IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 34, No. 1, February, 2020, pp. 111-116
Journal homepage: http://iieta.org/journals/ria

Supply-Demand Prediction of DiDi Based on Points of Interests Selection in Extreme Gradient Boosting Algorithm

Yonghong Tian[1*], Zeyu Li[2], Yue Zhang[1], Qi Wu[1]

[1] College of Data Science and Application, Inner Mongolia University of Technology, Hohhot 010080, China
[2] School of Computer Science and Technology, Xidian University, Xi'an 710071, China

Corresponding Author Email: tyh@imut.edu.cn

https://doi.org/10.18280/ria.340115

Received: 5 August 2019
Accepted: 10 November 2019

Keywords:
sharing economy, point of interest (POI), supply-demand prediction, extreme gradient boosting (XGBoost)

ABSTRACT

In recent years, DiDi, an online car-hailing (OCH) service provider, has emerged as a leader in the sharing economy. To improve user experience, the company must minimize the waiting time and optimize car utilization based on accurate estimation of supply-demand gap. This paper aims to develop a desirable model to select the most significant factors for OCH supply-demand estimation. Firstly, the correlation between the points of interest (POIs) and the supply-demand gap was proved through statistical analysis. Next, the number and type of POIs were found to have a slight impact on the estimation results. On this basis, the authors put forward a method called POI principal component extraction based on supply-demand gap (PPCE-SDG) to select the most significant POIs. The PPCE-SDG involves four steps: k-means clustering (KMC) of blocks based on supply-demand gap; creating a data vector of POIs after counting the POIs in each cluster; extracting the significant POIs through principal component analysis (PCA) of the data vector; importing the extracted POIs to extreme gradient boosting (XGBoost) for OCH supply-demand prediction. Finally, the POIs selected by the PPCE-SDG were proved superior than those collected by other methods in OCH supply-demand estimation, indicating that our model is a desirable tool for significant POIs selection. The research results lay a good basis for the optimization of OCH services.

1. INTRODUCTION

Recent years has seen an explosion of services that facilitate sharing economy. Take the field of transport for example. Leading online car-hailing (OCH) service providers (e.g. Uber, DiDi and Lyft) have developed popular mobile apps to provide on-demand transport services. Compared with traditional transport means like metro and buses, the OCH offers a convenient and flexible travel mode for passengers. The OCH apps incentivize private car owners who agree to provide car-hailing services, which promotes the sharing economy and enlarges urban transport capacity.

Currently, a large number of OCH orders are generated every day, fulfilling the travel demand of millions of passengers. The frequent uses of OCH services expose two problems: For one thing, some drivers find it hard to receive any order, because few people nearby use car-hailing apps; for another, it is extremely difficult to get a ride in bad weather or rush hours, due to the short supply in the surrounding areas. To solve the problems, OCH service providers must schedule the drivers reasonably to minimize the waiting time of passengers and maximize the use of drivers. The effectiveness of scheduling hinges on supply-demand estimation.

The OCH supply-demand is influenced by various factors, including but not limited to weather, time and traffic condition [1]. The analysis of these factors is the first step to setting up an estimation model of OCH supply-demand. To collect the necessary data for the analysis, the points of interest (POIs) in each block of the target city should be observed on a daily basis. After all, the POIs are the common destinations of passenger flows in urban areas. However, the POIs, being numerical identifiers, offer a limited amount of information. It is an arduous task to select those suitable for supply-demand estimation out of the various POIs.

To select the few significant POIs, this paper proposes a method called POI principal component extraction based on supply-demand gap (PPCE-SDG). Firstly, the k-means clustering (KMC) was adopted to cluster the blocks in the target city based on the supply-demand gap. Next, the POIs in each cluster were counted, and a data vector was obtained for the POIs. Then, the data vector was subjected to principal component analysis (PCA) to extract the POIs. Finally, the effectiveness of the PPCE-SDG in OCH supply-demand estimation was verified through experiments under the end-to-end framework of extreme gradient boosting (XGBoost) [2], using the public dataset released by DiDi. The dataset, as the result of Di-Tech Algorithm Challenge, contains 11,467,117 OCH orders over 7 weeks across 58 blocks in Hangzhou, China. The dataset was supplemented by a number of other features like the POI distribution in each block, weather (temperature and PM2.5), and the number of segments under each congestion level in each block.

The main contributions of our research are about the influence of POIs over OCH supply-demand:

(1) The POIs are correlated with OCH supply-demand. The POI-based models are more accurate than the models without POIs in the estimation of OCH supply-demand. Two blocks with similar POIs tend to bear high resemblance in OCH

supply-demand.

(2) Different types of POIs exert different impacts on OCH supply-demand. The estimation accuracy varies with the types of POIs, even if the number of POIs remains constant. With the rising number of POIs in the same type, the estimation accuracy first increases and then declines. Hence, the POIs must be selected carefully to estimate OCH supply-demand accurately.

(3) A few POIs play decisive roles in the estimation of OCH supply-demand. The influence of all POIs largely comes from the primary POIs extracted through the PCA. Besides, 90% of the improvement in estimation accuracy is attributable to the four most significant POIs.

(4) After clarifying the influence of POIs over OCH supply-demand, it is possible to predict the exact number of passengers in need of OCH services and the number of drivers available in each block, and strike a balance between supply and demand in advance through car dispatching, price adjustment and pick-up location recommendation.

The remainder of this paper is organized as follows: Section 2 reviews the previous studies on OCH supply-demand and POIs; Section 3 describes the research dataset and introduces the verification framework; Section 4 analyzes the impacts of POIs on OCH supply-demand, and selects the most significant POIs; Section 5 sets up the PPCE-SDG model; Section 6 verifies the proposed model through experiments; Section 7 puts forward the conclusions.

2. LITERATURE REVIEW

The existing studies on OCH supply-demand mostly focus on a specific aspect of the problem. For instance, Saadi et al. [3] compared several machine learning approaches in the prediction of OCH supply-demand. Chang et al. [4] and Yan et al. [5] forecasted the hotspots of OCH service users. Phithakkitnukoon et al. [6] identified and projected the positions of vacant taxis on the OCH platform.

There are many factors that affect the OCH supply-demand, such as POIs, weather and traffic condition. The OCH demand is high in a block with many POIs (e.g. malls and restaurants); more people will resort to OCH services in bad weather; traffic jam dampens the interest in OCH services. Many scholars [7-10] have designed excellent path planning models that recommend the best itinerary to drivers, yet failed to assess the contribution of POIs to OCH supply-demand. Chen et al. [2] noticed the relationship between POIs and supply-demand gap, but did not clearly demonstrate the relationship.

The current estimation methods for OCH supply-demand are bottlenecked by the lack of information or the absence of the POIs. As a location information, the POIs are an essential cause of the gap between supply and demand. The OCH demand may be affected differently by POIs from different categories: malls have a positive impact on the demand, while congestion exerts a negative impact. Therefore, this paper probes deep into the impacts of the POIs on OCH supply-demand.

3. PRELIMINARIES

3.1 Dataset description

Our dataset was released by DiDi after the Di-Tech Algorithm Challenge. The dataset contains 11,467,117 OCH orders over 7 weeks across 58 blocks in Hangzhou, China, and was supplemented by features like the POI distribution in each block, weather (temperature and PM2.5), and the number of segments under each congestion level in each block. The weather and traffic conditions could be directly imported to the XGBoost, but the semi-structured POIs must be pre-processed before use.

There are 173 types of POIs in the dataset. But it is unclear which POI types belong to which block. To make matters worse, the distribution and number of POIs differ from block to block. Here, the 173 types are numbered from 1# to 173#, respectively, and the number and types of POIs were counted for each block.

The data on the 24 days from February 23rd to March 17th were allocated to the training set. For each block, one training sample was generated every 10min from 0:00 to 24:00 on each of the 24 days, that is, each day was divided into 144 slices. Thus, a total of 200,448 training sample were obtained. The data on the 28 days from March 25th to April 31st were allocated to the test set. For each block, one test sample was generated every 10min from 0:00 to 24:00 on each of the 28 days. Thus, the test set contains a total of 58,464 test samples.

3.2 Verification framework

The XGBoost, an integrated learning framework good at classification and regression, was employed to verify our model. This machine learning (ML) algorithm was improved from gradient boosted decision tree (GBDT) [11]. As the name suggests, the GBDT integrates the decision tree and gradient boosting. The XGBoost selects the optimal decision tree by scoring the tree structure and leaf nodes. Unlike the GBDT, the XGBoost optimizes the loss through second-order Taylor expansion, and prevents overfitting with additional regularization terms. With high efficiency and good ability of parallel processing, the XGBoost is very suitable to handle big data problems. The pseudocode of the XGBoost is given below.

Algorithm: XGBoost
Inputs: I, instance set of the current node
d, feature dimension
gain $\leftarrow 0$
$G \leftarrow \sum_{i \in I} g_i, H \leftarrow \sum_{i \in I} h_i$
for k=1 to m do
$G_L \leftarrow 0, H_L \leftarrow 0$
for j *in sorted*$(I, by\ x_{jk})$do
$G_L \leftarrow G_L + g_j, H_L \leftarrow H_L + h_j$
$G_R \leftarrow G - G_L, H_R \leftarrow H - H_L$
$score \leftarrow \max(score, \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda})$
end
end
Output: Split with max. score

4. POIS ANALYSIS

4.1 Impacts of POIs on OCH supply-demand

Observations show that the OCH supply-demand is directly affected by time, weather and traffic condition. In general, it is difficult to receive OCH services in the rush hours on workdays, during the holidays or under bad weather. Meanwhile, fewer passengers have OCH demand under congestion and good weather. Nonetheless, the estimation of OCH supply-demand often ignores an important factor: the POIs.

Here, a set of five POIs and another set of ten POIs are established, and applied to estimate the OCH supply-demand. The estimation effects were compared with that of a dataset with zero POI. As shown in Table 1, the estimation based on POI sets were much better than that based on the dataset without POI, as measured by minimum, mean, maximum and standard deviation. This means the addition of POIs benefits the estimation of OCH supply-demand.

Table 1. Estimation effects with different number of POIs

Number of POIs / Metric	0	5	10
Minimum	0.59032	0.63359	0.63549
Mean	0.58839	0.63654	0.63787
Maximum	0.59261	0.63932	0.64019
Standard deviation	0.00624	0.00287	0.00235

To further disclose the impacts of POIs on OCH supply-demand, the mean supply-demand gap of each 10min in each block was calculated based on the training set, creating a 114-dimensional data vector of the supply-demand gap. Each dimension corresponds to one of the 144 time slices of each day. Another 114-dimensional data vector was created to reflect the variation in the supply-demand gap between workdays and holidays. The two vectors were combined into a 288-dimensional data vector.

Next, the KMC was introduced to cluster the data in 288 dimensions. In this way, the 58 blocks were allocated to five clusters, denoted as 0, 1, 2, 3 and 4, in turn. The five cluster centers of supply-demand gap were confirmed by the k-means-plus-plus clustering (KM++C). Extended from the KMC [12, 13], the KM++C overcomes an inherent defect of the KMC (the clustering effect heavily depends on the initial cluster centers, because the similarity is measured by Euclidean distance), and enhances the inter-cluster difference. The clustering results show that the 5 samples at cluster centers are highly representative, and the five clusters distinguish the 58 blocks well.

After that, the authors analyzed the intra- and inter-cluster similarities of POIs. The divergence of POIs between clusters is mainly reflected in quantity. Thus, the intra- and inter-cluster similarities were both measured by Euclidean distance. The Euclidean distance is negatively correlated with the two similarities. The Euclidean distances of clusters 0-3 were computed, for clusters 1 and 4 contain too few samples. The calculated results are listed in Table 2, where the numbers are the proportions of POIs falling in the range of Euclidean distance. Obviously, the intra-cluster similarities of POIs were all greater than the inter-cluster similarities, indicating that the POIs are closely correlated with OCH supply-demand gap.

Table 2. Intra- and inter-cluster similarities

Euclidean distance / Cluster	0.000001	0.0000015	0.000002	0.0000025	0.000003
Intra-cluster similarities	0	0.0085	0.1974	0.6131	0.8895
Cluster 0	0	0	0	0.4667	0.83
Cluster 2	0	0	0	1	1
Cluster 3					
Inter-cluster similarities	0	0	0.1354	0.4323	0.7813
Clusters 0, 2	0	0	0.0625	0.4028	0.8611
Clusters 0, 3	0	0	0	0.25	0.75
Clusters 2, 3	0.000001	0.0000015	0.000002	0.0000025	0.000003

4.2 Contributions of POIs to OCH supply-demand estimation

This subsection mainly explores how the number and type of POIs contribute to the estimation of OCH supply-demand. For this purpose, four sets of POIs were constructed, each of which contains 8 subsets. The number of POIs in the 8 subsets is 5, 10, 15, 20, 30, 50, 75 and 100, respectively. The four sets have different combinations of POI types. Then, the four sets of POIs were separately imported to the XGBoost for estimation of OCH supply-demand. The estimation accuracies are compared in Figure 1.

As shown in Figure 1, the estimation accuracy generally improved with the growing number of POIs. However, a large number of POIs did not necessarily lead to a high estimation accuracy. For example, the highest accuracy was not obtained in the case of 100 POIs. Besides, the increase in the number of POIs prolonged the estimation duration. Therefore, the number of POIs is not a decisive factor of OCH supply-demand estimation.

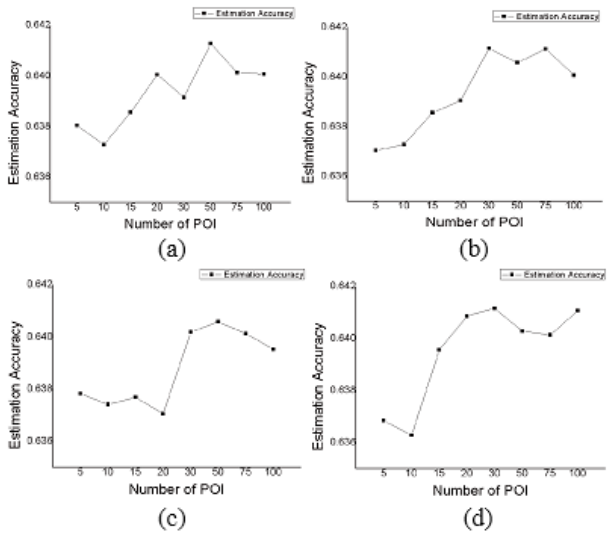

Figure 1. Estimation accuracies of different sets of POIs

From Figures 1(a), (b), (c) and (d), it can be seen that the estimation accuracy varied slightly between different combinations of POI types, when the number of POI were the same. This means the type of POIs is not a decisive factor of OCH supply-demand estimation, too.

Overall, the estimation of OCH supply-demand is greatly affected by the combined effects of the number and type of POIs. This calls for the selection of a few significant POIs before estimating OCH supply-demand.

5. MODEL CONSTRUCTION

This sections designs the PPCE-SDG model to extract the significant POIs from the 173 types of POIs distributed in the 58 blocks. After the blocks were clustered by the KMC, the PCA was performed to calculate the absolute eigenvalue of each POI. The PCA is a popular way to reduce data dimensionality. It transforms the original data into a set of linearly independent representations. Next, the significant POIs were selected based on the absolute eigenvalues. The PPCE-SDG is a data-centered logical model that makes strict calculation and derivation in each step. The model can be seamlessly coupled with XGBoost. The pseudocode of the PPCE-SDG is given in Algorithm 1.

Algorithm 1. PPCE-SDG

Inputs:
G: the set of supply-demand gaps
P: the set of POIs
B: the set of blocks
T: the set of training days
Output:
The selected POIs
1: Start:
2: for each block b_i in B do
3: for each time slice t_j in T do
4: b_i. calculate ($\frac{1}{T}\sum(gap_{i,j})$) //calculate the mean gap in the same time slice of T
5: end for
6: end for
7: gap vector = statistic (B. calculate ($\frac{\sum G}{T}$)) //integrate the mean gap into a gap vector
8: center points = k-means+(gap vector) //use KM++C to confirm center points
9: clusters = KMC (center points, B) //use KMC to cluster the blocks
10: for each cluster c_k in clusters do
11: for each POI p_m in c_k do
12: c_k. calculate ($\frac{1}{B_k}\sum(p_m)$) //calculate the mean of a type POI in each cluster
13: end for
14: end for
15: POI vector=statistic (cluster. Calculate (PB$\frac{\sum G}{B}$)) // integrate the mean POI into a POI vector
16: POI eigenvalue=PCA (POI vector) // use PCA to calculate the eigenvalues of each POI type
17: sort <- Max (POI eigenvalues)
18: return selected POIs
19: End

6. EXPERIMENTAL VERIFICATION

To verify its effectiveness, the proposed PPCE-SDG was applied to predict the OCH supply-demand with POIs of different contributions, and then compared with XGBoost feature selection method (XFSM) and GBDT feature selection method (GFSM) through experiments on the above-mentioned dataset. Each experiment was carried out 50 times and the mean of the ten best results were adopted for analysis.

Before the experiments, the supply-demand gap was defined as follows: On the d-th day, the supply-demand gap in the interval [t, t+C) of a block equals the total number of invalid orders in the interval. Here, the constant C is set to 10, and the gap is denoted as $\mathrm{gap}_a^{d,t}$.

6.1 Performance metrics

The performance of the PPCE-SDG, the XFSM and the GFSM was measured by four metrics, namely, the mean absolute error (MAE), the root mean squared error (RMSE), the accuracy and the F_1. The four metrics can be respectively calculated by:

$$MAE = \frac{1}{|T|}\sum_{(a,d,t\in T)}\left|gap_a^{d,t} - pred_a^{d,t}\right| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{|T|}\sum_{(a,d,t\in T)}\left|gap_a^{d,t} - pred_a^{d,t}\right|} \quad (2)$$

$$Accuracy = \frac{H}{N} \quad (3)$$

where, H is the hit number; N is the total number.

$$F_1 = \frac{2 * \mathrm{Pr}ecision}{\mathrm{Pr}ecision + \mathrm{Re}call} \quad (4)$$

where, Precision is the degree of prediction accuracy; Recall is the measure of completeness.

6.2 Results analysis

To measure its effectiveness, the PPCE-SDG was adopted for POI selection from three sets of POIs on three contribution levels: maximum contribution (Max), medium contribution (Med) and minimal contribution (Min). The top 10 POI that contribute the most to OCH supply-demand estimation were selected and imported to XGBoost for prediction. Figure 2 compares the estimation accuracies under the three POI sets.

As shown in Figure 2, the estimation accuracy was much higher under the Max set than that under the Med set, which was in turn far greater than that under the Min set. The results prove that the PPCE-SDG can extract the most significant POIs. However, once the number of POIs reached four, the estimation accuracy was basically stable, indicating that a few POIs make most of the contribution. In fact, a small number of POIs means a high efficiency of estimation. Hence, the above results fully demonstrate the effectiveness of the PPCE-SDG in POI selection.

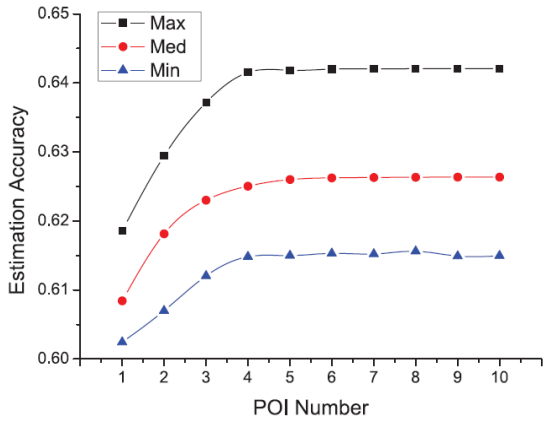

Figure 2. The estimation accuracies under the three POI sets

Next, the PPCE-SDG, XFSM and GFSM were separately applied to select the four POIs that contribute the most to OCH supply-demand estimation. The parameters of XFSM and GFSM were fine-tuned through grid search. For comparison, four POIs (Random) were randomly selected from our dataset. The selected POIs were imported in turn to XGBoost for OCH supply-demand estimation. The estimation effects under the four sets of POIs are compared in Table 3.

Table 3. The estimation effects under the four sets of POIs

POI sets	Performance metrics		
	MAE	RMSE	F_1
XFSM	3.47	14.88	0.67
GFSM	3.62	15.71	0.66
PPCE-SGD	3.41	14.33	0.68
Random	3.64	15.82	0.66

As shown in Table 3, the POIs identified by the PPCE-SDG achieved the smallest MAE and RMSE, and the highest F_1 among the four sets of POIs. The results manifest that the PPCE-SDG is superior in selecting the suitable POIs for the estimation of OCH supply-demand. The superiority of our model over XFSM and GFSM is attributable to the fact that our model can extract a few significant POIs and import them directly to the XGBoost, while the latter two need to calculate the information gain of each POI.

7. CONCLUSIONS

This paper puts forward an effective model called PPCE-SDG to select the significant POIs for the estimation of OCH supply-demand. Firstly, the correlation between the POIs and the OCH supply-demand gap was proved through statistical analysis. Next, the number and type of POIs were found to have a slight impact on the estimation results. On this basis, the PPCE-SDG was explained in details: the KMC was adopted to cluster the blocks in the target city based on supply-demand gap; the POIs in each cluster were counted, creating a data vector of POIs; the PCA was performed on the data vector to extract the POIs; the POIs were imported to the XGBoost to predict OCH supply-demand. Finally, the POIs selected by the PPCE-SDG were proved superior than those collected by other methods in OCH supply-demand estimation. The research results lay a good basis for the optimization of OCH services.

ACKNOWLEDGMENTS

The work was supported by the Natural Science Foundation of Inner Mongolia (Grant No.: 2013MS0920), and Science and Technology Planning Project of Inner Mongolia (Grant No.: 201502015).

REFERENCES

[1] Chen, T.Q., Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Cornell University, 785-794. https://doi.org/10.1145/2939672.2939785

[2] Chen, L., Mislove, A., Wilson, C. (2015). Peeking beneath the hood of uber. In ACM Conference on Internet Measurement Conference, pp. 495-508. https://doi.org/10.1145/2815675.2815681

[3] Saadi, I., Wong, M., Farooq, B., Teller, J., Cools, M. (2017). Aninvestigation into machine learning approaches for forecasting spatio-temporal demandin ride-hailing service. Cornell University, 2017.

[4] Chang, H.W., Tai, Y.C., Chen, H.W., Hsu, J.Y. (2008). iTaxi: Context-aware taxi demand hotspots prediction using ontology and data miningapproaches. Semantic Scholar, 1-8.

[5] Yan, Q., Wang, H.Y., Wang, H. (2015). Prediction of taxi passenger volumebasedon a gray linear regression combined model. In International Conference on Transportation Engineering, Dailan, China, pp. 335-341. https://doi.org/10.1061/9780784479384.042

[6] Phithakkitnukoon, S., Veloso, M., Bento, C.L., Biderman, A., Ratti, C. (2010). Taxi-aware map: identifying and predicting vacant taxis in the city. Ambient Intelligence: First International Joint Conference AmI 2010, Malaga, Spain, Springer Berlin Heidelberg, pp. 86-95. https://doi.org/10.1007/978-3-642-16917-5_9

[7] Zhang, X., Wang, X.R., Chen, W., Tao, J., Huang, W.J., Wang, T.J. (2017). A taxi gap prediction method via double ensemble gradient boosting decision tree. In IEEE International Conference on Big Data Security on Cloud, pp. 255-260. https://doi.org/10.1109/BigDataSecurity.2017.27

[8] Wang, D., Cao, W., Li, J., Ye, J.P. (2017). Deepsd: Supply-demand prediction-for online car-hailing services using deep neural networks. In IEEE International Conference on Data Engineering, San Diego, CA, USA, pp. 19-22. https://doi.org/10.1109/ICDE.2017.83

[9] Chen, X.Q., Zahiri, M., Zhang, S.C. (2017). Understanding ridesplitting be-havior of on-demand ride services: An ensemble learning approach. Transportation Research Part C Emerging Technologies, 76: 51-70. https://doi.org/10.1016/j.trc.2016.12.018

[10] Tong, Y.X., Chen, Y.Q., Zhou, Z.M. (2007). The simpler the better: A unified approach to predictingoriginal taxi demands based on large-scale online platforms. ACM SIGKDD In-ternational Conference on Knowledge Discovery and Data Mining ACM, pp. 1653-1662. https://doi.org/10.1145/3097983.3098018

[11] Ye, J., Chow, J.H., Chen, J., Zheng, Z.H. (2009). Stochastic gradient boosted distributed decision trees. ACM Conference on Information & Knowledge Management, pp. 2061-2064.

[12] Bahmani, B., Moseley, B., Vattani, A., Kumar, R., Vassilvitskii, S. (2012). Scalable k-means. Proceedings of the VLDB Endowment, 5(7): 622-633. https://doi.org/10.14778/2180912.2180915

[13] Arthur, D., Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding. Eighteenth ACM-Siam Symposium on Discrete Algorithms. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January, pp. 1027-1035. https://doi.org/10.1145/1283383.1283494