SUMMARY :
A TAXI GAP PREDICTION METHOD VIA DOUBLE ENSEMBLE GRADIENT BOOSTING DECISION TREE

Xiao Zhang∗, Xiaorong Wang†, Wei Chen∗, Jie Tao∗, Weijing Huang∗ and Tengjiao Wang∗
∗Key Laboratory of High Confidence Software Technologies (MOE), School of EECS, Peking University,
Beijing, 100871, China
{xiao.zhang, pekingchenwei}@pku.edu.cn
†Technology & Strategy Research Center, China Electric Power Research Institute,
Beijing, 100192, China

Amina Benzerga

OVERVIEW

Paper goal :

Predicting the GAP between taxi demand and supply in taxi booking app.

Gap ?

predict the number of passengers who launched orders in taxi booking app but have not received responses from any drivers.

WHAT MAKES THIS PAPER DIFFERENT FROM THE OTHERS ?

- Computes the **concrete value** of gap
- Takes into account other **features** s.a. : weather info, traffic condition the points of interest (POI),…
- Proposing a new **feature selection** method to exploiting appropriate features
- **Double Ensemble** idea to solve **dataset sparsity** and **missing values** in data mining task (Do not ignore missing data)
- Offering **large-scale** taxi **data to verify** assumption of taxi gap prediction approaches

TASK DESCRIPTION - 1

District ID :

Split a city into **n non-overlapping** square area D = {d1, d2, ..., dn}

District d as the index of these square areas.

Time Slot :

Partition one day into **144 continuous time** slots, each time slot represents 10 minutes.

Time slot t as the sequence of all 10-minute time slots in one day.

TASK DESCRIPTION - 2

Demand-Supply Gap :

For the given date time k (d_t = k), district id i (d = i) and time slot j (t = j),

$demand_{i,j,k}$ is the number of passengers' orders launched at district id i, in time slot j of the date time k

$supply_{i,j,k}$ is the number of taxi drivers who accepted the orders generated at district id i and in the time slot j of the date time k.

$$Gap_{i,j,k} = \begin{cases} demand_{i,j,k} - supply_{i,j,k} & \text{When } demand_{i,j,k} > supply_{i,j,k} \\ 0 & \text{Otherwise} \end{cases}$$

PROBLEM FORMULATION

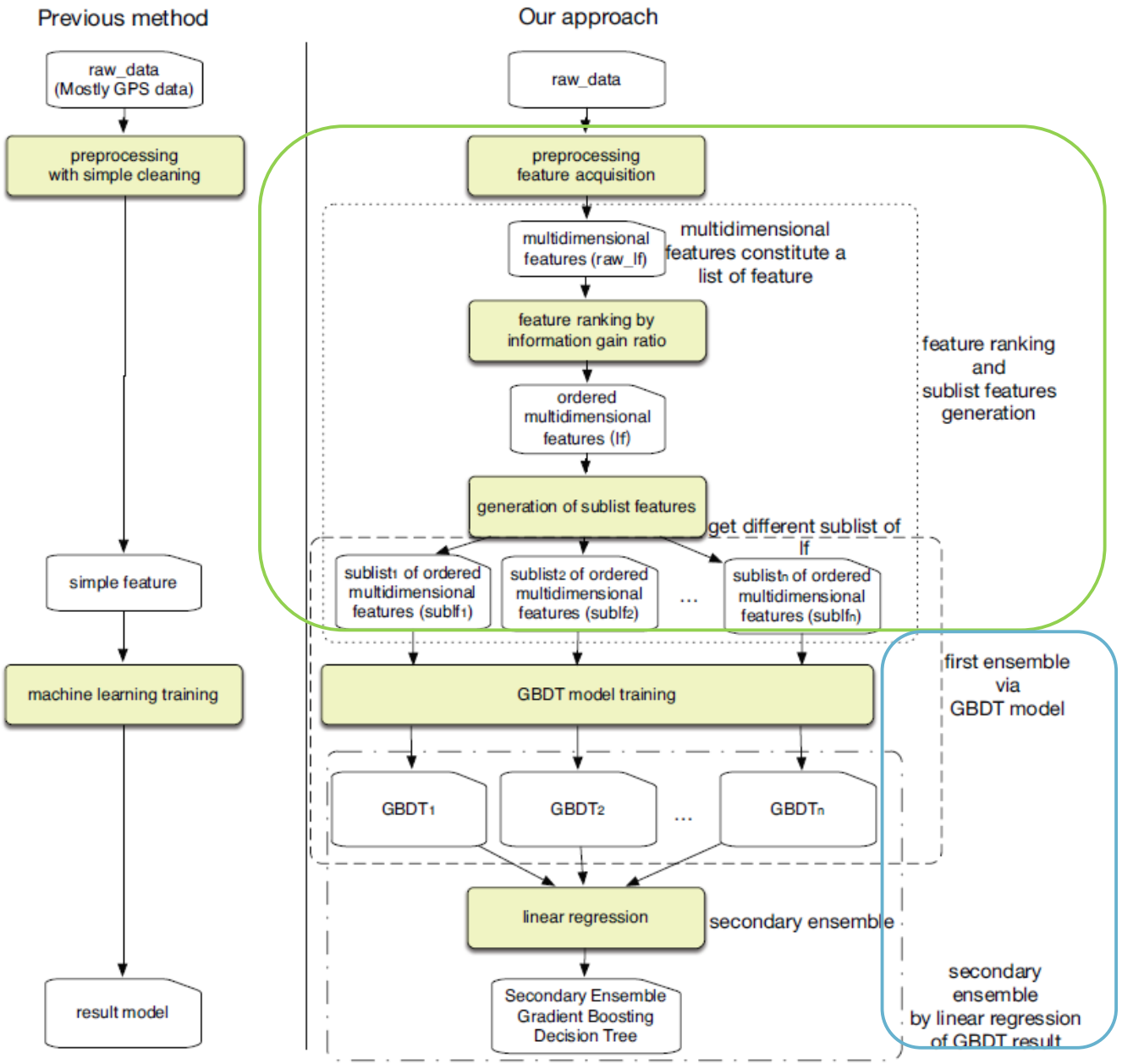
Training model X from given raw data which would be introduced at competition website.

X input : date time k dt = k,

time slot j t = j

district id i d = i

X output : one real number : gap_ijk.



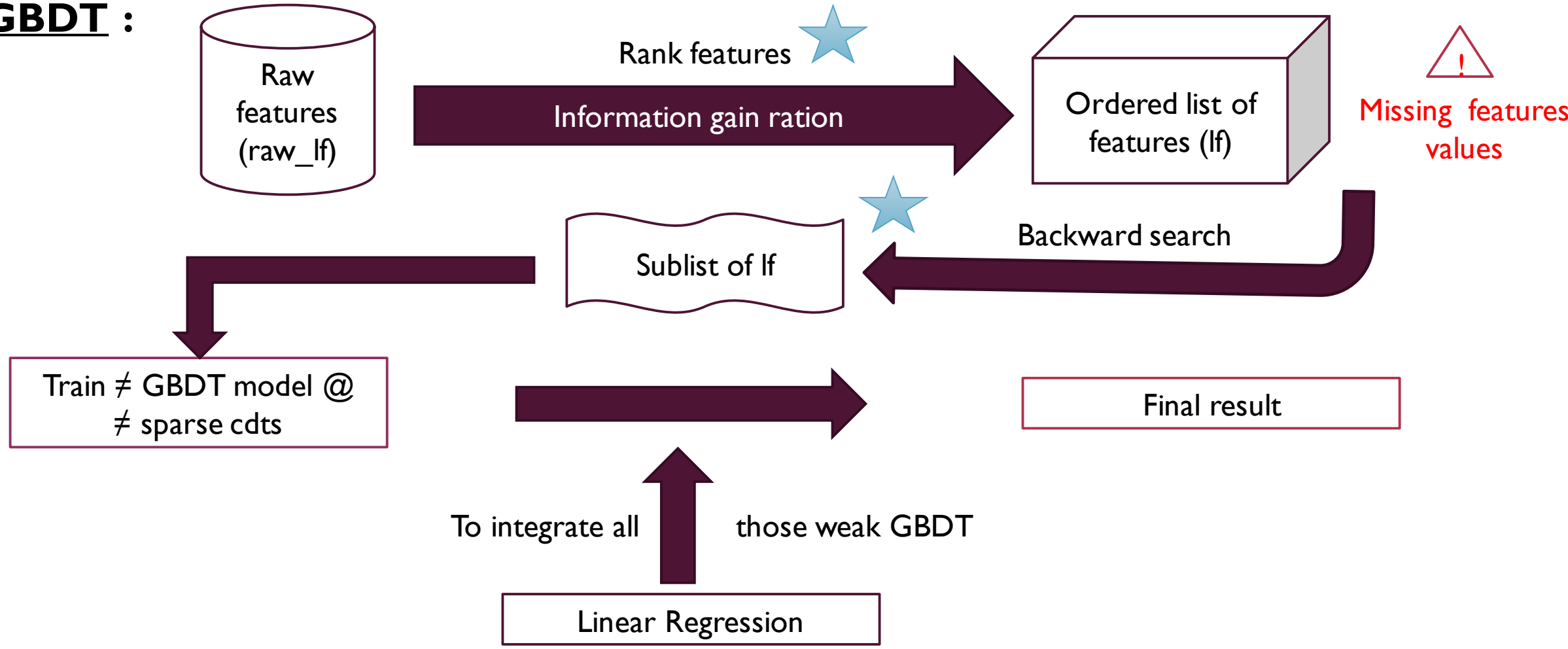
METHODOLOGY

1. Feature engineering before model's contruction

2. Stronger double ensemble technique to deal with missing data

METHODOLOGY - FRAMEWORK OF THEIR APPROACH

DEGBDT :



METHODOLOGY - FEATURE RANKING BY INFORMATION GAIN

1. **Split continous features** into three parts according to the uniform distribution
2. Keep discrete features unchanged.
3. **Calculate every feature's rate of information gain**[17] to rank features (helps determine which feature is effective)
4. **Sort the raw If** according to the information gain rate to get the If (the information gain ratio is considered as sorting indices from top to bottom)

METHODOLOGY - SUBLIST GENERATION BY GREEDY BACKWARD SEARCH & DOUBLE ENSEMBLE GRADIENT BOOSTING DECISION TREE

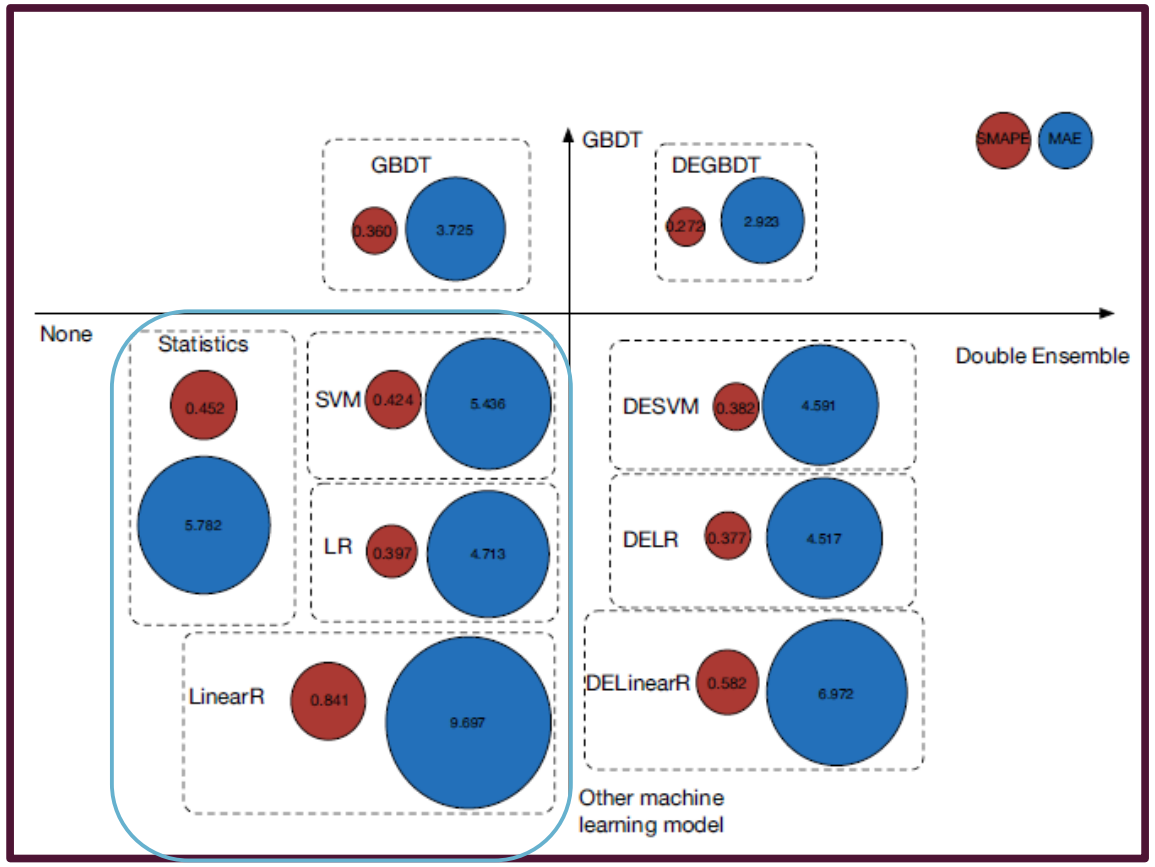
Sublist Generation By Greedy Backward Search :

After ranking features, We get **different effective features' combination**, called sublf which is **sublist of lf,** to train different GBDT model in this subsection.

Double Ensemble Gradient Boosting Decision Tree :

GBDT is a **gradient boosting algorithm** that utilizes **decision stumps** or **regression trees** as weak classifiers. In DEGBDT, the weak learners **measure the error observed in each node**, **split the node using a test function** $\kappa: R^n \rightarrow R$ with a threshold τ , and return values η_l and η_r. The **optimal split** : triplet (τ , η_l, η_r) to **minimize** the **error after split**.

EXPERIMENT



Lower SMAPE and MAE mean better method.

Symmetric mean absolute percentage error(sMAPE) : The limitation of sMAPE is that if the actual gap or forecast gap is 0, the value of error will boom up to the upper-limit of error.

Mean Absolute Error (MAE)

Basic method : neither using GBDT nor using weak classifier ensemble. It contains LR, SVM, linear regression, and statistical average (as the baseline). As expected, linear regression was significantly lower than baseline because of non-direct linear correlation. And LR and SVM were almost identical.

Ensemble learning is superior to simple machine learning in this problem that all the base models have been improved after using the DE method.

To summarize, the **Double Ensemble method** has a **better** performance than the general method in solving the **problem of data sparsity** when the feature quantity satisfies a certain number.