Collaborative Embedding Features and Diversified Ensemble for E-Commerce Repeat Buyer Prediction

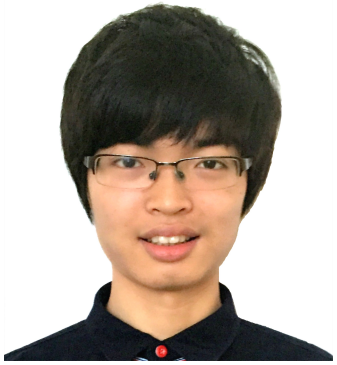
Zhanpeng Fang*, Zhilin Yang*, Yutao Zhang
Department of Computer Science and Technology
Tsinghua University
* Indicates equal contributions

Results

- Team “FAndy&kimiyoung&Neo”
- 2nd place in stage 1
- 3rd place in stage 2
- The only team marching in top 3 of both stages

Team Members

- Zhanpeng Fang
  - Master student, Tsinghua-Carnegie Mellon dual program



- Zhilin Yang
  - PhD student, Carnegie Mellon Univ.

- Yutao Zhang
  - PhD student, Tsinghua Univ.

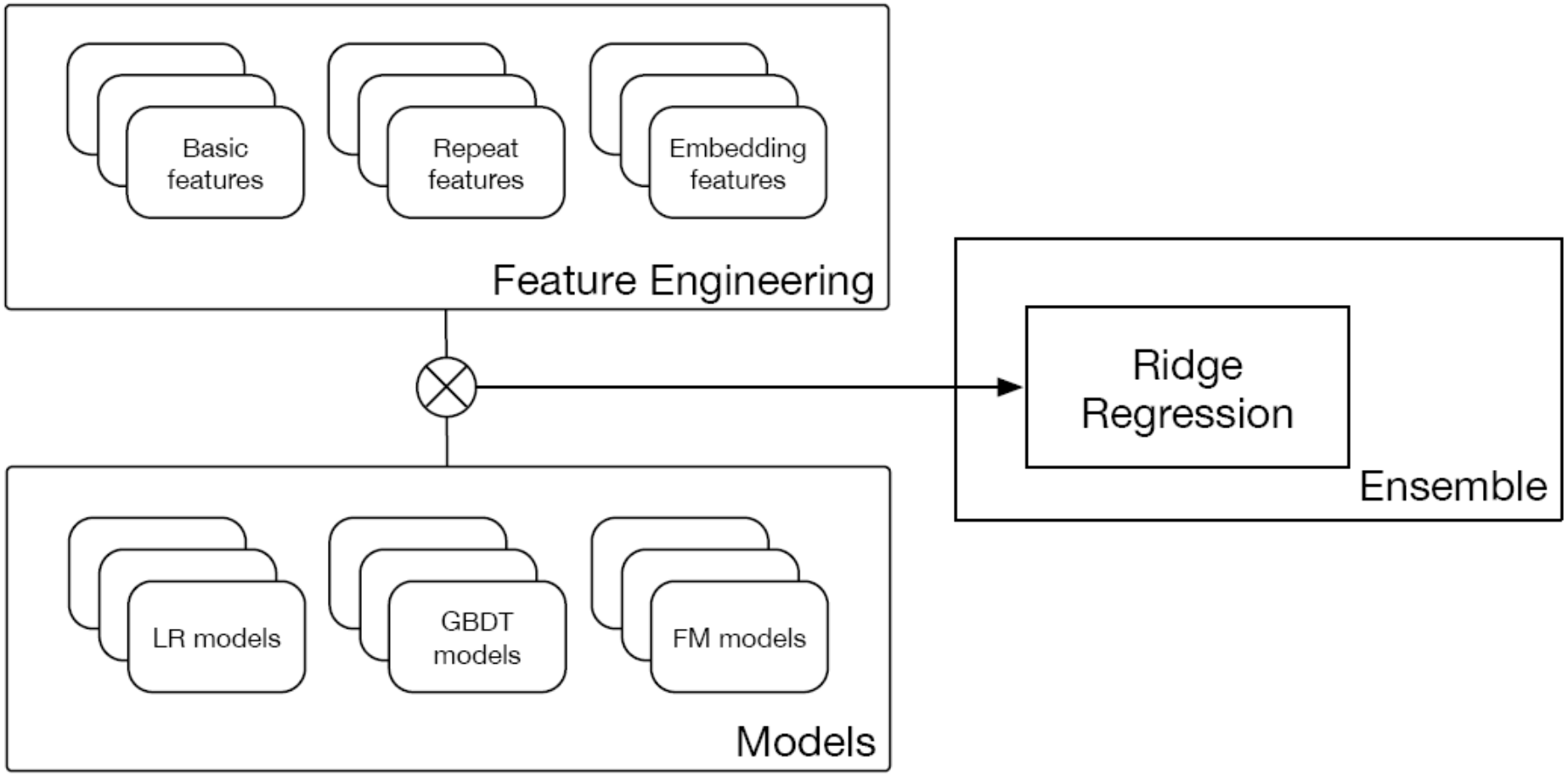
Presenter: FreakOut Inc.

- A digital marketing technology company focusing on online advertisements
  - Launched the first demand side platform in Japan
  - Utilizing machine learning for prediction of click-through rate and conversion rate

Challenges

- Heterogeneous data
  - User, merchant, category, brand, item
- Repeat buyer modeling
  - What are the characteristic features for modeling repeat buyer?
- Collaborative information
  - How to leverage the collaborative information between users and merchants [in a shared space]?

Framework

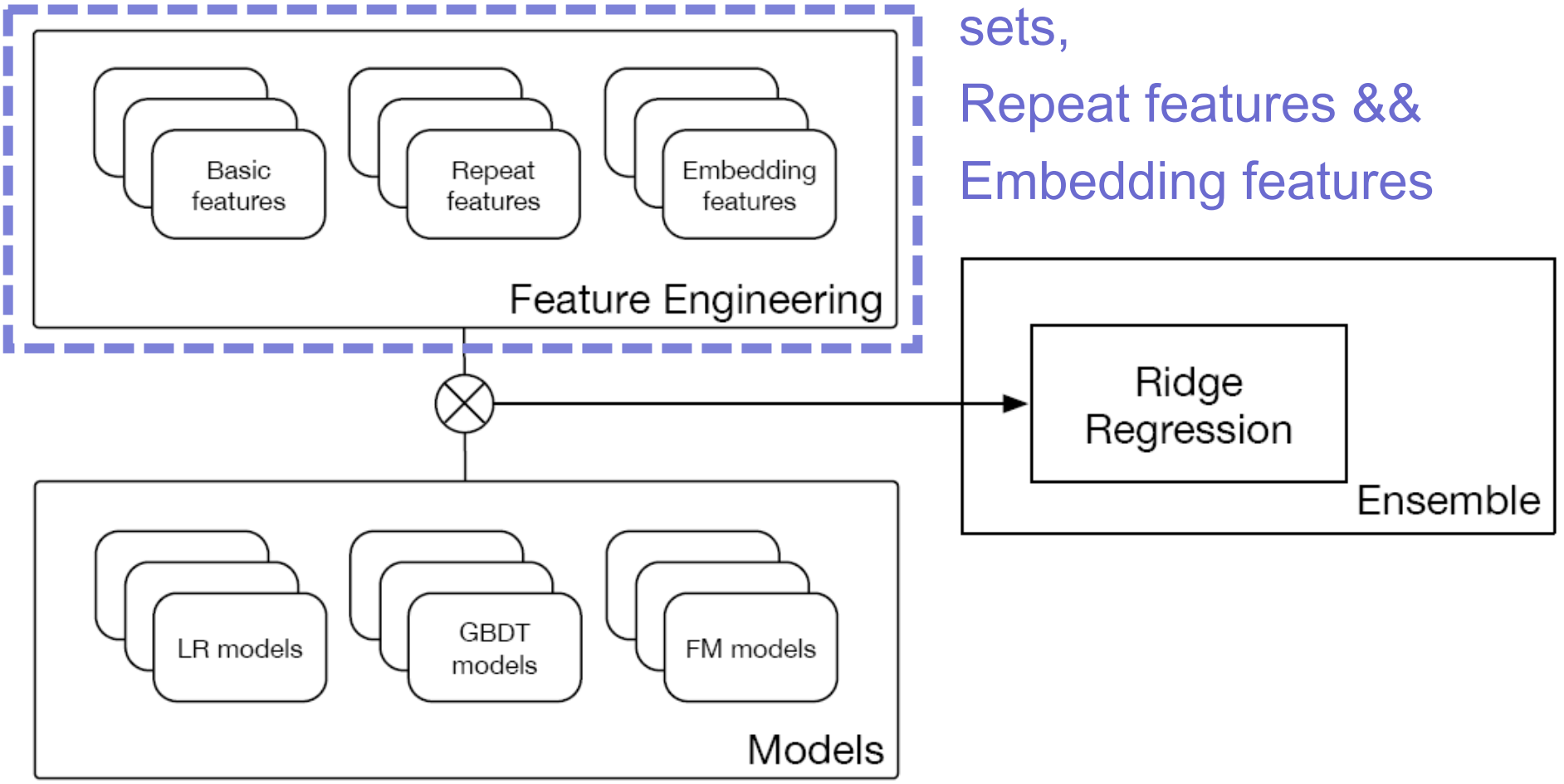

Basic features
Repeat features
Embedding features
Feature Engineering
Ridge Regression
Ensemble
LR models
GBDT models
FM models
Models

Framework

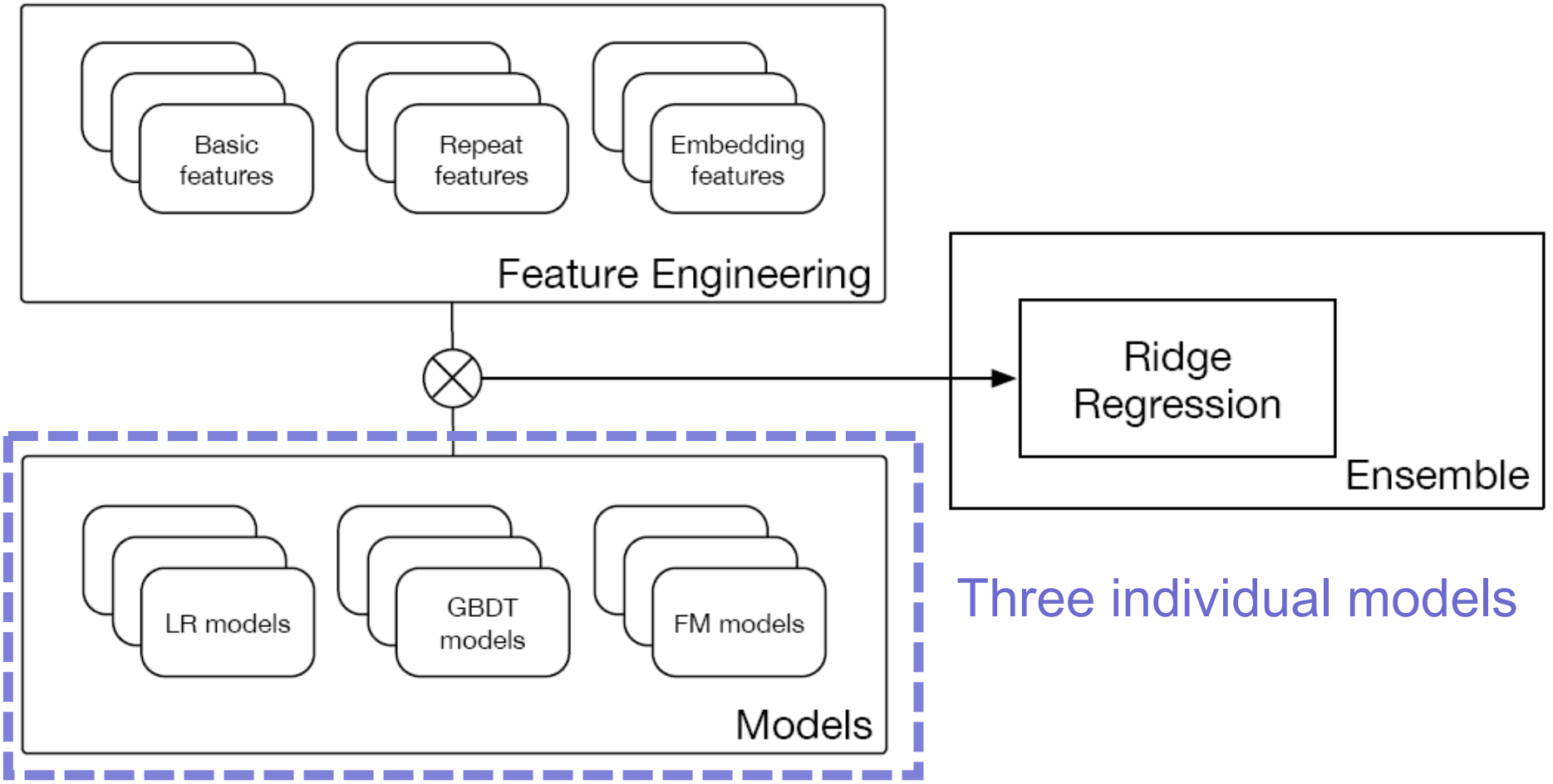
Two novel feature sets,

Repeat features &&

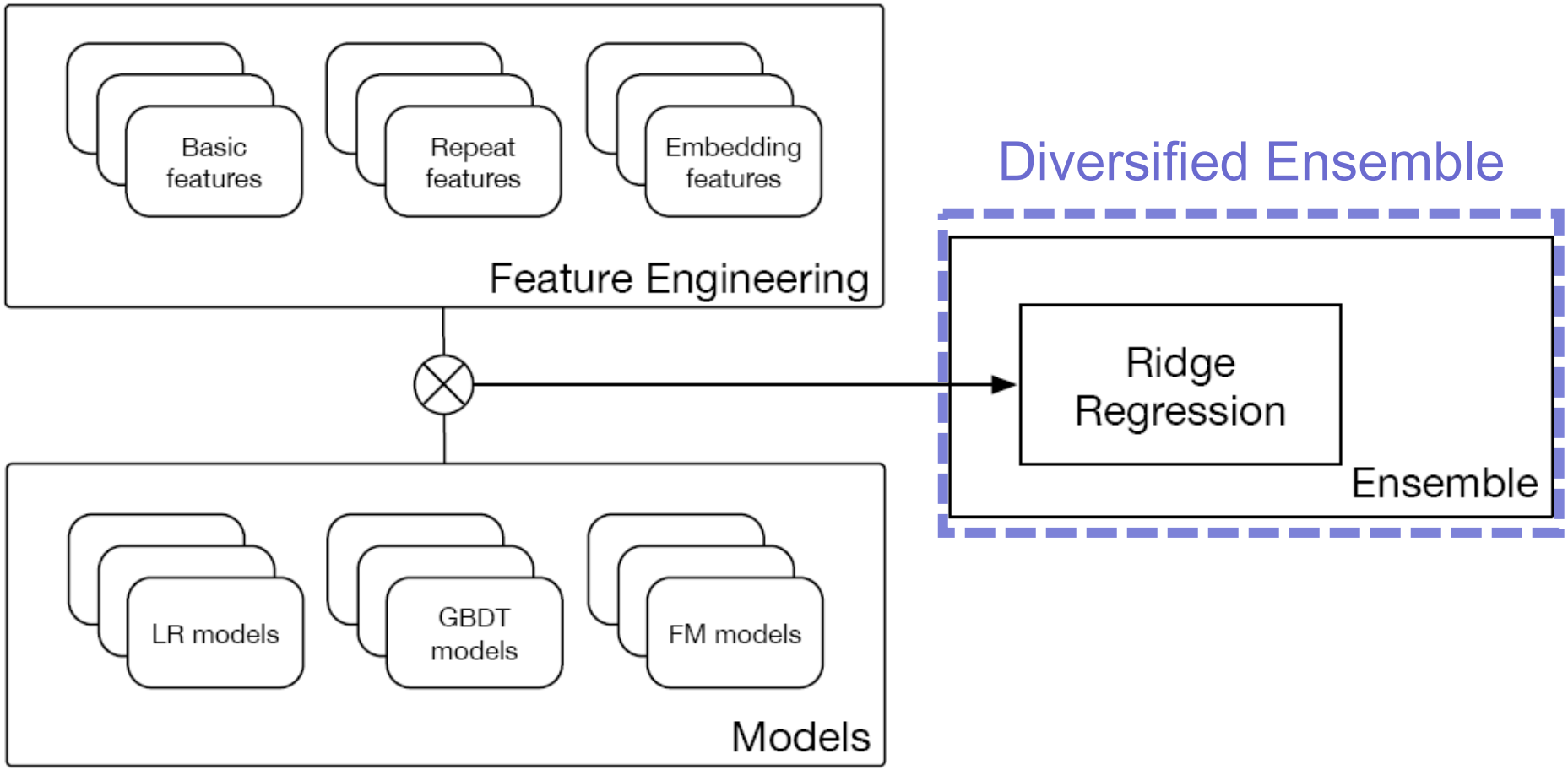
Embedding features



Framework


Basic features
Repeat features
Embedding features
Feature Engineering
Ridge Regression
Ensemble
LR models
GBDT models
FM models
Models
Three individual models

Framework


Basic features
Repeat features
Embedding features
Feature Engineering
Diversified Ensemble
Ridge Regression
Ensemble
LR models
GBDT models
FM models
Models

Feature Engineering – Basic Features

- User-Related Features
  - Age, gender, # of different actions
  - #items/merchants/… that clicked/purchased/favored
  - **Omitting add-to-cart in all actions related features increases performance (since almost identical to purchase)**

- Merchant-Related Features
  - Merchant ID
  - #actions and #distinct users that clicked/purchased/ favored (only in Stage 1)

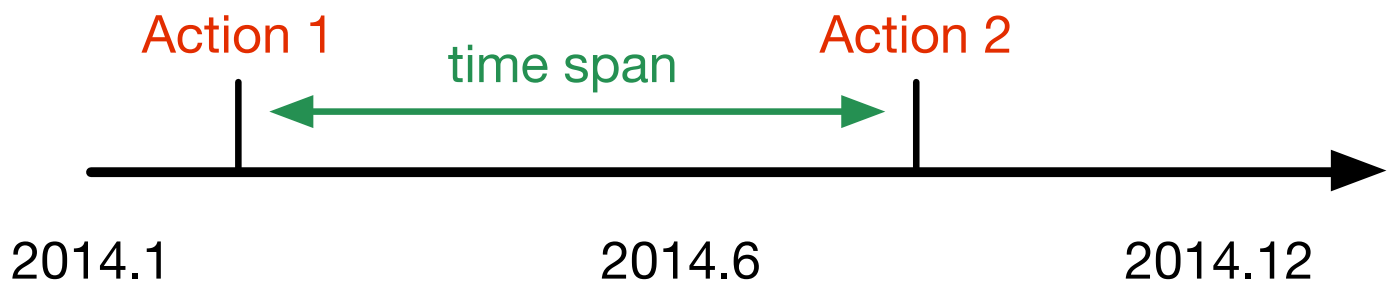
Feature Engineering – Basic Features

- User-Merchant Features
  - # different actions
  - Category IDs and brand IDs of the purchased items

- Post Processing
  - Feature binning in Stage 1
  - Log(1+x) conversion in Stage 2
  - **Perform similarly. Both much better than raw values.**

Repeat Features

- User Repeat Features
  - Average span between any two **actions**
  - Average span between two **purchases**
  - How many days since last **purchase**



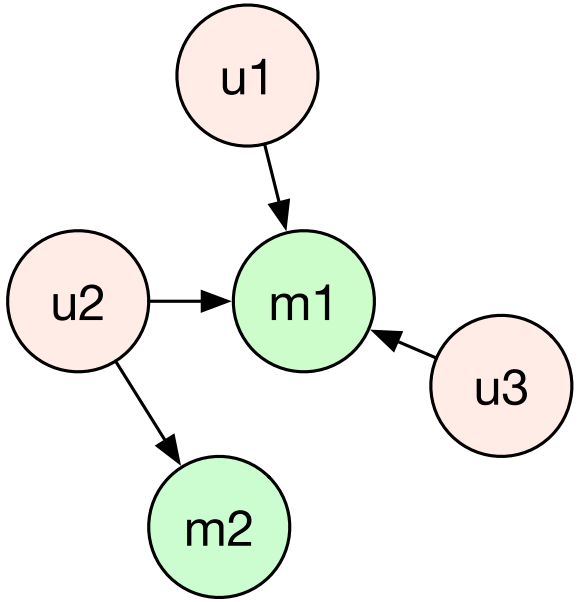
Repeat Features

- User-Merchant/Category/Brand/Item Repeat Features
  - **Average active days** for one merchant/ category/brand/item
  - **Maximum active days** for one merchant/ category/brand/item
  - **Average span** between any two actions for one merchant/category/brand/item
  - **Ratio** of merchants/categories/brands/items with **repeated actions**

Repeat Features

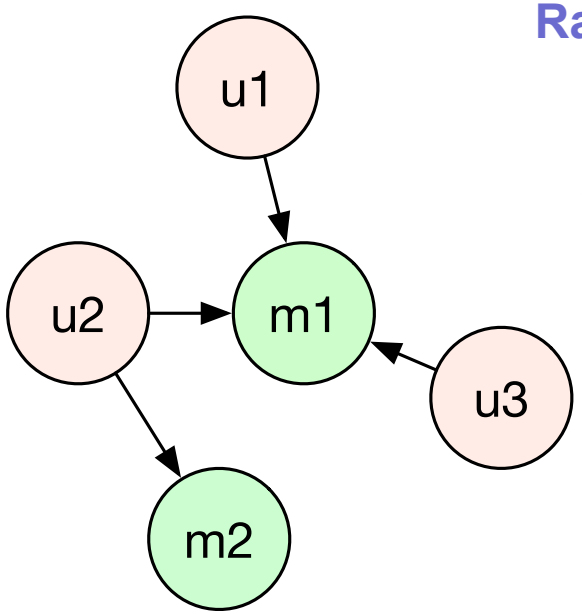
- Category/Brand/Item Repeat Features
  - **Average active days** on given category/category/brand/item of all users
  - **Ratio** of **repeated active users** on given category/brand/item
  - **Maximum active days** on given category/brand/item of all users
  - **Average days** of purchasing the given category/brand/item of all users
  - **Ratio** of users who purchase the given categories/brands/item **more then once**
  - **Maximum days** of purchasing the given category/brand/item of all users
  - **Average span** between two actions of purchasing the given category/brand/item of all users

Embedding Features



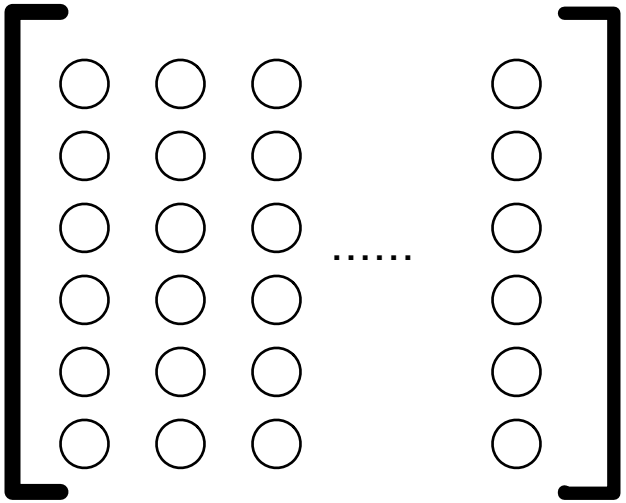
Heterogeneous interaction graph

Embedding Features



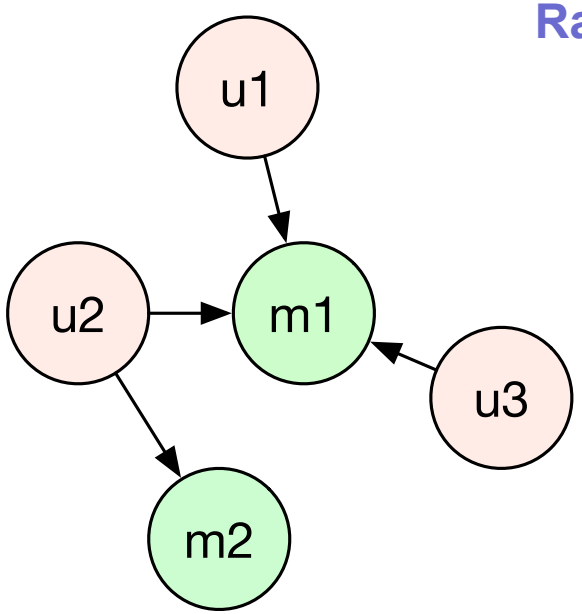
Random walk

W =



Heterogeneous interaction graph

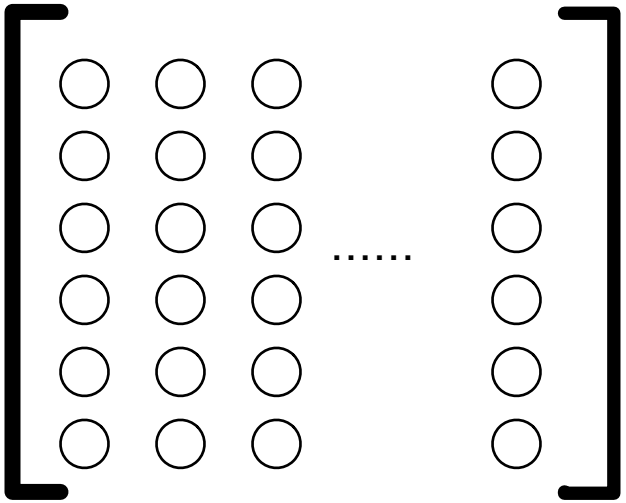
Embedding Features



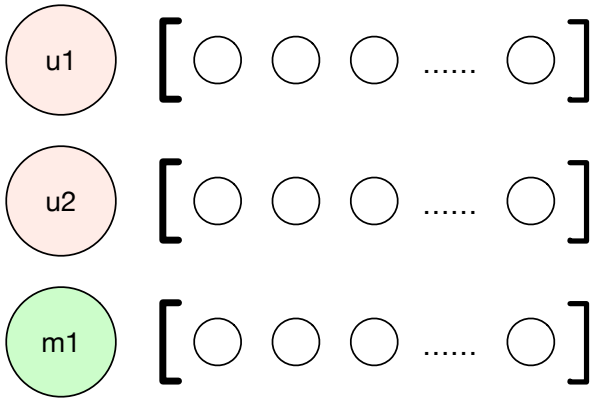
Heterogeneous interaction graph

Random walk

W =

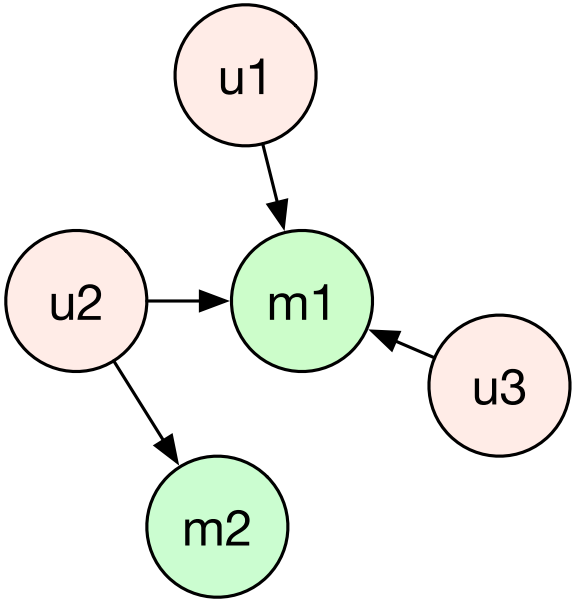


Skipgram model



Embedded vectors

Embedding Features: Interaction Graph

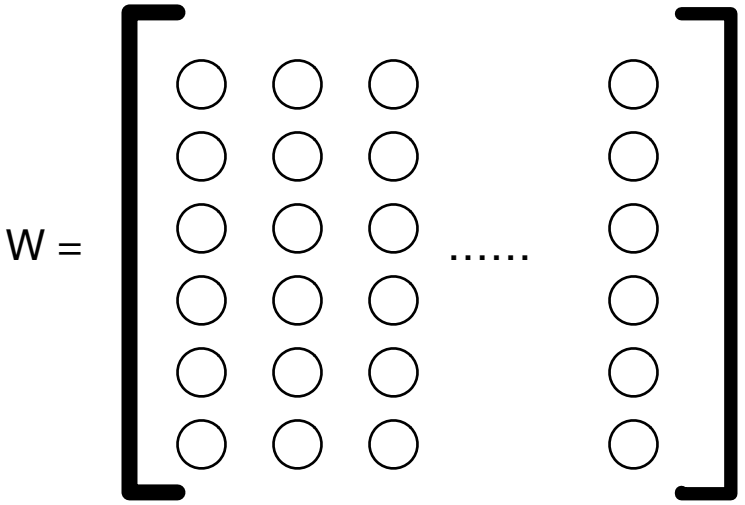


- Let the graph G = (V, E)
  - V is the vertex set
  - E is the edge set

- V contains all users and merchants
- If user u interacts with merchant m, then add an edge <u, m> into E

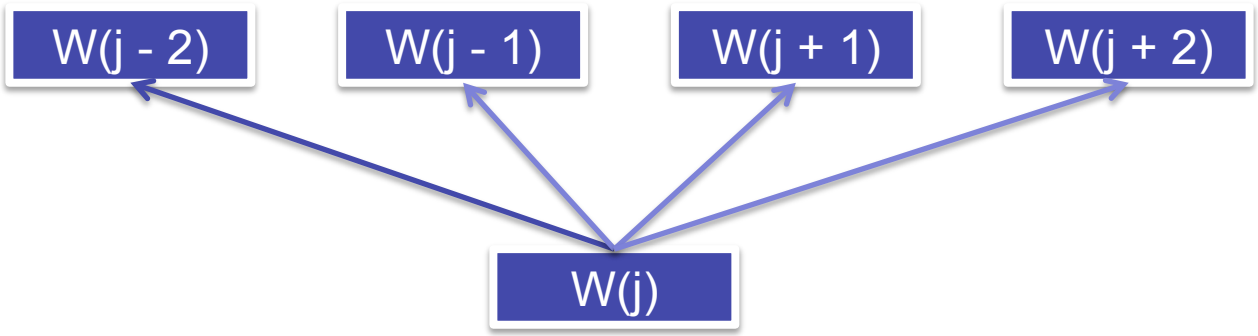
Embedding Features: Random Walk

- Repeat a given number of times
  - For each vertex v in V
    - Generate a sequence of random walk starting from v
    - Append the sequence to the corpus



Generate random walk corpus

Embedding Features: Skipgram



Use the current word W(j) to predict the context.

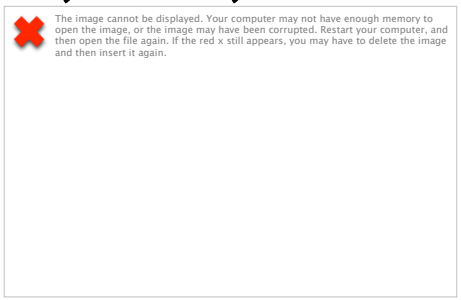
Objective function:

$$L = -\sum_{W \in \mathcal{W}} \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} (f'_{W_{t+j}}{}^{\intercal} f_{W_t} - \sum_{w \in V} f'_w{}^{\intercal} f_{W_t})$$

Use SGD to optimize the above objective and obtain embeddings for users and merchants.

Embedding Features: Dot Products

- Now we have embeddings of all users and merchants.
- Given a pair <u, m>, we derive a feature



- to represent the semantic similarity between u and m.
- f means embeddings.

Embedding Features: Diversification

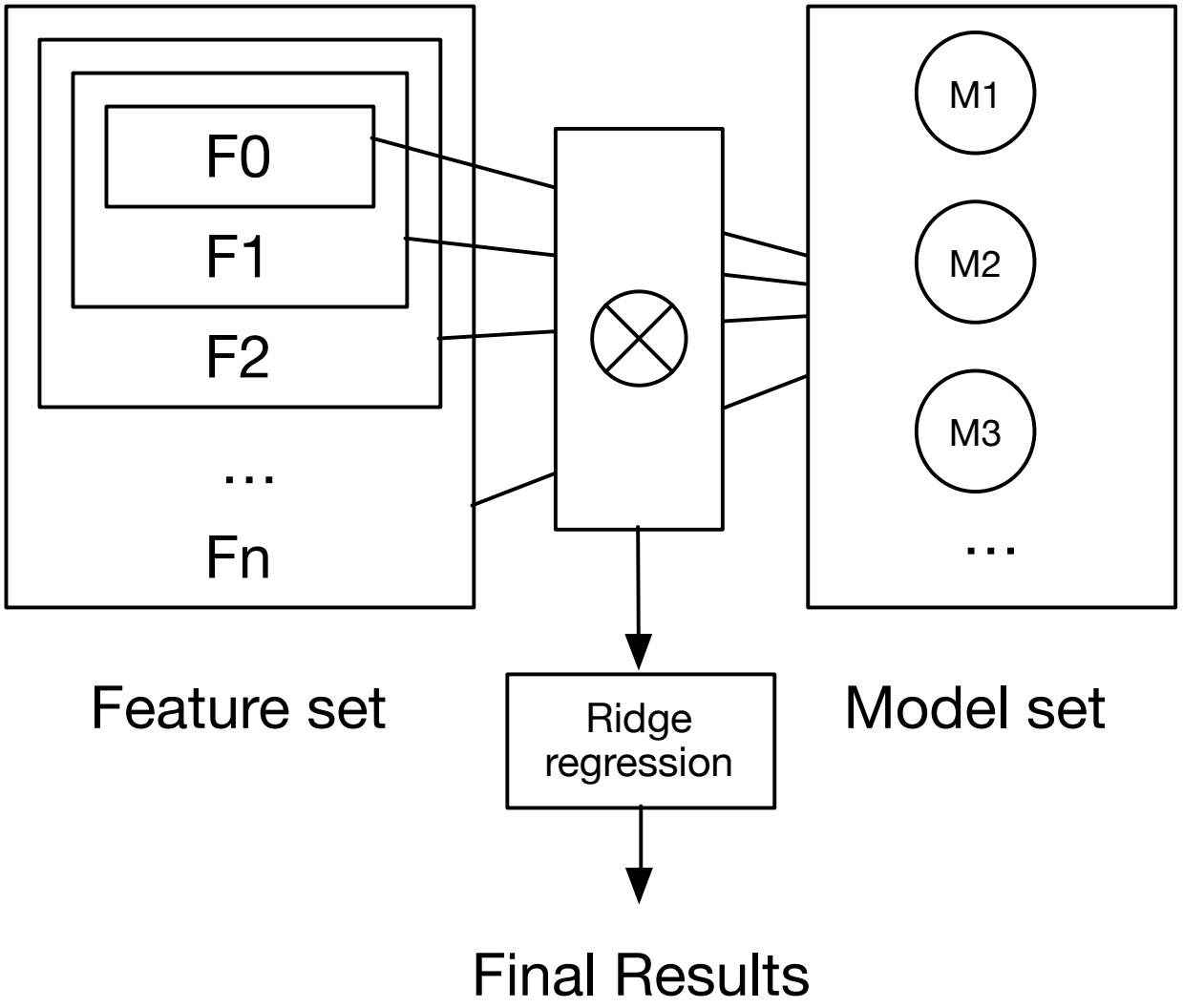
- Simply applying the dot product of embeddings is not powerful enough.
- Recall that we use SGD to learn the embeddings.
- We use embeddings at different iterations of SGD.
- An example
  - Run 100 iterations of SGD.
  - Read out embeddings at iteration 10, 20, …, 100.
  - Obtain a 10-dim feature vector of dot products
- **Intuition: similar to ensemble models with different regularization strengths**

Individual Models

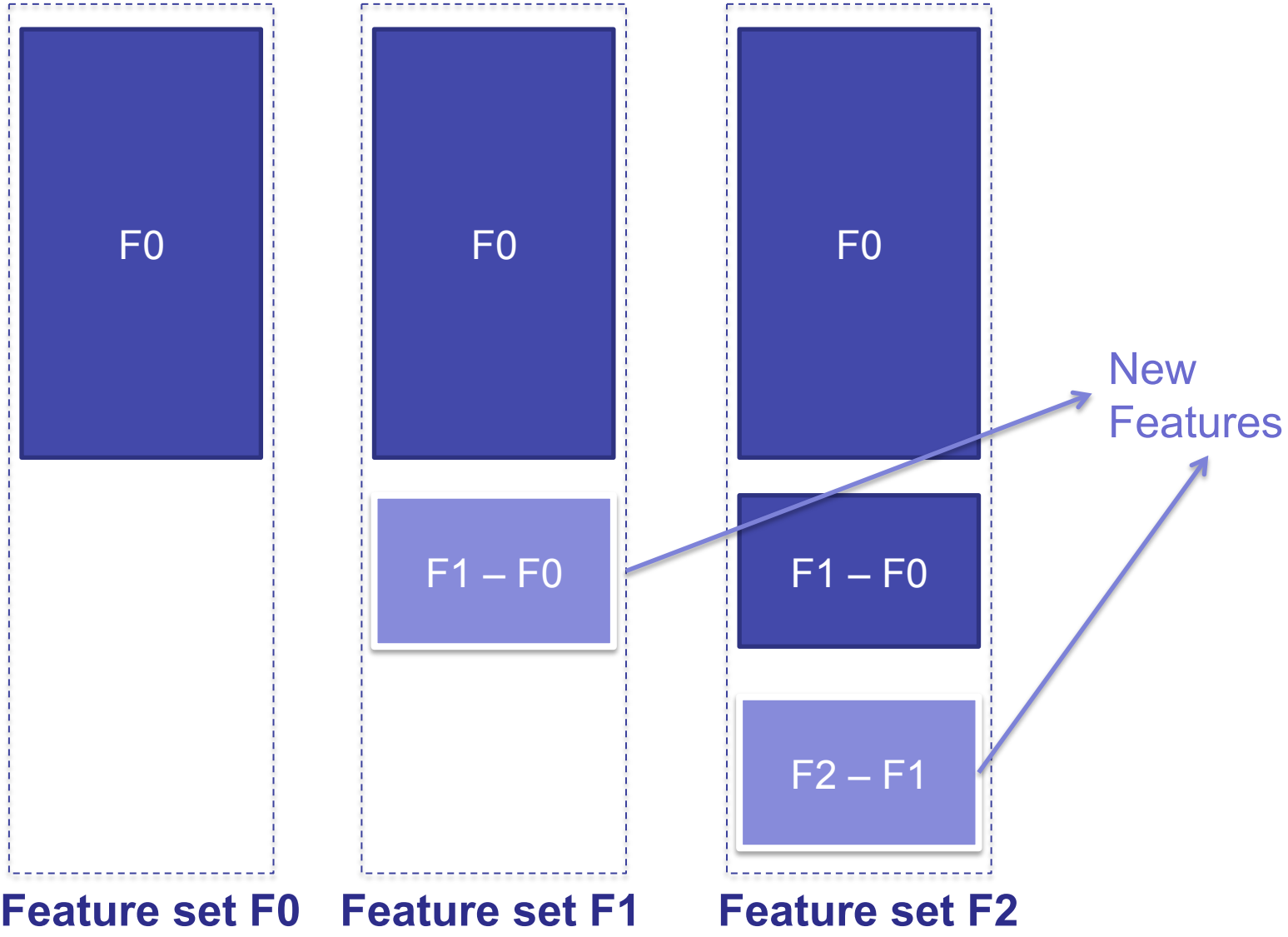
- Logistic regression
  - Use the implementation of Liblinear
- Factorization machine
  - Use the implementation of LibFM
- Gradient boosted decision trees
  - Use the implementation of XGBoost

Method	Implementation	Best AUC in Stage 1 (%)
Logistic Regression	Liblinear	69.782
Factorization Machine	LibFM	69.509
GBDT	XGBoost	69.196

Diversified Ensemble



Diversified Ensemble: Appending New Features



Diversified Ensemble: Cartesian Product

	Model 1	Model 2	Model 3
Feature Set F0	Ensemble 1	Ensemble 2	Ensemble 3
Feature Set F1	Ensemble 4	Ensemble 5	Ensemble 6
Feature Set F2	Ensemble 7	Ensemble 8	Ensemble 9

Diversified Ensemble Results

- Simple ensemble: Only ensemble the top 3 models
- Diversified ensemble outperforms simple ensemble

Method	Implementation	Best AUC in Stage 1 (%)
Logistic Regression	Liblinear	69.782
Factorization Machine	LibFM	69.509
GBDT	XGBoost	69.196
Simple Ensemble	Sklearn Ridge	70.329
Diversified Ensemble	Sklearn Ridge	**70.476**

Factor Contribution Analysis

- Clear performance increase after adding each feature set
- Both embedding features and repeat features provide **unique information** to help the prediction

No.	Method	Stage 1 AUC (%)	Gain
1	Basic features	69.369	-
2	1 + Embedding features	69.495	**0.126**
3	2 + Repeat features	69.782	**0.287**

Stage 2 Performance

- Repeat features are consistent in both stages
- **Data cleaning** is important
  - duplicated/inconsistent records exist in this stage

No.	Method	AUC (%)	Gain
1	Basic features	70.346	-
2	1 + Repeat features	70.589	0.243
3	2 + Data cleaning & more features	70.898	0.309
4	3 + Fine-tuning parameters	71.016	0.118

Summary

- “Tricks” on how to win top 3 in both stages
  - **Diversified ensemble**
  - **Novel embedding features**

Thank you!

Questions ?