

# RECOMMENDATIONS OF D & C / TAX BREAK MAXIMIZER

COMPANY/TT VISION: Taxes done for everyone with ease  
& STRATEGIC ALLARS

2) Confidence

3) putting more money into the pockets of our customers

CONTEXT: Over 400 D&C

Matching customer to D&C that can be claimed requires asking a ton of questions (rule-based) → Friction + makes tax

Instead if we can predict D&C customers are most likely to claim, we restrict questions to only applicable ones (we will need to ask questions since it is a compliance business) filing time-consuming

CURRENT SOL: Using customer's last year's tax data, predict D&C for (MULTI-LABEL CLASSIFICATION since told D&C ~400) this tax year using Item & User features

PROBLEMS IN CURRENT SOL: 1) Only applicable to returning users, New users  
2) Returning users tax sit. Changed drastically from last year.  
(50% of TT customers have life events esp. in age-grp 25-40) + COVID

BUSINESS: Take-rate of recommended D&C ↓ drastically  
740% to less than 10%  
→ 2% S2C ↓

SOLUTION: Section based Recommendation since 1) Not utilizing any prior data of customer from last year (prev. history does not matter)  
2) Standard Ded. Vs Itemized Ded: changes the universe of D&C applicable  
Scope:  
3) Only Federal + Itemized other team 12-13% of users: Itemized Ded



# TECHNICAL DESIGN - (SESSION BASED RECOMMENDATION)

ONLINE METRICS: 1) Take-rate of recommended D&C + in filed tax  
(users may drop D&C after clicking Haking it in the final filing process)

2) Counter Metric: Hunt-rate

3) S2C

→ sequential Recommendation

TRAINING DATA: ~~By~~ ✓ (Transformers4Rec): NVIDIA library using HF Transformers works with PyTorch  
+ user ID + features found in final filed tax form available  
~~user features available~~ ~~not completed~~

- Not → i) seq. of (D&C id + features) taken / clicked + abandoned  
user features available  
ii) seq. of (D&C id + features) where ~~where~~ than 50% recommended D&C not taken  
user features available

Seq of D&C id + other user features found in "completed" & filed tax returns  
(+ category id) (profile + W-2 cy)  
↓  
binarized initially + continuous  
e.g. age  
Age  
Has children  
No. of children  
Investments  
Debts  
Home loan  
Home loan amt  
↓  
hierarchical classification of D&C (using tax experts domain knowledge)  
prev. year data  
trained on regular intervals when substantial complete happen

At the time prediction, rule-based (profile + W-2 cy) to recommend first 3 D&C, then ML based recommendations  
(only added/taken D&C from recommended D&C become the features)

TRAINING METHODOLOGY: CAUSAL LM (autoregressive) → next word/item prediction  
Though other techniques such as MASKED LM (auto-encoding) can also be used (randomly masking items & using both left & right items for prediction)



## TASK & LOSS FUNCTION:

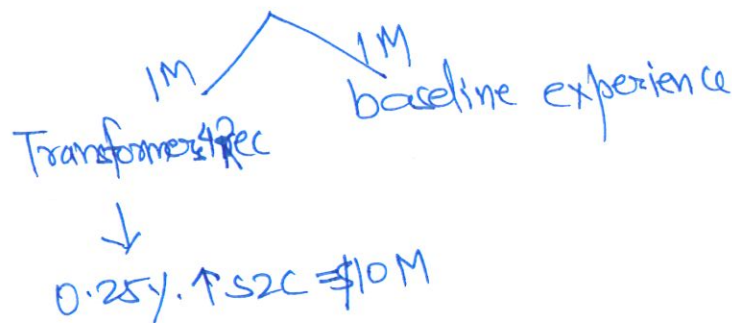
## Next-item Prediction

internally formulated as multi-class pred  
with cross-entropy loss

## OFFLINE METRICS:

Recall @ 10 → Provide as many tax breaks as possible  
NDCG @ 10

A/B TESTING & IMPACT DELIVERED: 1M users (both new & returning) who selected itemized deduction in each bucket



✓ uses masked self-attention for understanding contextual relationships b/w items in a session

ADDITIONAL NOTES: ✓ Transformer4Rec uses only encoder block of transformer architecture

1) Supports incremental finetuning: Models can be trained incrementally i.e. on data generated everyday

✓ 2) Supports leveraging additional tabular features of item metadata & user context (along with seq of items etc)

✓ 3) Dynamically creates all necessary layers (e.g. embedding layers) to encode, normalize and aggregate categorical & continuous features (using TabularSequenceFeatures block) e.g. uses embedding table learnt during training at the time of prediction using lookup table

4) NextItem Prediction head projects the output of Transformer block to items space, followed by softmax layer to produce relevance scores for all items

✓ 5) supports multi-task learning

CHALLENGES: 1) Cold start of "Seq": <sup>Sol<sup>n</sup></sup> Heuristic based first 3 D&C recomm.  
(profile + W2 uploaded)

2) D&C recommended could be very different: Category Id  
Using tax experts help, devised Hierarchical  
classification of D&C and assigned  
a number to every D&C + appended  
it as a feature  
This number signifies how close each D&C  
is to other D&C based on experts  
domain knowledge

TRADEOFFS: Build embeddings for each D&C  
using data / IRS publications available  
(time-intensive) v/s Use just D&C id  
to build & launch first  
version

YOUR CONTRIBUTION: 1. Product Strategy & RoadMap : Customer problem,  
opp. sizing  
2. E2E ML System Design  
3. Prototyping the first Transformer4Rec model  
4. Leading E2E development + deployment + integration  
with other subsystems  
(Cross-collab)

WHAT COULD HAVE BEEN DONE DIFFERENTLY? Investing in incremental  
fine-tuning  
earlier