Use Rate Prediction For Charging Stations

Jiaxiang Huang, Junjie Yao
School of Computer Science and Software Engineering
East China Normal University
Shanghai, P.R.China
xxxx@xxx, junjie.yao@sei.ecnu.edu.cn

3rd Yingxia Shao
BUPT
BUPT
Beijing, Country
yxshao@bupt.edu.cn

***Abstract*—With the development of electric vehicles, there comes a great need of charging stations for the recharging demand. However, where to locate a station and what are the main elements that operators should take into account when planning a setting, are still bothering problems that wait to be solved. In a common view, a better place to set a station ought to gurantee a relatively higher use rate of that station. Therefore, the problem changes into how to gain a higher use rate, and what are the factors that have important impact on it. In this paper, we propose a spatio temporal data based prediction framework of use rate for charging stations in Shanghai. The approach proposed in this paper takes both station's geographical information, such as longitude, latitude, surrounding Point of Interests(POIs), and working elements including price, AC/DC type and private or public to use, into consideration. After prerocessing, we seperate our datasets into urban area, suburb area and different time frames including total time, weekday, weekend, morning, evening, moring_rush hours, evening_rush hours and travel hours. We evaluate our method in two tasks, including districts prediction and time frames prediction. The aim is to classify what the level of a charging station's use rate is in different area districts and during different time periods. Experimental results show that our method performs well on SVM, Random Forest and MLP(ANN), which demonstrates that featrues as geographical information and working elements play important roles in use rate of charging stations. What's more, in the second task, we also find that the prediction accuracy is strongly attached to time span that a time frame covers.**

***Index Terms*—charging station, use rate, POIs, price, AC/DC, private, public, time frames**

I. Introduction

An electric vehicle charging station, also called EV charging station is an element in an infrastructure that supplies electric energy for the recharging of electric vehicles, such as plug-in electric vehicles and plug-in hybrids. At home or work, some electric vehicles have onboard converters that can plug into a standard electrical outlet or a high-capacity appliance outlet.

However, in most cases, others require a charging station that provides electrical conversion, monitoring, or safety functionality. These stations are also needed when travelling, and many support faster charging at higher voltages and currents that are available from residential EVSEs. Public charging stations are typically on-street facilities provided by electric utility companies or located at retail shopping centers and operated by many private companies.

Nowadays, electric vehicles become more and more popular as people want to cut down the pollution and cost of the usage of traditional energy. When most people considering whether to switch to an electric car, the most worrying aspect is the development of charging stations, which directly impact the usability of the electric car.

EVs in China is experiencing an overall growth in the past few years, which directly results in the massive construction and modification of charging stations and current road networks. Fig.1 shows the global distribution of an operators' charging stations in Shanghai, where there are over 600 stations. However, the cost of construction of charging stations is considerable and are often very costly or even impracticable to reallocate. This raise the question of how to select the locations for building the charging stations. In a typical view, a charging station no matter where it is located, its 'success' is often determined by the use rate of a station.

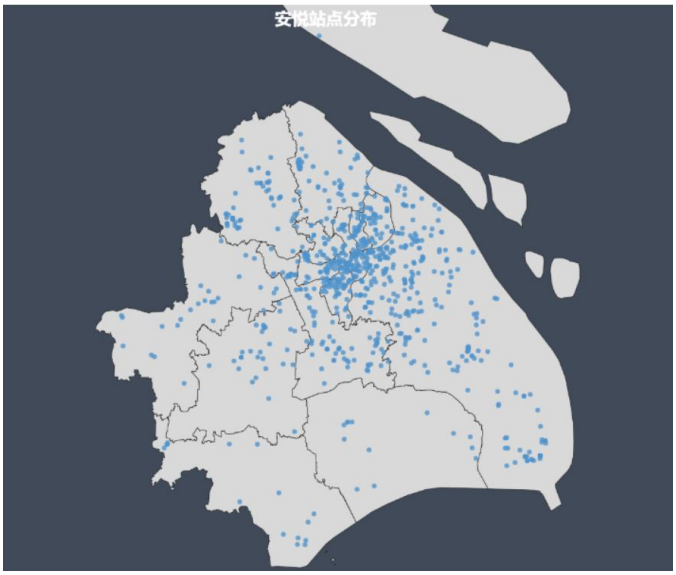


Fig. 1. Distribution of charing stations in Shanghai, China

In order to help with station setting strategies, we explore a time frame based prediction framework to classify stations into high use rate and low use rate in various time frames, using features like georaphical information as well as other working elements. We set 8 time frames in total, they are total time, weekdays, weekends, mornings, evenings, moring_rush hours, evening_rush hours and travel hours. As for features, we choose longitude, latitude, Point of Interests(POIs), charging

price, AC/DC charging types and whether it's private or public to use.

We make use of an operator's charging station data in Shanghai and have kept collecting the use rate value for about a month, then we seperate the whole use rate dataset into various time frames that we have already planned. At the same time, we also collect important featrure data as formerly said. Both of the two datasets require a data cleaning procedure, in order to filter some invalid data or outliers. Furthermore, we make some analyses on features that we confirm to have significant impact on station's use rate.

We evaluate our method with four machine learning algorithm, they are RL, Random Forest, SVM and XGBOOST respectively. In each time frame, with features put into it, the model will tell which level of use rate that a station belongs to. Experimental results show that geopraphical information as well as working elements of a station do have great influence on its use rate, which can bring operators some enlightment on location choosing for station construction.

In summary, the contributions of this work are listed as follows:

- We propose a time frame based prediction framework to classify whether a staion is of high use rate or low use rate based on operator's charging station data and important features.
- We make detailed analyses on both of the two datasets to obtain basic information and find the relationships between station's use rate and those features.
- We make use of four learning models for improving the accuracy of use rate prediction and achieve relatively favorable results.

The rest of this paper is organized as follows: Section 2 gives definition of the problem. Section 3 reviews some related work done before. Section 4 describes work on collected dataset and feature analyses. Section 5 provides experiments with four machine learning algorithms and the results. In Section 6, we draw a conclusion to the paper and make discussions on future work.

II. Related Work

Hence, we studied the optimization problem of how to deploy charging stations. Existing works mainly fall into the domain of bike-sharing. [1] provieds a data-driven apporoach to deal with bike lane construction problem. It takes government constraints of planning bike lanes, such as budget limitations, construction convenience and bike lane utilization into consideration to formulate the problem. Furthermore, the problem is proved to be NP-hard so that they propose a greedy network expansion algorithm to help work out a scalable and approximate solution to bike lane planning problem. The approach performs well in the given problem, however it doesn't make use of learning models. [2] introduces a reinforcement learning algorithm to help solve the problem of repositioning sharing-bikes. First it uses an inner-balance clustering algorithem to cluster stations into groups, then the reinforcement learning algorithm is conducted in each group to learn a reposition policy. They make a good use of spatio-temporal data while don't take advantages of useful geographical and station-self features.

Current works of location selection are usually based on the flow prediction of a single station. Futhermore, they rely heavily on the historical data. [3] introduces a model for bicycle mobility prediction. It relis on historical bike-sharing data and a per-station basis with sub-hour granularity. It makes use of the randonm forest prediction model to implement their experiments and obtain a rather good result. [4] gives an optimization to this problem. In this work, traffic prediction no longer focus on the history data only, but can use location-based socail media to collect a much larger area of the traffic data for predicting traffic conditions. [5] is also a good example of prediction model for spatio-temporal mobility event. It encodes each POI's spatio and temporal dependencies rather than neglect the correlations between POIs. In this paper, we argue that the surrounding point of interests(POIs), distances to important POIs(e.g, metro stations, estates, etc.), station charging price, AC/DC station types as well as whether a station is private of public for use, play important roles in selecting the optimal location for stations.

III. Design Overview

A. Problem Formulation

There are many elements that affect location selection for charging stations when operators make the decision. In this paper, we hold that use rate of a station is the key factor which determines the 'success' of the station. Therefore, the original problem turns into how to get a higher use rate and what are the factors behind it.

The main objectives of our work is three-fold. First, we aim to explore some important featrues that have great impact on use rate of a station using spatio-temporal data of operator's charging station data. Second, we propose to study stations' different 'behaviours' during diverse time frames. Finally, on the basis of time frame based framework, we aim to predict that one station is of high use rate or low use rate according to its georaphical information and working elements during different time periods.

B. Design Methodology

Since we consider that diffrent features may have different influence on station's use rate, we make detailed analyses on features like geographical information and stations' working elements. Furthermore, station's use rate may differ during different time periods, so that we also study on each time frame to find the changes. Based on all the analyses mentioned above, we propose the time frame based framework to help with use rate prediction for charging stations. We then apply three machine learning algorithms including SVC, Random Forest and MLP(ANN) to implement our experiments on different area districts and time frames. Fig.3 gives the overall data processing and learning pipelines of our work.

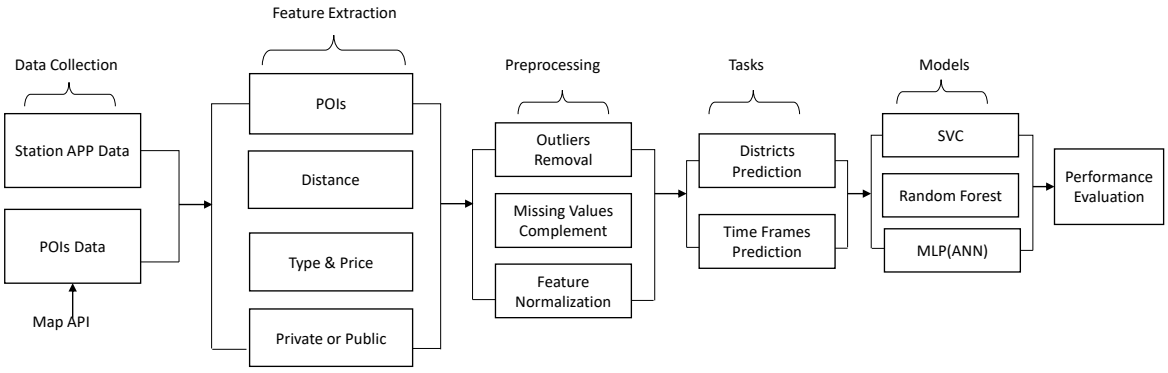


Fig. 2. Overall Pipelines of Data Processing and Learning

IV. Methodology

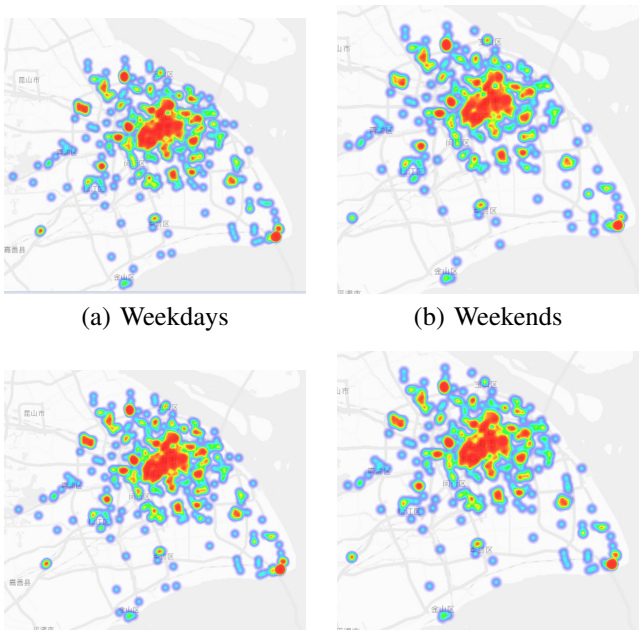
A. Feature Extraction

1) Datasets: In this paper, we gather the charging stations logs from all the existing charging station companies who provide their services in Shanghai, with a total of over 2,000,000 lines. The log has a length of one month, from 2018/10 to 2018/11, in which an hourly summarize of each charging station is recorded, showing whether it's occupied or not.

2) Hotspots: From the data we gathered, we made several observations that benefits the features to be included in the model we present. We separate the data into different time frames in order to determine the overall difference among them, since charging network is a dynamic system. From Fig.**??**, it is easy to notice that in different time frames, the charging hotspots stays at almost the same locations, meaning that during different time periods, the use rate of a given charging station is determined by its spatiotemporal context.

3) Point Of Interest: To the prosperity of Shanghai, there are so many points of interest(e.g., shopping malls, schools. estates, companies, etc.) located in the city. Understanding the purpose of the trip by each person who participates in charging system will help us to analyze the use rate prediction. So, we need to extract the POI around each existing charging station. There are a lot of POIs in Shanghai. The POIs are very close to each other. We set a radius around each station and then collect the POIs within the radius. Based on our experiment, choosing a radius as 300 meters is proper. In our work, we get 80 different types of POI totally. Some of the POIs are very similar to each other. Therefore, we group 80 POIs in further step. By grouping POIs, we get 10 groups at last. Table.I gives the groups and POIs in detail.

4) Distance: In use rate prediction, we need to consider distance. People will not choose to park their electric cars for charging if the destination they planned to go is far away. In the system, a station with nearer distance to metro stations, financial centers and major functional buildings would easily be used more often. We select the nearest distance to the following to be the distance we considered as features: company, estate, hospital, metro station, shopping center and university. Fig.4 shows the total count of these important 'distance' POIs, from the figures we can see that these features almost satisfy the long tail distribution, which will be normalized in later work, see Preprocessing part.



(a) Weekdays (b) Weekends

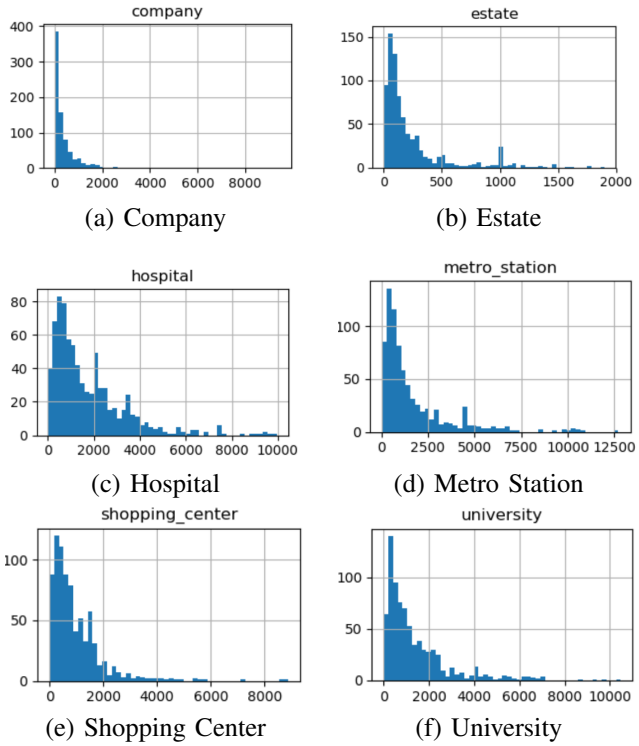
(c) Mornings (d) Evenings

Fig. 3. Charging Hotspots in Shanghai in different time frames

5) Type & Price: By futher digging into the data, we find that there is a 0.3 correlation between price for charging and the use rate. Since there are two types of charging ports: DC and AC. We would include the number of ports and the price of both types in a charging station as one of its feature. Fig.5 shows the number of AC/DC type of stations's charging ports,

TABLE I
GROUPS OF POIS

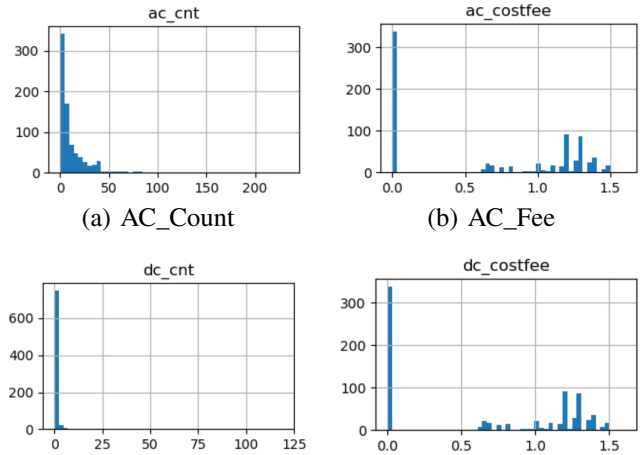
Group	Points of interests
Food	chinese & foreign restaurant, snack bar, cake & dessert shop, cafe, bar
Hotels	star hotel, express hotel, apartment hotel
Shopping	shopping centers, department stores, supermarkets, convenience stores, home building materials, home appliances digital, shops, markets
Education	institutions of higher learning, middle schools, primary schools, kindergartens, adult education, parent-child education, special education schools, study agencies, research institutions, training institutions, libraries, science and technology museums
Cultural venue	press and publication, radio and television, art groups, art galleries, exhibition halls, cultural palaces
Medical	general hospitals, specialist hospitals, clinics, pharmacies, medical examination institutions, nursing homes, emergency centers, disease control centers
Car service	car sales, car repair, car beauty, auto parts, car rental, car inspection field
Transportation	airport, railway station, subway station, subway line, long-distance bus station, bus station, bus line, port, parking lot, refueling station, service area, toll station, bridge, charging station, roadside parking space
Estates	Office building, residential area, dormitory



(a) Company (b) Estate (c) Hospital (d) Metro Station (e) Shopping Center (f) University

Fig. 4. Count of Important POIs

as well as each type's charging cost fee. From the figures we observe that there are much more AC type charging ports than DC type ones, while the charging cost fee of them two are almost the same.



(a) AC_Count (b) AC_Fee (c) DC_Count (d) DC_Fee

Fig. 5. Charging Hotspots in Shanghai in different time frames

6) Private or public: By observing the data we've collected, it can be seen that most of the charging stations are private charging stations, which means they are typically used by electric buses and rent cars, or used by specific companies for their employees, accounting for over 70% of the total charging stations. Also, since the private are used by more regular users(e.g., buses, companies employees), its use rate are 5% higher compared to public ones. This alongside other observations will be taken into consideration.

B. Spatio Temporal Data Based Framework

1) Ditricts: There are 16 districts of station data in total, Fig.6 shows the distribution of charging stations in Shanghai. And for experiment , we separate them into two parts as urban area and suburb area. The final dataset for district prediction task contains the two parts of data, in urban part, 7 districts are included while in suburb part, 9 left districts are included.

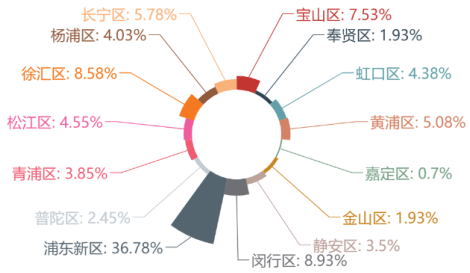


Fig. 6. District Distribution of Charging Stations

2) Time Frames: In order to further study stations' use rate in different time periods, we set different time frames, including total time, weekday, weekend, daytime, evening time, morning_rush hours, evening_rush hours and travel_hours, to help observe the use rate discrepancy during these phases. Fig.7 shows the average use-rate of the time frames above.

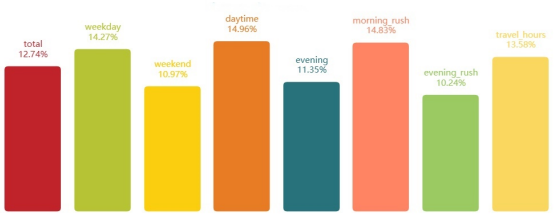


Fig. 7. Average use-rate of different time frames

C. Preprocessing

1) Data Cleaning: The original dataset may exist some outliers and missing values, there should be a data cleaning procedure to deal with these invalid data. For outliers, if detected, we remove them from the dataset; for missing values, we use the median filling method to fill blanks with median values computed via 'Imputer' function.

2) Feature Zooming: In 'nearest important POIs', the value of each feature is over hundred, while in other features like 'dc cost fee', the mean value is just about 0.85 per hour. In order to achieve a better performance using machine learning algorithms, we take a feature zooming procedure to normalize feature values so that they can range in [0,1]. Different models may use different zooming strategies, which will be discussed in detail in experiment section. Fig.8 shows an example of the feature 'university' after normalization step.

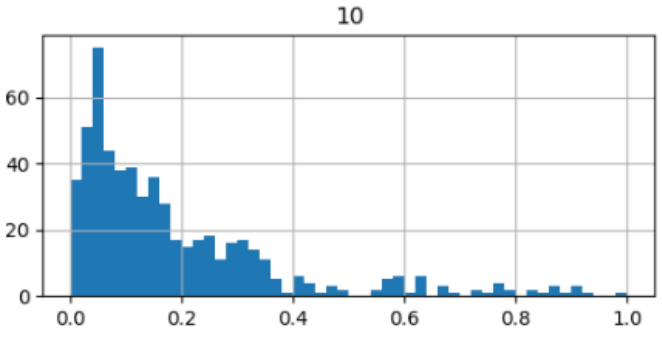


Fig. 8. An Example of Feature Normalization

D. Implementation

We aim to implement that when given important features, the model can tell what's the level of a station's use rate. Because the use rate of charging stations range in [0,1), so we set three levels as labels for classification task. In detail, we set use rate from 0 to 20% as low_use_rate; 20% to 50% as mid_use_rate; and over 50% as high_use_rate. For prediction tasks in both districts dataset and time frames dataset, we take features including stations' surrounding POIs, the number of nearest important POIs(e.g., company, estate, hospital, metro station, shopping center, university), charging port type, charging cost fee and whether it's for private or public use, into consideration. Then we make use of three classification models: SVC, Random Forest and MLP(ANN). With features added into these models, we evaluate the final results and find relationships between classification accuracy and features. The detailed experiment procedure will be discussed in next section.

V. Experiments

A. Parameter Settings

We run our dataset on SVM, Random Forest and MLP respectively. For SVM, we make use of its 'SVC' module for classification task and choose 'linear' function for its kernel, the punishment parameter C is set to default as 1.0. For Random Forest, we set the number of decision-making trees 'n_estimators' as 8. For MLP, we choose 'lbfgs' as solver and the activation function is 'relu', in order to avoid overfitting, we set the regular term alpha as 1e-4, the hidden_layer_sizes is set to (5,3) to achieve the best performance.

B. District Prediction

The task of district prediction aims to classify stations into high, middle or low use rate in urban and suburb areas of Shanghai.

We first separate our dataset into two parts: urban areas data and suburb areas data. After this operation, about 30% of station data is in the urban one, and other 70% of data is in the suburb one. Then, for each part, we randomly choose 80% of the data as training set, and 20% left as test set. For SVC and MLP, there shoule be a feature normalization procedure, we conduct MinMaxScaler for SVC and StandarScaler for MLP to set constraints on feature value so that they can range in (0,1). For Random Forest, there's no need for normalization, so we keep the feature data as original when we implement our experiment on it. We use mean average precision(MAP) as evaluation method which is commonly used in multilabel classification task.

Table.II and Table.III show the prediction accuracy of different models on urban areas data and suburb areas data. From the results we observe that: (1) SVC and Random Forest can perform stably on both two tasks. (2) MLP performs best in suburb prediciton while performs worst in urban prediction. The reason we infer is that because of the distribution of station data mentioned above, the urban part has lower data volume than the suburb part, this may influence the model performance due to MLP's neural network characteristic, while SVC and Random Forest are not sensitive to this difference. Table.II shows the prediction accuracy of different models. We can see that they can all perform well based on our settings, and XGBOOST achieves the most favourable result.

TABLE II
EVALUATION RESULTS ON URBAN PREDICTION

Model	Accuracy
LR	60.9%
SVM	68.9%
Random Forest	70.1%
XGBOOST	73.6%
SVC	84.62%
MLP	74.35%
Random Forest	79.52%
(Mean)	79.50%

TABLE III
EVALUATION RESULTS ON SUBURB PREDICTION

Model	Accuracy
SVC	79.52%
MLP	80.72%
Random Forest	78.31%
(Mean)	79.52%

C. Time Frames Prediction

Stations' use rate in different time frames is a more concerned problem we want to explore. According to time frames partition mentioned in Section 4, we also separate station data into 7 time frames: weekday, weekend, morining, morning_rush_hours, evening, evening_rush_hours and travel hours. In this task, features like POIs, charging port type and charging price still stay the same, while the use rate itself will vary in different time frames, so that the level of use rate, also the classification labels will differ from each time frame dataset. After preprocessing, we also randomly choose 80% of valid data as training set, and the left as test set. We use the same models as used in districts prediction task, SVC, Random Forest and MLP(ANN), and then conduct our experiments.

Table.IV and Table.V show the accuracy of different models on weekday data and weekend data. We can see that in weekday time, all the three model just perform ordinarily, achieve an average accuracy of 74.71%. However, in weekend time, all of them perform pretty well and achive a more than 90% score. We assume that classification models can perform better in a shoter time span dataset. And we continue the further experiments.

TABLE IV
EVALUATION RESULTS ON WEEKDAY PREDICTION

Model	Accuracy
SVC	75%
MLP	76.72%
Random Forest	72.41%
(Mean)	74.71%

Then, we conduct experiments in morning, morning_rush_hours, evening, evening_rush_hours datasets respectively to verify our assumption. Table.VI and Table.VII show the results on the former two datasets. As the result shows, models perform much better in rush_hours which has a shorter time span.

TABLE V
EVALUATION RESULTS ON WEEKEND PREDICTION

Model	Accuracy
SVC	96.55%
MLP	99.13%
Random Forest	98.27%
(Mean)	97.98%

TABLE VI
EVALUATION RESULTS ON MORNING PREDICTION

Model	Accuracy
SVC	75%
MLP	77.59%
Random Forest	74.14%
(Mean)	75.58%

TABLE VII
EVALUATION RESULTS ON MORINING_RUSH_HOURS PREDICTION

Model	Accuracy
SVC	95.69%
MLP	95.69%
Random Forest	96.55%
(Mean)	95.98%

The results on evening and evening_rush_hours datasets are according with our observation, as Table.VIII and Table.IX show. It can also be seen that because evening hours are a little shorter than morning hours, the classification accuracy is some higher in evening time. When comparing morning and evening rush_hours, it's common sense that the two time frames cover almost the same time span, so the results differ scarcely.

TABLE VIII
EVALUATION RESULTS ON EVENING PREDICTION

Model	Accuracy
SVC	82.76%
MLP	81.90%
Random Forest	83.62%
(Mean)	82.76%

TABLE IX
EVALUATION RESULTS ON EVENING_RUSH_HOURS PREDICTION

Model	Accuracy
SVC	98.28%
MLP	96.55%
Random Forest	97.41%
(Mean)	97.41%

Finally, we conduct experiments on travel_hours dataset. The time span mainly falls in the National Day holidays.

Table.X shows the classification results. It also verifies our observation, since its time span is as long as one week in 'weekday' time.

TABLE X
EVALUATION RESULTS ON TRAVEL_HOURS PREDICTION

Model	Accuracy
SVC	75.86%
MLP	79.31%
Random Forest	76.72%
(Mean)	77.30%

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose the time frame based prediction model for use rate of charging stations. We study on some important features like station's surrouding POIs, charing price and station type. In experiments, we separate our dataset into different time frames and add the features into them, which obtains a relatively favourable result, indicating that the use rate of charing station is highly influenced by its geographical information and working elements. Furthermore, it also has a greate impact on location choosing problem.

There is still a lot of work to be continued in the future. First, we only consider three main types of features that might affect station use rate, many other important features also need to be extracted and incluede. Furthermore, we might explore a properly modified model to execute those features and gain a more satisfactory result.

REFERENCES

[1] J. Bao, T. He, S. Ruan, Y. Li, and Y. Zheng, "Planning bike lanes based on sharing-bikes' trajectories," in *Proc. of KDD*, 2017, pp. 1377–1386.

[2] Y. Li, Y. Zheng, and Q. Yang, "Dynamic bike reposition: A spatio-temporal reinforcement learning approach," in *Proc. of KDD*, 2018, pp. 1724–1733.

[3] Z. Yang, J. Hu, Y. Shu, P. Cheng, J. Chen, and T. Moscibroda, "Mobility modeling and prediction in bike-sharing systems," in *Proc. of MobiSys*, 2016, pp. 165–178.

[4] X. Liu, X. Kong, and Y. Li, "Collective traffic prediction with partially observed traffic history using location-based social media," in *Proc. of CIKM*, 2016, pp. 2179–2184.

[5] B. Shen, X. Liang, Y. Ouyang, M. Liu, W. Zheng, and K. M. Carley, "Stepdeep: A novel spatial-temporal mobility event prediction framework based on deep neural network," in *Proc. of KDD*, 2018, pp. 724–733.