2021-Fall AI+X Final Report

Finance: Risk Evaluation and Prediction

JIAJIE ZHANG　YANCHENG LIANG

Advised by: YI WU
JINYAO ZHANG

Index

1. Introduction

Unsecured Personal Loan

- A small amount of money
- Not secured by property

Contradiction

Many people need it
Issuance of loan is strict

Financial Risk Management

- Crucial for all finance institution

the percentage of bad debts

◆ In this case, a key factor for financial risk

Lessons: P2P companies

Better risk prediction model → Lower percentage of bad debts

better risk management direct outcome by reducing financial loss
more people not satisfied by traditional financial institute can be covered

Risk Prediction Model

Finance Side

Calls for Breakthrough

Algorithm Stagnated

- XGBoost has been used for many years

Hard for Small Banks

- do not have ability to build a complex model, many use logistic regression
- Seek collaboration with fintech firms

Data-Driven

- Acquire more data is nearly the only way to make better evaluation

Machine Learning Side

General Real-world Data

- **Multimodal**

  Have both sequential and non-sequential

- **General Industrial Irregular Data**

  A lot of noise, missing values in data

  Both categorial and real-valued data

2. Specific Problem

Build an end-to-end risk prediction model

Credit Report
Basic information and past credit behavior

Future Performance
Risk of future overdue behavior

Modern Deep Learning Techniques

Inspired by recommend system, where deep learning has already revolutionized the recommendation algorithm.

Goal: make improvement over past algorithm (XGBoost), and use deep learning techniques to utilize sequential data which is hard for XGBoost

Input: Credit Report

Seq of Loan

- Due Date
- Issuance Date
- Type
- Repay Period
- Repay Frequency
- …

Seq of Query

- Query Organization
- Reason
- …

Seq of Credit Card

- Card Type
- Issuance Date
- Organization
- Currency Type
- …

Non-Seq

- Location
- Age
- Education
- …

Train data: 430865 users (20.06 – 21.05)

Test data: 152131 users (21.06 – 21.07)

Out-of-time (OOT) Evaluation

Output: Predict Overdue Behavior

Labels (overdue behaviors in different degree) including:

- i1label15: The first payment is overdue for more than 15 days
- i2label30: The Second payment is overdue for more than 30 days
- overdue15: Any payment is overdue for more than 15 days

Finally it is a binary classification problem

Evaluation Metric: AUC

Area under curve: $\mathbb{E}_{x \in D^{+}, y \in D^{-}} [\text{score}(x) > \text{score}(y)]$

An increase of 0.01 of AUC is significant, which can reduce roughly 5% of bad debts

3. Methods & Experiments

Challenges & Solutions

1. Data imbalance

- i1label15: negative **43 : 1** positive; overdue15: negative **6 : 1** positive;

➢ Oversampling (bootstrap) / Weighted BCE loss;

2. Complex and noisy data, hard to learn.

- Many NAN & 0; Category + Real Value;

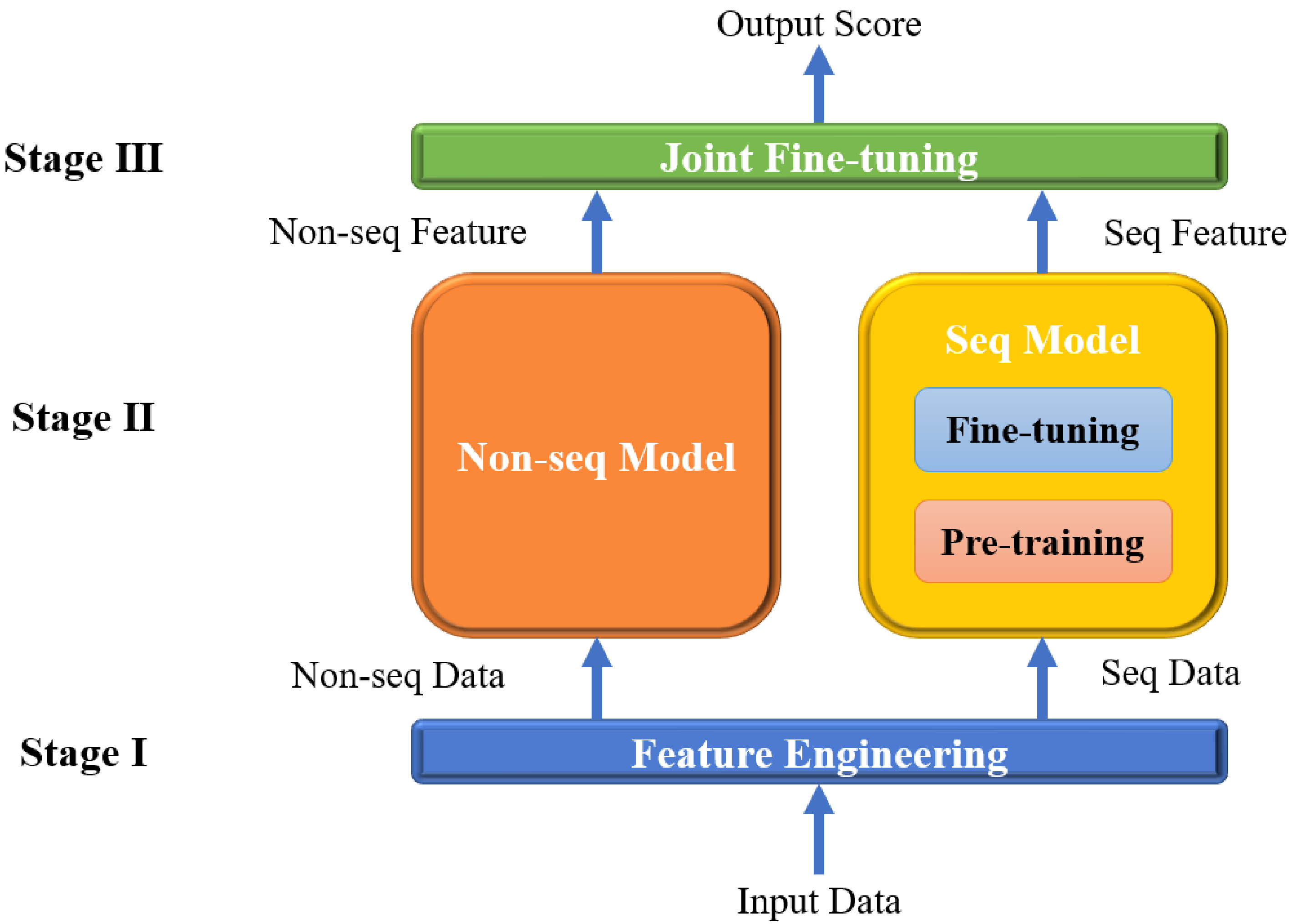
➢ Feature engineering / Pre-training on seq data.

3. Data distribution varies over time.

- Consumers are first filtered by company's ground decision model which becomes better over time.

➢ Try more general (but harder) label (overdue15) in training.

Overall Pipeline



Feature Engineering

- There are enormous “non sense”: create indicator

1

332000+ “0”s	13000+ “NaN”s	less than 600 1, 0.5, 0.1, …
Create [x!=0]	[x==NAN]	x

2

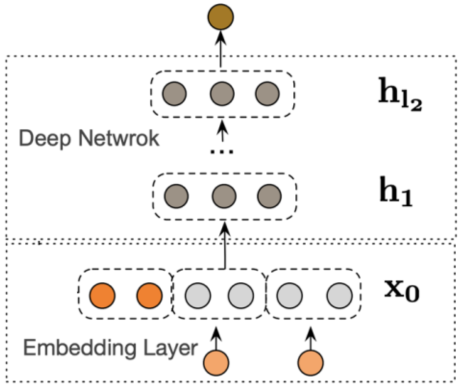
Outlier: will be clipped after normalization

- Processed non-seq data is more than 10,000 dimensions. So we first utilize XGBoost to select the most important 900 dimensions as input of non-seq data.

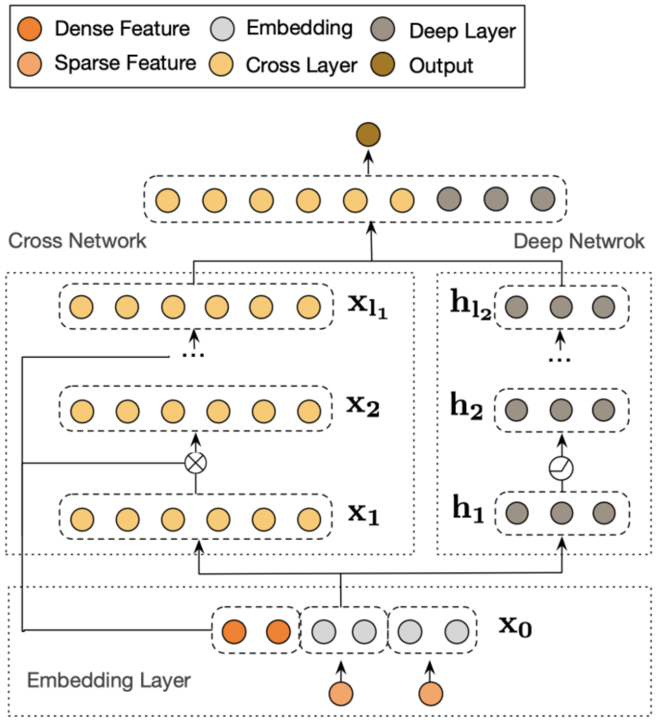
Non-seq Model

- Refer to popular models in recommender system:

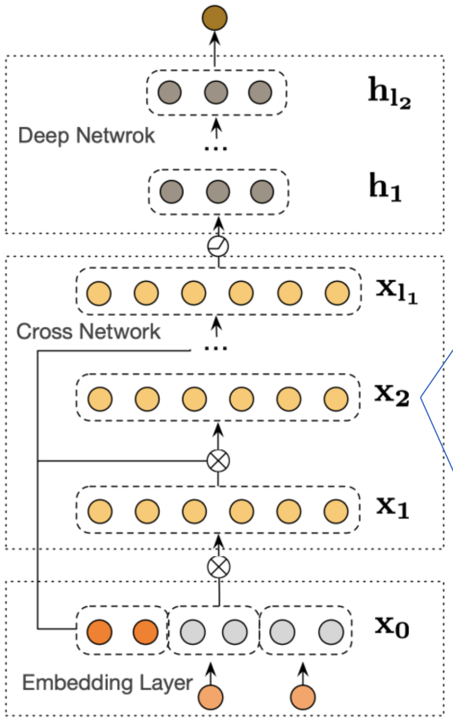
MLP



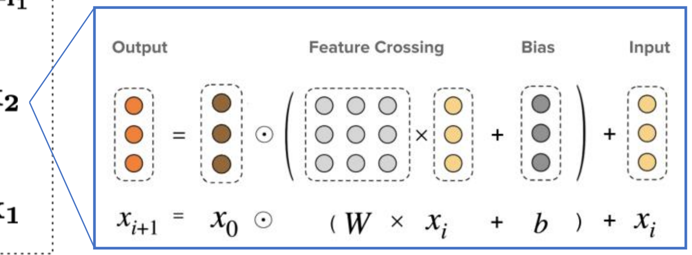
Parallel DCN



Stacked DCN



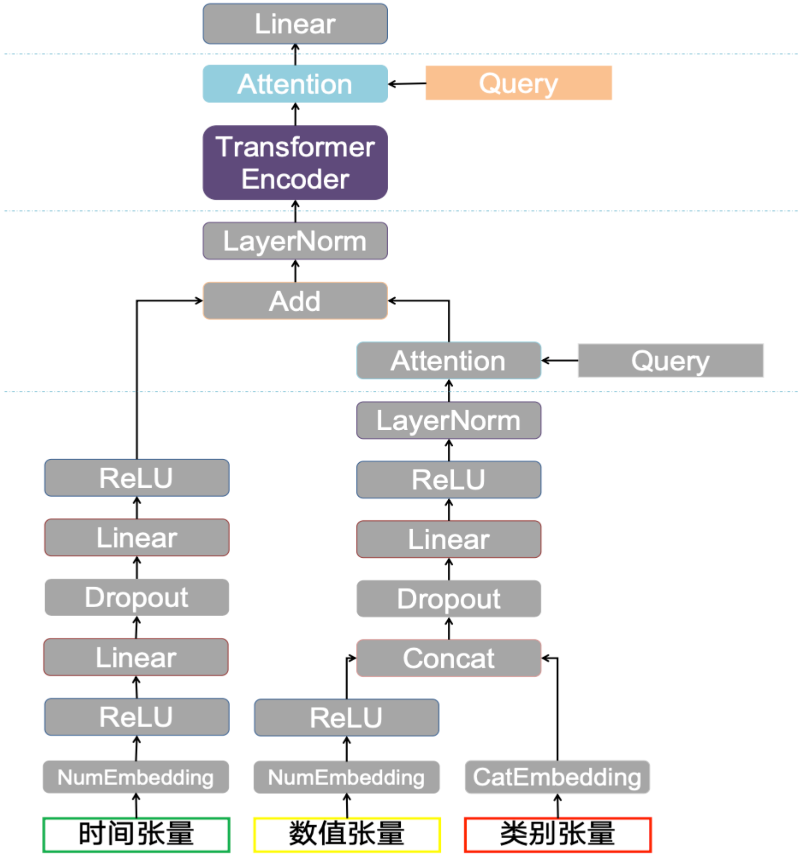
cross layer



Wang, R., Shivanna, R., Cheng, D., Jain, S., Lin, D., Hong, L., & Chi, E. (2021, April). DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems. In *Proceedings of the Web Conference 2021* (pp. 1785-1797).

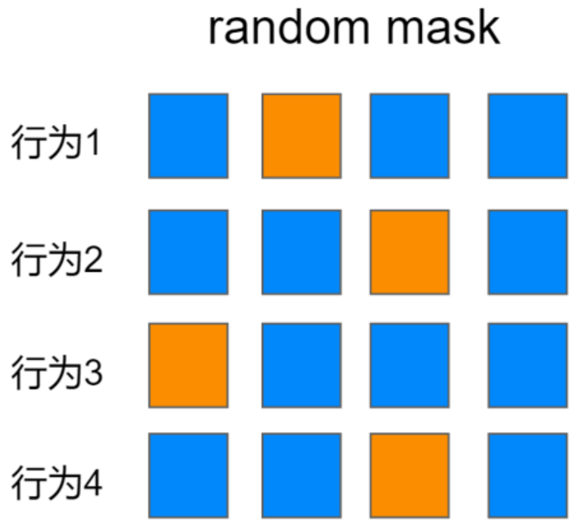
Seq Model

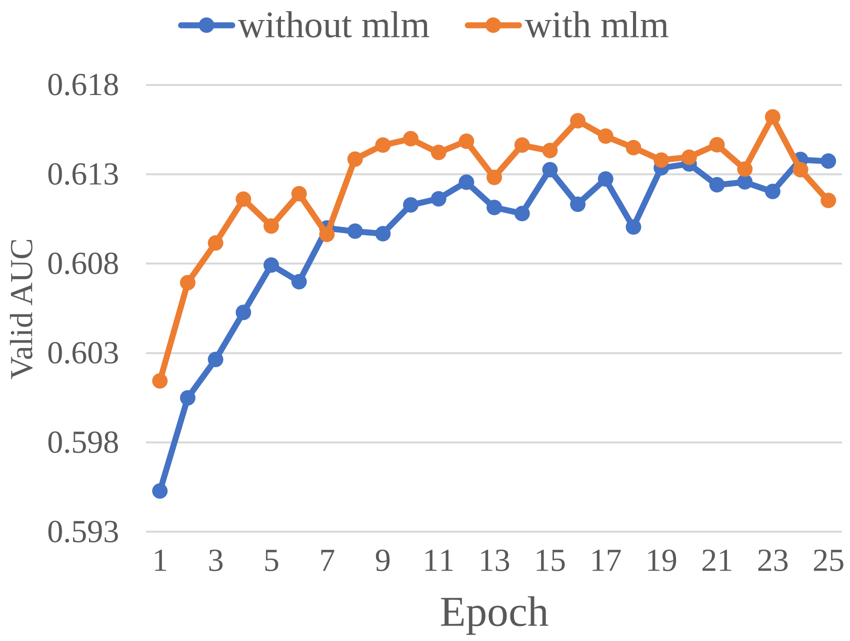
- Transformer based model
  - Time data → position embedding
  - Use attention to integrate feature embeddings.



Masked Language Model Pre-training

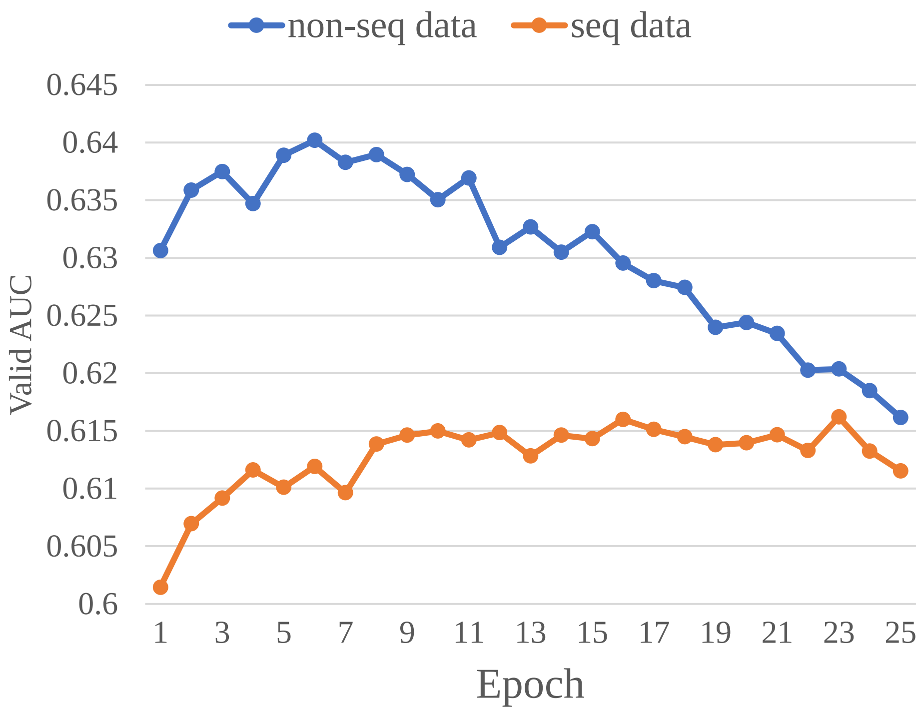
- First learn general knowledge about credit data by MLM to make the downstream classification task easier to learn.
- Output seq feature is input into different cls. heads for different type of data.

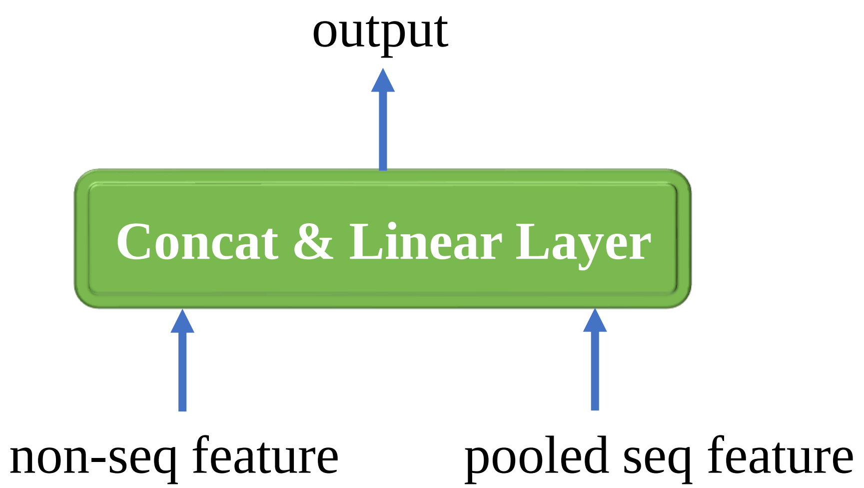




Joint Fine-tuning

- The hardness of learning non-seq and seq data is different;
- First train non-seq and seq models respectively, then jointly fine-tune them.





Main Results

Model	Training label	Oversample	i1label30	i2label30	i3label30
non-seq model					
XGBoost(Baseline)	overdue15	no	0.6418	0.6282	0.6187
SDCN	overdue15	no	0.6450	0.6319	0.6236
PDCN	overdue15	no	0.6483	0.6343	**0.6254**
MLP	overdue15	no	**0.6499**	**0.6349**	**0.6254**
seq model					
Baseline	i1label15	yes	0.5803	0.5711	0.5630
Pooled MLP	i1label15	yes	0.6065	0.5855	0.5736
LSTM	overdue15	no	0.6108	0.5936	0.5859
Transformer	overdue15	yes	0.6132	0.5941	0.5871
MLM + Transformer	overdue15	yes	**0.6156**	**0.5971**	**0.5885**
joint model					
Add Attn Net	overdue15	no	0.6504	0.6369	0.6285
Mul Attn Net	overdue15	no	0.6520	0.6377	0.6278
Concat Net	overdue15	no	**0.6546**	**0.6398**	**0.6297**

Table 1: Main results and best training configurations (training label, oversampling) for all models. When we adopt oversampling, we use weighted BCE loss. Otherwise we use BCE loss.

- Our best non-seq and seq models improve i1labe30 AUC by **0.0081** and **0.0353** over baselines, respectively;
- Complex models do not necessarily perform better;
- Joint fine-tuning of non-seq and seq models can achieve better results.

4. Summary & Discussion

Summary & Discussion

- In financial projects, data noise, complicated distribution, data imbalance and other problems are very common, so it is very important to use feature engineering to get clean data.

- Compared with XGBoost, deep learning can achieve comparable or even better results, and can deal with sequential data well. But more complex networks do not necessarily perform better, and feature engineering and training parameters have more obvious effects.

- XGBoost has excellent model extensibility. Just adding the output score of deep network into training of XGBoost as a new feature works well. It can be seen that the future trend will be the combination of deep learning and traditional ML algorithms.