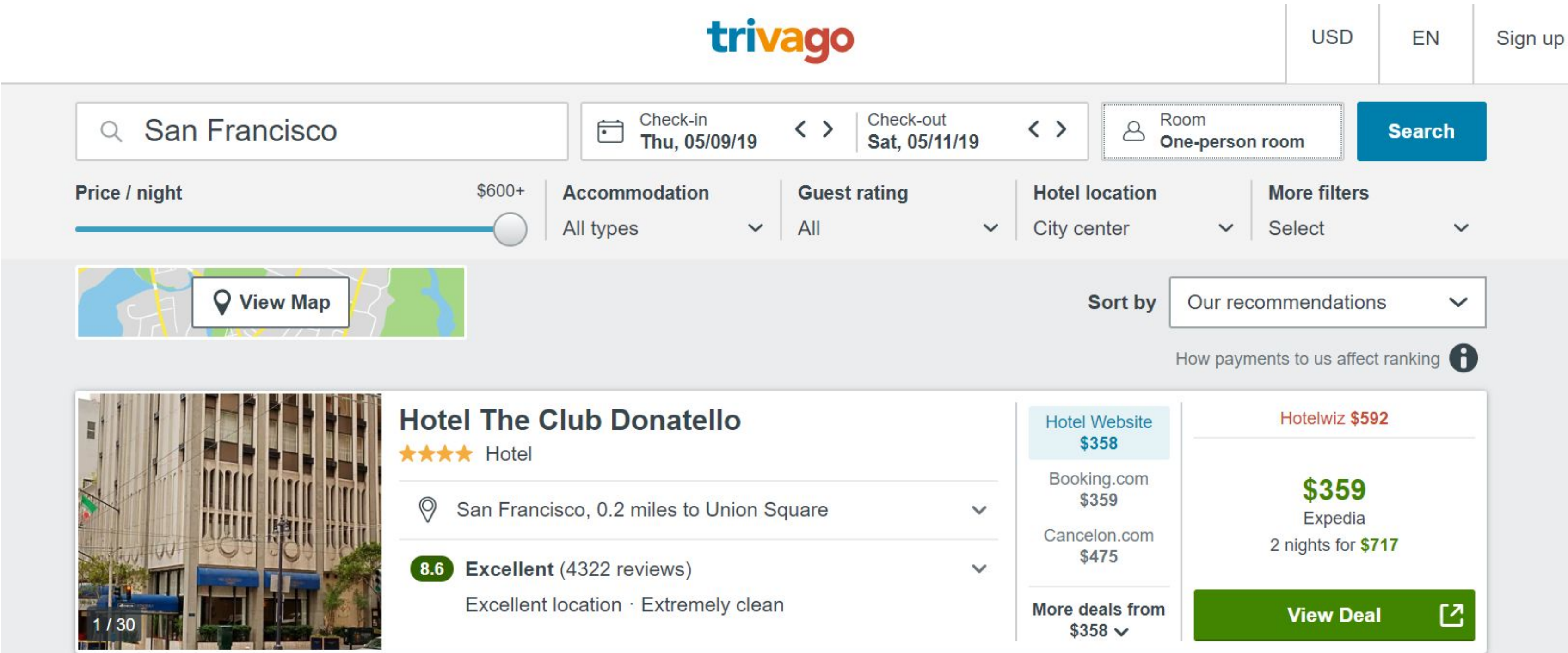
Two-stage Model for Automatic Hotel Recommendation at Scale

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



Dataset and Evaluation Metrics

<table>
<tr><th>user_id</th><th>session_id</th><th>timestamp</th><th>step</th><th>action_type</th><th>reference</th><th>platform</th><th>city</th><th>device</th><th>current_filters</th><th>impressions</th></tr>
<tr><td>93F7WGHBPO3A</td><td>569f5ea70df51</td><td>1541543231</td><td>1</td><td>search for destination</td><td>Barcelona, Spain</td><td>US</td><td>Barcelona, Spain</td><td>desktop</td><td></td><td></td></tr>
<tr><td>93F7WGHBPO3A</td><td>569f5ea70df51</td><td>1541543269</td><td>2</td><td>filter selection</td><td>Focus on Distance</td><td>US</td><td>Barcelona, Spain</td><td>desktop</td><td>Focus on Distance</td><td></td></tr>
<tr><td>93F7WGHBPO3A</td><td>569f5ea70df51</td><td>1541543269</td><td>3</td><td>search for poi</td><td>Port de Barcelona</td><td>US</td><td>Barcelona, Spain</td><td>desktop</td><td>Focus on Distance</td><td></td></tr>
<tr><td>93F7WGHBPO3A</td><td>569f5ea70df51</td><td>1541543371</td><td>4</td><td>interaction item deals</td><td>40255</td><td>US</td><td>Barcelona, Spain</td><td>desktop</td><td></td><td></td></tr>
<tr><td>93F7WGHBPO3A</td><td>569f5ea70df51</td><td>1541543425</td><td>5</td><td>clickout item</td><td>40255</td><td>US</td><td>Barcelona, Spain</td><td>desktop</td><td></td><td>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</td></tr>
<tr><td>93F7WGHBPO3A</td><td>569f5ea70df51</td><td>1541543741</td><td>6</td><td>search for item</td><td>81770</td><td>US</td><td>Barcelona, Spain</td><td>desktop</td><td></td><td></td></tr>
<tr><td>93F7WGHBPO3A</td><td>569f5ea70df51</td><td>1541543770</td><td>7</td><td>interaction item info</td><td>81770</td><td>US</td><td>Barcelona, Spain</td><td>desktop</td><td></td><td></td></tr>
<tr><td>93F7WGHBPO3A</td><td>569f5ea70df51</td><td>1541543813</td><td>8</td><td>clickout item</td><td>81770</td><td>US</td><td>Barcelona, Spain</td><td>desktop</td><td></td><td>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</td></tr>
</table>

- **Mean Reciprocal Rank (MRR)**

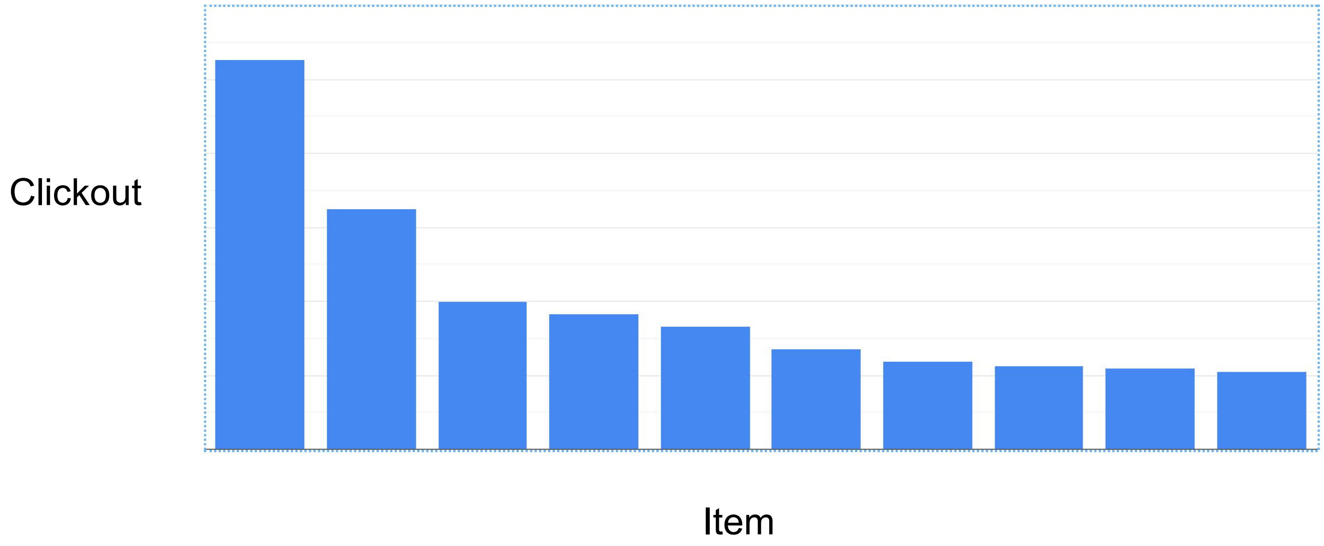
$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}.$$

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

(⅓ + ½ + 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

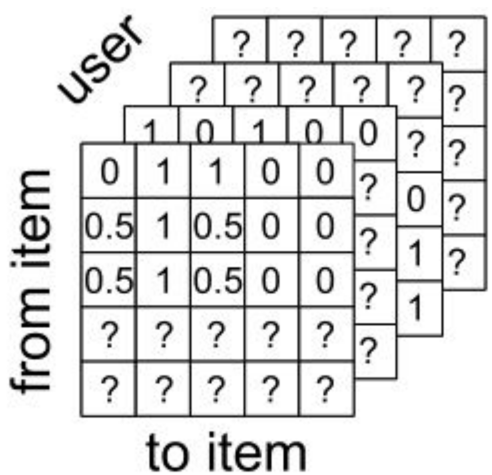
Transition Matrix

- Item-Item transition (0.3)

$$P = \begin{bmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,j} & \dots & P_{1,S} \\ P_{2,1} & P_{2,2} & \dots & P_{2,j} & \dots & P_{2,S} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{i,1} & P_{i,2} & \dots & P_{i,j} & \dots & P_{i,S} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{S,1} & P_{S,2} & \dots & P_{S,j} & \dots & P_{S,S} \end{bmatrix}.$$

$$\hat{a}_{l,i} = \hat{p}(i \in B_t | l \in B_{t-1}) = \frac{\hat{p}(i \in B_t \wedge l \in B_{t-1})}{\hat{p}(l \in B_{t-1})} = \\ = \frac{|\{(B_t, B_{t-1}) : i \in B_t \wedge l \in B_{t-1}\}|}{|\{(B_t, B_{t-1}) : l \in B_{t-1}\}|}$$

- Personalized item-item transition matrix (0.5)



Results

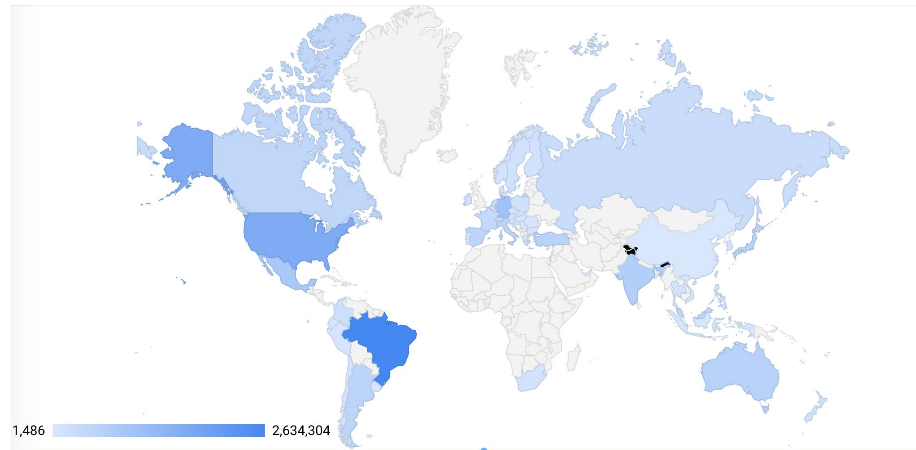
Leaderboard

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

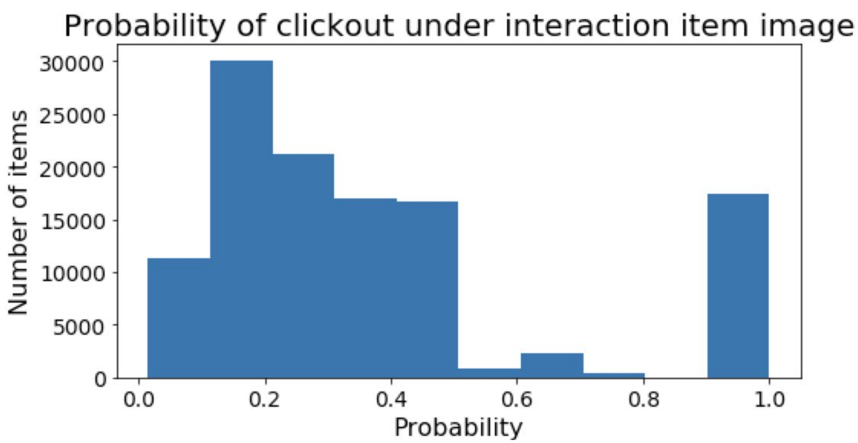
Feature Engineering - Session Features

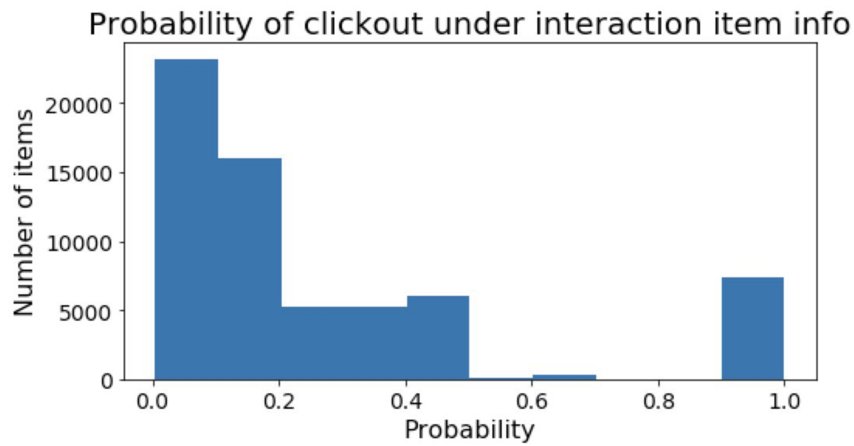
- Interact before: The user interacted with the item before in another session
- Position in the list: The front positions are more likely to be chosen
- First interact: The item first interaction in one session period

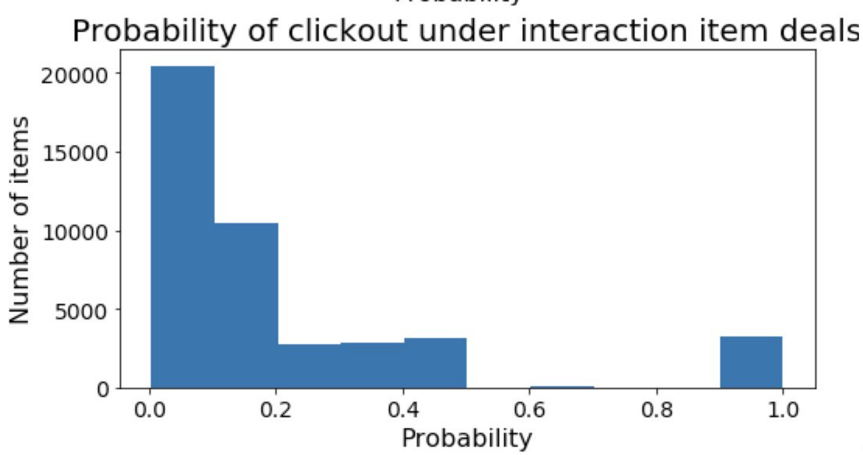
session_id	step	reference	item id
f7c78f27	1	interaction item info	7818446
f7c78f27	2	interaction item image	7818446
f7c78f27	3	interaction item image	7818446
f7c78f27	4	interaction item deals	7818446
f7c78f27	5	interaction item deals	7818446
f7c78f27	6	interaction item deals	7818446
f7c78f27	7	interaction 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

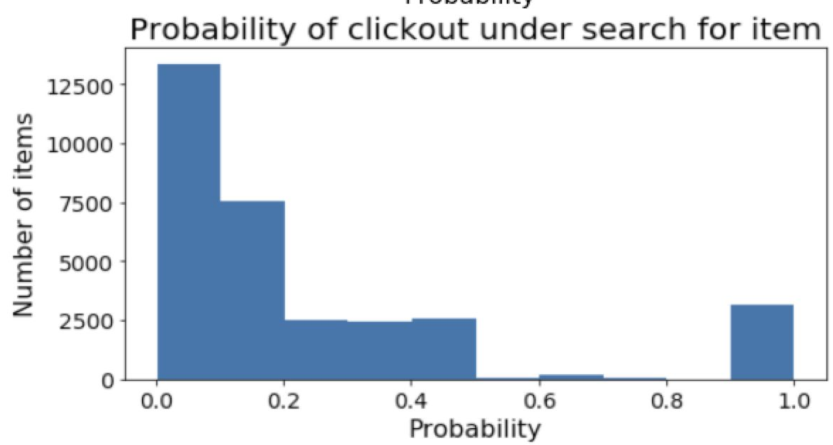
Feature Engineering - Interaction Features

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

Train | Test

90% | 10%

Train | Validation

Downsample

0: 90% | 1: 10%

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

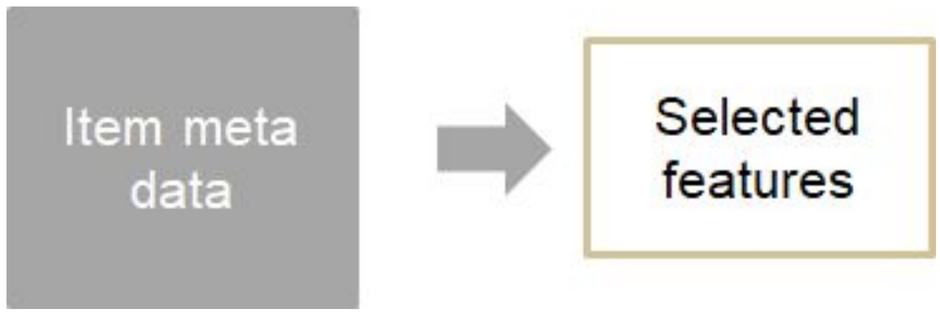
LR, DT: 0.57

Transition Matrix 0.50

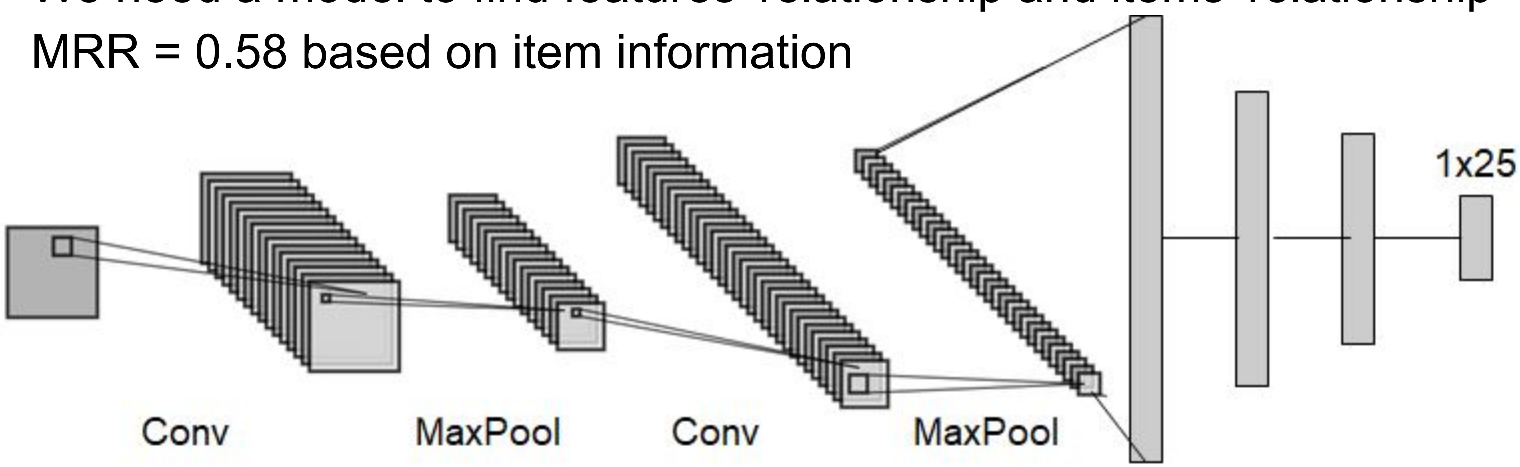
Baseline 0.28

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

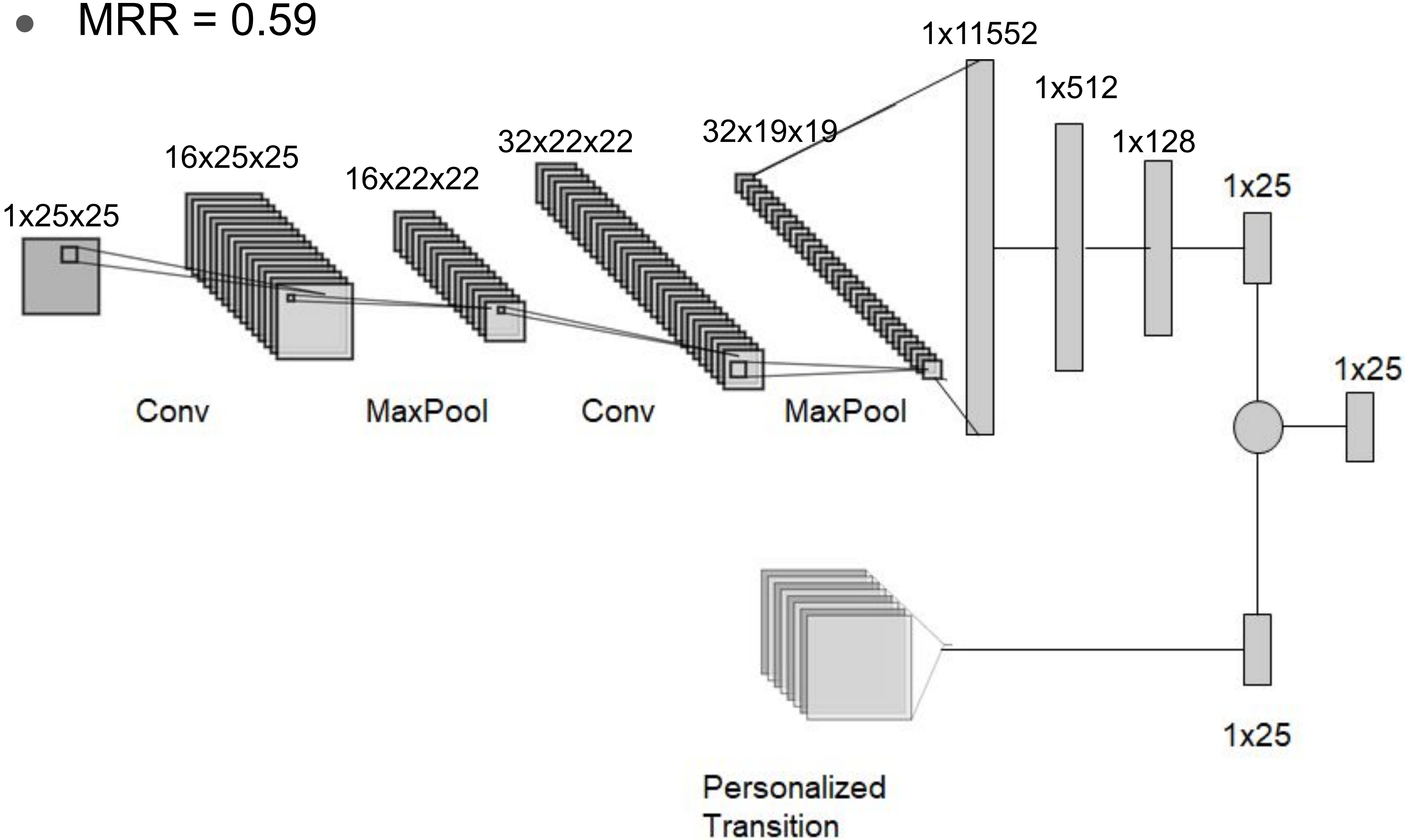


- We need a model to find features' relationship and items' relationship
- MRR = 0.58 based on item information

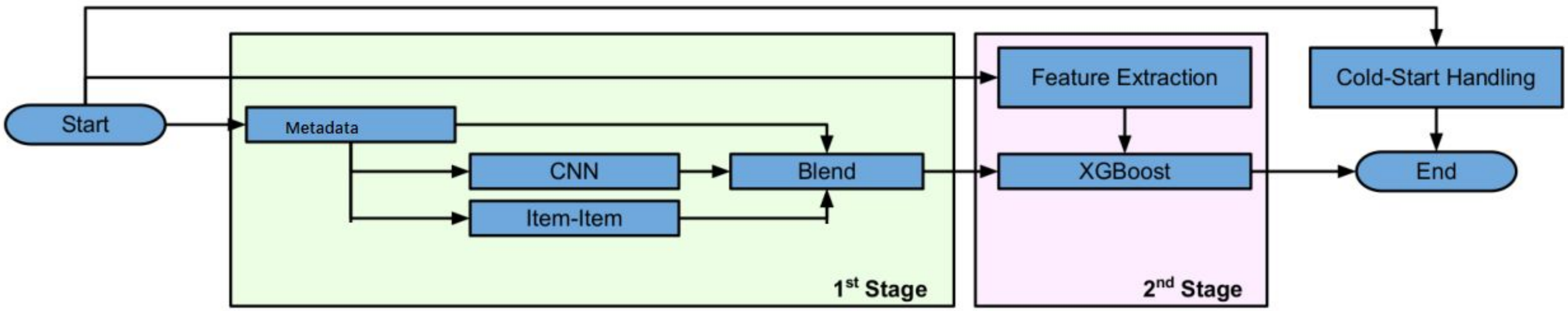


CNN based Model

- 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