Demand prediction of mobile clinics using public data

Haipeng Chen
Harvard University
hpchen@seas.harvard.edu

Susobhan Ghosh
Harvard University
susobhang70@gmail.com

Gregory Fan
The Family Van & Harvard Medical School
Gregory_Fan@hms.harvard.edu

Nikhil Behari
Harvard University
nikhilbehari@college.harvard.edu

Arpita Biswas
Harvard University
arpitabiswas@seas.harvard.edu;

Mollie Williams
The Family Van & Harvard Medical School
Mollie_Williams@hms.harvard.edu;

Nancy E. Oriol
The Family Van & Harvard Medical School
Nancy_Oriol@hms.harvard.edu

Milind Tambe
Harvard University
milind_tambe@harvard.edu

Abstract

The advent of mobile clinics plays an important role in enhancing health equity, as they can provide easier access to preventive healthcare for patients from marginalized populations. For effective functioning of mobile clinics, accurate prediction of demand (expected number of individuals visiting mobile clinic) is the key to their daily operations and staff/resource allocation. Despite its importance, there are very limited studies on predicting demand of mobile clinics. We are among the first few to explore this area, using AI-based techniques. A crucial challenge in this task is that there are no known existing data sources from which we can extract useful information to account for the exogenous factors that may affect the demand. We propose a novel methodology that uses public data to obtain the features, with several innovations that are designed to improve prediction. Empirical evaluation on a real-world dataset from the mobile clinic The Family Van shows that, by leveraging publicly available data (which introduces no extra monetary cost to the mobile clinics), our method achieves $26.4\% - 51.8\%$ lower Root Mean Squared Error (RMSE) than the historical average-based estimation (which is presently employed by mobile clinics like The Family Van).

1 Introduction

The disproportionate impact of the COVID-19 pandemic on historically disadvantaged populations has exemplified long-standing health inequities in the United States [3, 5, 9]. These health disparities are often the result of marginalized populations facing increased barriers to healthcare access, including fear or mistrust of the medical system, and prohibitive travel times [1, 8]. In response to the pandemic and the health disparities, healthcare providers and public health organizations have begun implementing novel solutions such as the mobile health clinic, which is a large bus or van that is converted to provide medical care at a location physically and socially closer to at-risk communities.

A main benefit of using mobile clinics is their ability to relocate to areas of high demand. It is critical for mobile clinic administrators to understand what key factors affect client demand and what future demand might look like. Predictive demand models could then be used to optimize van scheduling as

35th Conference on Neural Information Processing Systems, workshop on Machine Learning in Public Health (NeurIPS-MLPH), Sydney, Australia.

well as the allocation of staff and healthcare resources. Despite its importance, only a limited number of studies discuss forecasting client demand for mobile clinics. (see Appendix A for a discussion of related works). Developing such demand prediction models has several challenges: i) It is not clear which factors affect demand. ii) Unlike hospitals, data from mobile clinics are not as available. iii) For the factors that are identified as good indicators of demand, we have no prior knowledge of their ground truth values on the forecast date.

We propose a novel prediction framework that addresses the above challenges, with the following innovations. (i) We find features from public data sources that we hypothesize to be good indicators of demand. We determine which factors should be included in the prediction model by performing a correlation analysis of each factor with respect to mobile clinic demand. Very interestingly and surprisingly, we find that *factors such as ferry and shared bike usage, which are proxies of foot traffic, are highly correlated with client demand*. (ii) Though we can gain insights from the correlation analysis, a major obstacle to using these insights to extract features for the prediction model is that the values of these features on the forecast date are not known a-priori. Therefore, we propose to first make intermediate predictions for the future values of these variables, and then use the predicted variables as features. (iii) We observe that the demand patterns of repeat/non-repeat clients are distinct. To further improve prediction, we separate out the demand predictions for each type of client, and then combine the predictions. Finally, we integrate these components with different ML models to make the final prediction. Whereas we do not claim novelty for these ML models, our method is an innovative combination of the known ML models with the above mentioned components.

Our contributions include: (i) **Impact.** We are among the first to develop ML tools to automatically predict client demand of mobile health clinics. Except for the demand data (which is accessible to parties in the field), our model uses publicly available data to extract features. (ii) **Technical novelty.** We propose a novel prediction model which has several novel components that are designed to address the new challenges in the demand prediction task of interest. (iii) **Effectiveness.** We run experiments using real-world data from 4 locations of the mobile health clinic The Family Van in the Boston area. Results show that our new AI-based approach has a $26.4\% - 51.8\%$ lower RMSE than the traditional practice of The Family Van.

2 Demand prediction for mobile clinics

The Family Van is a non-profit mobile health clinic in the Boston area, designed to increase access to health care and improve the health of Boston's most under-served neighborhoods. It sends out medical vans to 4 major locations in Boston, one day per location, thus operating 4 days per week. Our goal is to predict the demand of the four locations for a future time (e.g., next week), given the historical demand data. Abstracting from the above scenario, we are given a time series $1, \ldots t, t+1, \ldots$, where each t is a week. There are various locations $i = 1 \ldots N$, one for each day of a week t. The demand at time t and location i is denoted as a non-negative integer $y_{i,t} \in \mathcal{Z}^+$. Suppose we are at time $t-1$, our goal is then to predict the mobile clinic's demand $y_{i,t}$ for each of the N locations, at the next time period t. Essentially, we want to learn the following demand function $f_\theta(\cdot)$:

$$\hat{y}_{i,t} = f_\theta(y_{i,1}, \ldots, y_{i,t-1} |\, \forall i = 1 \ldots N), \tag{1}$$

where $f_\theta(\cdot)$ denotes a certain ML model.

Location specific model Before determining the right ML model for prediction, our first observation, as shown in Figure 1 of Appendix B, is that the demand curves of different locations are distinct from each other. For example, the demand scales of different locations are different – the demand of location 1 can reach 30+, whereas the demand of location 2 is mostly under 15. In addition, the demand trend of location 3 is generally decreasing, whereas such a trend is not obvious for the other locations. Because of this, though a uniform and generalized model that works for all locations is desirable from a technical perspective, we choose to use a location specific model instead. In this case, Eq.(1) is re-written as: $\hat{y}_{i,t} = f_{\theta_i}(y_{i,1}, \ldots, y_{i,t-1}),\ \forall i = 1 \ldots N$. In other words, we will learn one separate model $f_{\theta_i}(\cdot)$ for each location $i = 1 \ldots N$.

Time-series model Without considering the exogenous factors, time-series models seem to be the immediate fit to our task. Autoregressive integrated moving average (ARIMA) [2] is arguably the most classical model for predicting time series data. The underlying ideas of ARIMA include auto-regression (i.e., the output variable is regressed on its own historical values), moving average (i.e., the regression error is a linear combination of error terms whose values occurred currently

and at various times in the past), and integration (i.e., the data values have been replaced with the difference between their values and the previous values). Each $f_{\theta_i}(\cdot)$ is represented as ARIMA(p,d,q) model. The preliminary prediction results of the ARIMA model are shown in row 2 of Table 1. For locations 1 and 2, the best prediction is obtained by ARIMA(0,0,0), meaning that the fitted model is a constant value plus a white noise. For locations 3 and 4, the best results are respectively obtained by ARIMA(2,2,4) and ARIMA(0,2,4), indicating that there may be a certain trend that is captured. Considering the scales of demand for different locations in Figure 1, the ARIMA-based predictions are reasonable for locations 3 and 4, but can be substantially improved for the other locations.

Extract and filter information from public data Training a model solely from historical demand would miss contextual information about demand dynamics, where the exogenous factors are completely neglected. This could lead to arbitrarily bad predictions, especially when there is no significant trend of any order (e.g., for locations 1 and 2). After discussing with practitioners from The Family Van, we identify two important types of exogenous factors that can affect demand dynamics, i.e., weather and foot traffic. It is intuitive that weather can affect the client demand. For example, people usually tend not to go out when the temperature is either too high or too low. In the meanwhile, we observe that the client demand is higher when there is more crowd around the vans. This indicates that *foot traffic* is an important factor of demand. Unfortunately, to the best of our knowledge, no foot traffic data in Boston are publicly available. Therefore, we use 3 other types of traffic information as surrogates of foot traffic, namely the public library usage data, the ferry usage data, and shared bike usage data. The links to the data sources are described in Appendix C.

Though we have a long list of features that could be potentially useful for prediction, it is often harmful to include any feature in a model, especially when a feature is not or very weakly correlated to the output. To measure the dependency of demand w.r.t. the features and therefore filter out informative features, we perform an Ordinary Least Squares (OLS) linear regression of demand vs each individual feature. Figure 3 of Appendix C shows the scatter plots of demand vs the feature value for some selected features. Each dot in the scatter plots represents a day's data point. The fitted linear curves are shown as the lines in the scatter plots. The p-values of the fitted univariate linear regression models are shown in Table 2 of Appendix C. We then extract features whose associated p-values of the linear regression are smaller than 0.05. This returns 5 features: solar radiation, humidity, temperature, ferry usage and Blue Bike usage. *Surprisingly, 2 of the 3 features (Blue Bike usage ferry usage) that we consider as proxies for the foot traffic factor are highly correlated with demand.*

Intermediate predictions The previous section shows important insights on which features are informative of the mobile clinics demand. Though these features are desirable, the main barrier from actually using these features is that, the *future* values of these features are not available at *the time of prediction*. For example, when making predictions on Sunday, the blue bike usage information of the next Wednesday is not known a-priori. For weather related features, fortunately, we can obtain reasonably accurate weather forecasts as the estimated future values. However, there are no known tools or sources to obtain the estimated future values for the surrogate foot traffic features. To overcome this issue, instead of assuming the future values are available, we first use the time series models (e.g., ARIMA) to make *intermediate* predictions of the future values for these features, and then use the predicted values as input. The advantage of intermediate prediction vs an end-to-end architecture is two-fold. First, the model is trained with loss defined directly on the intermediate variable, instead of the final client demand. The latter introduces more noise. Second, we can use more data (rather than the target time period) to train the intermediate models. Appendix D discusses the intermediate prediction results.

Separate predictions for repeat and non-repeat clients With the intermediate predictions, we integrate them with different types of ML models to make the prediction. For non-time-series (NTS) models, this means $\hat{y}_{i,t} = f_\theta(\hat{x}_{i,t}; t)$, where $\hat{x}_{i,t}$ is the estimated feature values based on the intermediate predictions. Here we still feed the time information t to this model to compensate for the loss of temporal information. Alternatively, we have also used recurrent neural networks which combine the historical demand and estimated contextual features: $\hat{y}_{i,t} = f_\theta(y_{i,1}, \ldots, y_{i,t-1}; \hat{x}_{i,t})$.

Another important observation, as shown in Fig. 2 of Appendix 2, is that the demand patterns of the repeat and non-repeat clients for different locations are distinct. Take location 2 as an example, the demand curve of the non-repeat clients has a significantly larger fluctuation over time, whereas the demand of the repeat clients is more stationary. Inspired by this, we propose an alternative training method, where we first train 2 separate models for each group of the repeat and non-repeat clients,

and then sum the predictions of the two models as the final prediction. This essentially means that we will train a model for each of $\hat{y}^R_{i,t}$ (repeat client demand) and $\hat{y}^N_{i,t}$ (non-repeat client demand).

Table 1: Demand prediction of our methods. The best result for each location is highlighted.

row number	method \ location	Location 1	Location 2	Location 3	Location 4
1	Historical average	7.06	4.20	8.91	4.45
2	ARIMA	7.11	4.28	**5.71**	3.18
3	NTS-**AB**	6.14 ± 0.0000	3.95 ± 0.0000	6.47 ± 0.0000	3.17 ± 0.3776
4	NTS-**Ab**	6.17 ± 0.0000	$\mathbf{3.09} \pm 0.5938$	6.48 ± 0.0000	$\mathbf{2.90} \pm 0.0000$
5	NTS-**aB**	$\mathbf{3.40} \pm 0.0000$	3.77 ± 0.4502	6.80 ± 0.0000	3.39 ± 0.0000
6	NTS-**ab**	3.65 ± 0.0000	3.93 ± 0.0000	6.69 ± 0.0000	3.31 ± 0.0199
7	RNN-**AB**	7.96 ± 1.8481	10.31 ± 1.1406	6.03 ± 0.2806	4.96 ± 0.3454
8	RNN-**Ab**	7.64 ± 0.7485	7.39 ± 1.4918	6.66 ± 0.2225	6.75 ± 1.0613
9	RNN-**aB**	8.33 ± 0.7146	10.28 ± 1.2289	7.38 ± 0.6014	5.31 ± 0.3027
10	RNN-**ab**	7.47 ± 0.5451	7.11 ± 1.8134	6.53 ± 0.1624	5.63 ± 0.4278

3 Main results

Our demand dataset spans from July 2019 to March 2020. It contains daily demand data for 4 locations in Boston. The Family Van operates on 4 days a week, one day per location. Therefore, we have one data point per week per location during the target time period, totalling 20-30 data points per location and 110 data points for all locations. We do a train/test split of 80/20. In practice, the set of NTS models we implement are: Linear, Ridge, Lasso, Lasso LARS, Tweedie, SGD, Logistic, MLP, Adaboost, Decision Tree, and XGBoost. For RNNs, we use 3 structures: vanilla RNN, LSTM and GRU. We run 20 times of training, and report the average RMSE and standard deviation of RMSE of the best NTS or RNN model, respectively. For intermediate features, we can either train with ground truth feature values, and test with predicted feature values, or both train/test with predicted feature values. We use **A** to denote the former setting and **a** to denote the latter. Similarly, we can either train a single model for both types of the repeat and non-repeat clients, or train 2 separate models. We denote them as **B** and **b**, respectively. Therefore, a training method **Ab** means we train with ground truth feature values, and test with predicted feature values, and at the same time train 2 separate models for the repeat/non-repeat clients. The main results are shown in Table 1.

Are predictions accurate in general? Comparing the best result and the historical average in row 1 (which is the current practice of The Family Van), we can see that the RMSE respectively decreases by 51.8%, 26.4%, 35.9%, 34.8% for the 4 locations. Combing the overall scales of the demand for the 4 locations (see Figure 1), this demonstrates that *our prediction is reasonably accurate.*

What is the best ML model (if there is any)? Most surprisingly, RNNs are dominated by NTS and ARIMA. Our hypothesis is that deep models have more parameters, and may overfit in small datasets like ours. Second, NTS models are significantly better than ARIMA, except for location 3. Our hypothesis is that the demand curve (see Figure 1) for location 3 is noticeably declining, whereas this is not obvious for other locations, especially 1 and 2. *There is no single best model for all locations.*

Are exogenous factors and intermediate predictions helping? Comparing the best results for each location in Table 1 and the results obtained by the pure time-series ARIMA model (row 2 in Table 1), there are substantial improvements for 3 locations. Notably, for location 1, the RMSE of the ARIMA model is more than twice of the NTS model. ARIMA performs slightly better in location 3. *This shows that exogenous factors are critical to the demand prediction.* Another interesting observation is that, when both training and testing on the intermediate feature values, there is a substantial gain in accuracy for location 1 (comparing row 3 vs 5 or row 4 vs 6). Our hypothesis is that the intermediate prediction of features like Blue Bike usage is biased. Therefore, when training on observed features and testing on predicted features, there is a gap between the two values. This gap is somehow removed when both training and testing on the intermediate feature values.

Is separating predictions for repeat and non-repeat clients helping? Comparing row 3 vs row 4, we can see that separating predictions leads to substantial improvements in locations 2 (3.95 to 3.09) and location 4 (3.17 to 2.9), while maintaining very close results in the other 2 locations. Our hypothesis for the improvement is that for locations 2 and 4, the demand patterns of the repeat and non-repeat clients are more distinct, as shown in Figure 2. Comparing row 5 vs row 6, we can see that the results are close for both training methods in all of the locations (separate prediction is

slightly better in locations 3 and 4, and slightly worse in locations 1 and 2). This shows that *separate prediction helps in improving the overall prediction accuracy*.

4 Conclusion

We are among the first to explore demand prediction of mobile clinics using AI. We propose a novel learning framework that uses publicly available data for prediction, together with multiple innovations that are customized into solving the demand prediction problem. Empirical results on real-world datasets from The Family Van demonstrate that our proposed approach has substantial improvement in accuracy compared to the experience-based estimation. Our study provides a brand-new angle to mobile clinics demand prediction with completely public data, which has a huge potential impact when broadly deployed. As a future work, we are actively exploring how our prediction algorithm can be deployed to help The Family Van's daily scheduling of staff and healthcare resources.

References

[1] S. D. Bolen, P. Sage, A. T. Perzynski, and K. C. Stange. No moment wasted: the primary-care visit for adults with diabetes and low socio-economic status. *Prim. Health Care Res. Dev.*, 17(1):18–32, Jan. 2016.

[2] J. D. Hamilton. *Time series analysis*. Princeton university press, 2020.

[3] K. Mackey, C. K. Ayers, K. K. Kondo, S. Saha, S. M. Advani, S. Young, H. Spencer, M. Rusek, J. Anderson, S. Veazie, M. Smith, and D. Kansagara. Racial and ethnic disparities in COVID-19-Related infections, hospitalizations, and deaths : A systematic review. *Ann. Intern. Med.*, 174(3):362–373, Mar. 2021.

[4] B. Majeed, J. Peng, A. Li, Y. Lin, and R. I. Delgado. Forecasting the demand of mobile clinic services at vulnerable communities based on integrated multi-source data. *IISE Transactions on Healthcare Systems Engineering*, 11(2):113–127, 2021.

[5] M. Odlum, N. Moise, I. M. Kronish, P. Broadwell, C. Alcántara, N. J. Davis, Y. K. K. Cheung, A. Perotte, and S. Yoon. Trends in poor health indicators among black and hispanic middle-aged and older adults in the united states, 1999-2018. *JAMA Netw Open*, 3(11):e2025134, Nov. 2020.

[6] D. Qian, R. W. Pong, A. Yin, K. Nagarajan, and Q. Meng. Determinants of health care demand in poor, rural china: the case of gansu province. *Health Policy and Planning*, 24(5):324–334, 2009.

[7] C. Reed, F. A. Rabito, D. Werthmann, S. Smith, and J. C. Carlson. Factors associated with using alternative sources of primary care: a cross-sectional study. *BMC health services research*, 19(1):1–9, 2019.

[8] S. T. Syed, B. S. Gerber, and L. K. Sharp. Traveling towards disease: transportation barriers to health care access. *J. Community Health*, 38(5):976–993, Oct. 2013.

[9] F. J. Zimmerman and N. W. Anderson. Trends in health equity in the united states by Race/Ethnicity, sex, and income, 1993-2017. *JAMA Netw Open*, 2(6):e196386, June 2019.

A Related work

We provide a brief discussion of related work on health care demand forecasting. Majeed et al. [4] focus on predicting demand for mobile clinics that provide free vaccination services in schools and regional areas that are at high risk of infection. They estimate the demand using non-temporal (static) census features and school-level data. In contrast, our work considers temporal features and makes real-time forecasting possible. Reed et al. [7] strictly use patient survey data and focus only on the perception of clinic quality. The paper examines associations between multiple features, such as perceived quality of care, travel distance, and patient utilization of alternative health care clinics, using multi-regression models on the survey data. Although the perception of alternate health care has an impact on the demand of mobile clinics, but relying on patient survey data may violate patients' privacy requirements. Moreover, such findings may suffer from selection bias. Qian et al. [6] focus on the factors that influence the health care demand of public and private clinics in rural areas of Gansu province, China. However, they do not consider mobile health clinics or temporal features.

B Demand curves

The overall demand curves and the demand curves for the repeat vs non-repeat clients are shown in Figures 1 and 2, respectively.

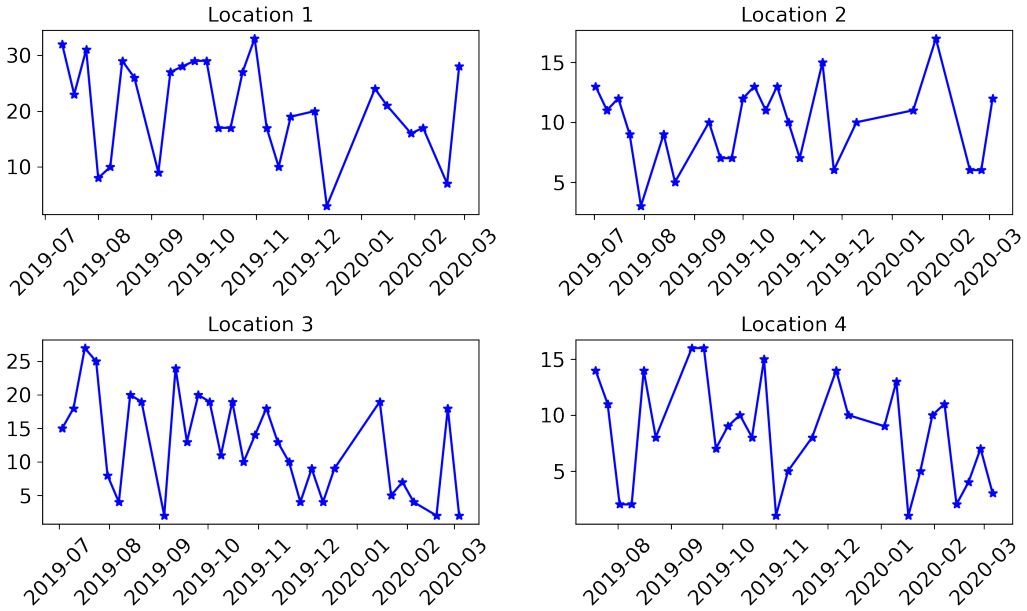


Figure 1: Distinct demand curves of different locations. The x-axis is the date, and the y-axis is the demand.

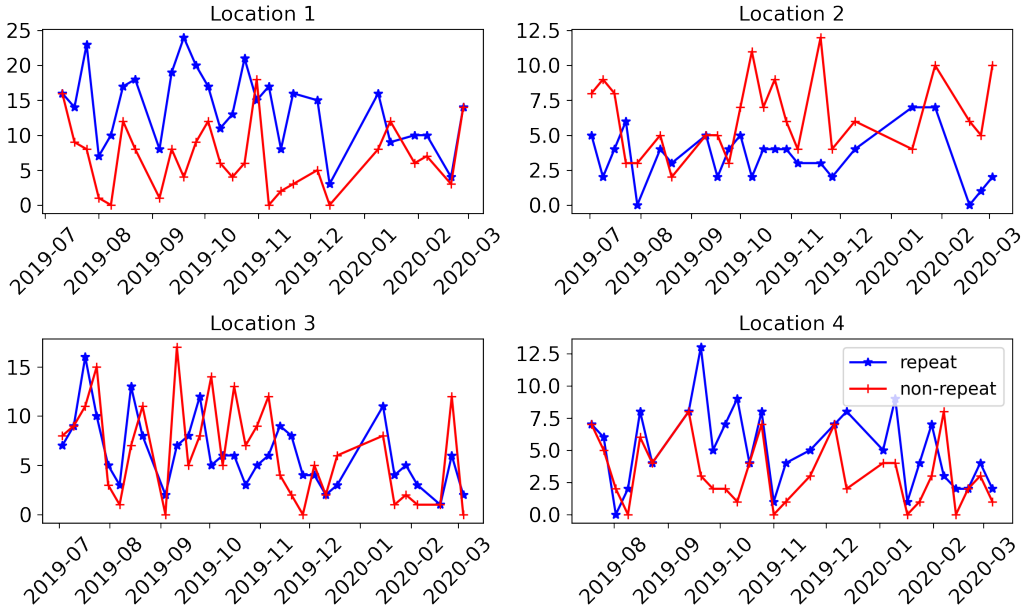
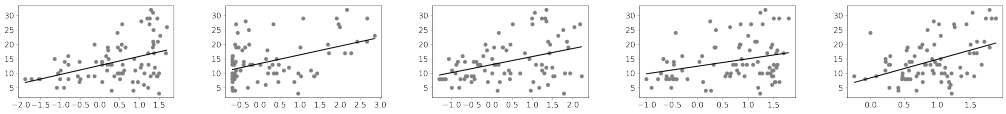


Figure 2: Demand curves of repeat vs non-repeat clients

C Data sources, plots and results related to exogenous factors

To extract weather information, we use the public weather data from the (US) National Oceanic and Atmospheric Administration (NOAA)[1] and Copernicus.[2] The 3 types of traffic-related data include the public library usage data which is obtained from Analyze Boston,[3] the Massachusetts Bay Transportation Authority (MBTA) ferry usage data from Open Data MBTA, [4] and Blue Bike usage data from Bluebikes.[5]

Figure 3 shows the scatter plots of demand vs different features. Table 2 shows the p-values of the OLS linear regression for different univariate models.



(a) Temperature (b) Solar radiation (c) Humidity (d) Bike usage (e) Ferry usage

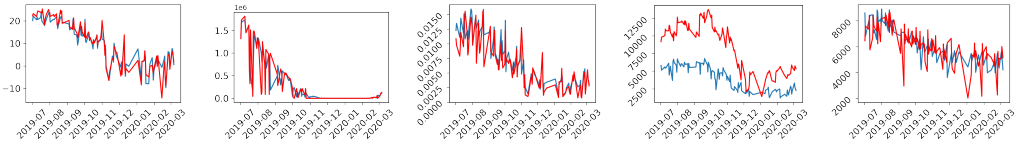
Figure 3: Scatter plots of demand vs different factors. x-axis is the normalized value of the underlying factor, y-axis is demand.

Table 2: P-values of univariate linear regression. Features with p-values < 0.05 are highlighted.

Feature	**Radiation**	**Humidity**	**Temperature**	Wind	**Ferry**	Library	**Bike**
p-value	0	0	0	0.597	0	0.732	0.007
Feature	Snow cover	Snow depth	Snowfall	Snowmelt	Pressure	Precipitation	Cloud
p-value	0.072	0.235	0.373	0.100	0.324	0.636	0.077

D Intermediate prediction results

Figure 4 shows the predicted and ground truth curves for the 5 selected features. We can see that the predictions are overall reasonably capturing the trends of the ground truth curves. For temperature, solar radiation and humidity, the intermediate predictions (obtained from weather forecast) are highly aligned with the ground truth values. This demonstrates that the weather forecasts are reliable. The same holds for ferry usage, where the prediction curve is also highly aligned with the ground truth. For bike usage intermediate prediction, there is a noticeable bias due to the discrepancy between the training and test data. Nonetheless, the trend is still largely captured by the prediction.



(a) Temperature (b) Solar radiation (c) Humidity (d) Bike usage (e) Ferry usage

Figure 4: Curves of ground truth variable (blue) and values based intermediate predictions (red).

[1] https://www.noaa.gov/
[2] https://www.copernicus.eu/en
[3] https://data.boston.gov/
[4] https://mbta-massdot.opendata.arcgis.com/
[5] https://www.bluebikes.com/