COMP3901 Research Project

Predicting Station-level Hourly Bike-Sharing Demand Using XGBoost and LSTM

Mike Gysel (z5251938)

Photo Credit: Jeff Greenberg, Getty Images

The Problem

- World population expected to increase by 2 billion people over next 30 years [1]
- Increased motorized vehicle use will lead to more congestion, noise, pollution, and greenhouse gas emissions [2, 3]
- Bike-Sharing Systems (BSS) pose one possible solution

Photo Credit: Craig Cole, Getty Images

Bike-Sharing Systems

- BSS: Shared transport service that allows for short-term bike rental at unattended urban locations [4]
- Benefits
  - Reduce vehicle miles traveled [5]
  - Improved public health [6]
- Requires fleet rebalancing

Photo Credit: Dereje, Shutterstock.com

Use of Machine Learning to Predict Bike-Sharing Demand

Table 1: Use of Machine Learning to Predict Bike-Sharing Demand				
Authors	**Article**	**Demand Level**	**Variables**	**Machine Learning Models Used**
Ashqar et al.	Modeling Bike Availability in a Bike-Sharing System Using Machine Learning [7]	Station	- Meteorological - Temporal	- Random Forest - Least-Squares Boosting - Partial Least Squares Regression
Choi and Han	The Empirical Evaluation of Models Predicting Bike Sharing Demand [8]	Station	- Meteorological - Temporal	- Random Forest - Gradient Boosting Machine (XGBoost) - Long Short-Term Memory - Gated Recurrent Units
Yang et al.	Using graph structural information about flows to enhance short-term demand prediction in bike-sharing systems [9]	Groups of Stations	- Meteorological - Temporal	- Gradient Boosting Machine (XGBoost) - Multilayer Perceptron - Long Short-Term Memory
Lin et al.	Predicting station-level bike-sharing demands using graph convolutional neural network [10]	Station	- Spatial - Temporal	- Geometric Convolutional Neural Network - Gradient Boosting Machine (XGBoost) - Long Short-Term Memory
Singhvi et al.	Predicting Bike Usage for New York City's Bike Sharing System [11]	Neighborhood	- Meteorological - Spatial - Taxi Usage	- Linear regression
Sathishkumar et al.	Using data mining techniques for bike sharing demand prediction in metropolitan city [12]	City	- Meteorological - Temporal	- Linear regression - Gradient Boosting Machine - Support Vector Machine - Boosted Trees - Extreme Gradient Boosting Trees

XGBoost (Extreme Gradient Boosting)

- XGBoost is an implementation of Gradient Boosting Machine (GBM)
- Uses an ensemble of decision trees
- Loss function, weak learner, additive model

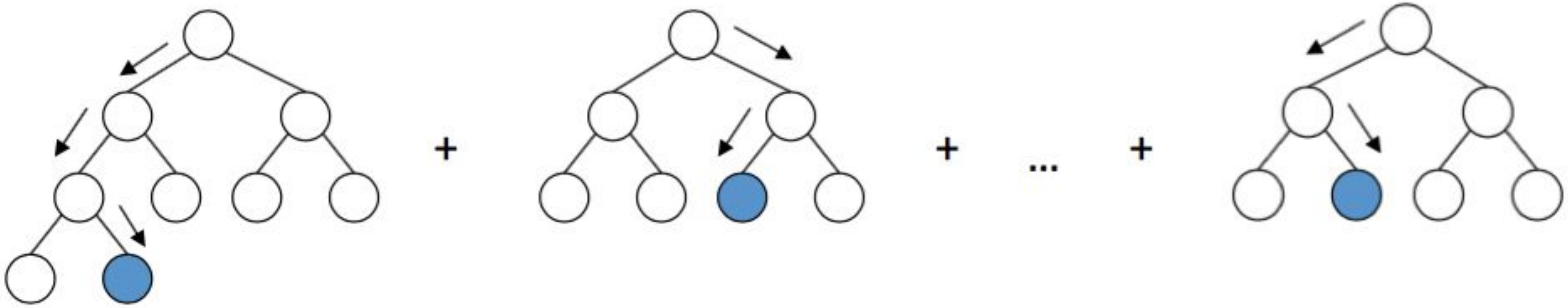

Figure 1: Model of Gradient Boosting Machine [13]

Long Short-Term Memory (LSTM)

- LSTM is a form of Recurrent Neural Network (RNN)
- RNNs allows past outputs to be used as inputs, useful in predicting time-series data

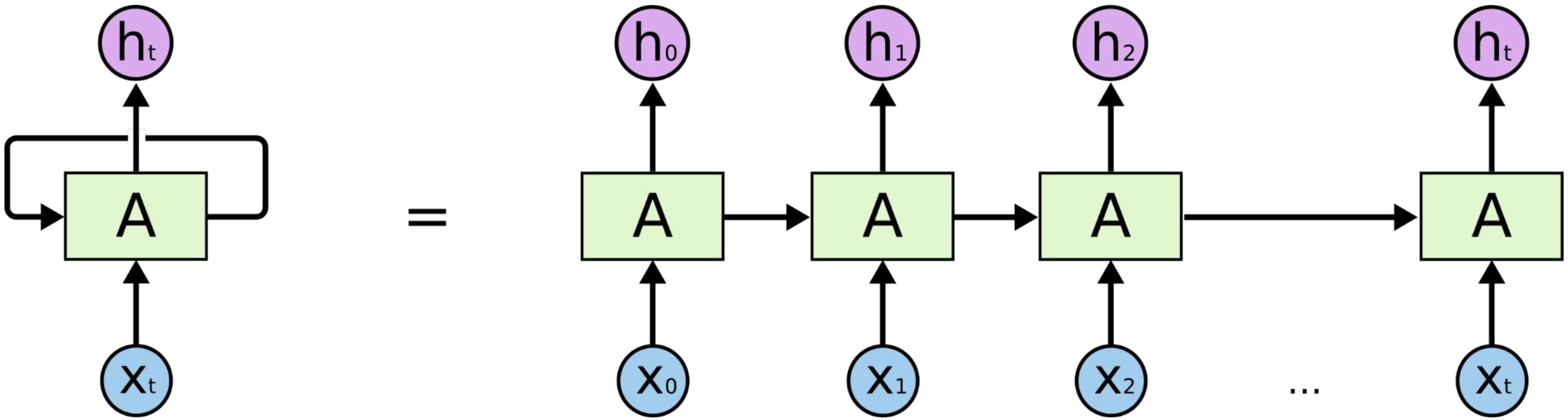

Figure 2: Model of Recurrent Neural Network [14]

RNN: Exploding and Vanishing Gradient Problems

- Neural networks train through gradient descent using backpropagation
- Backpropagation can cause exploding and vanishing gradients

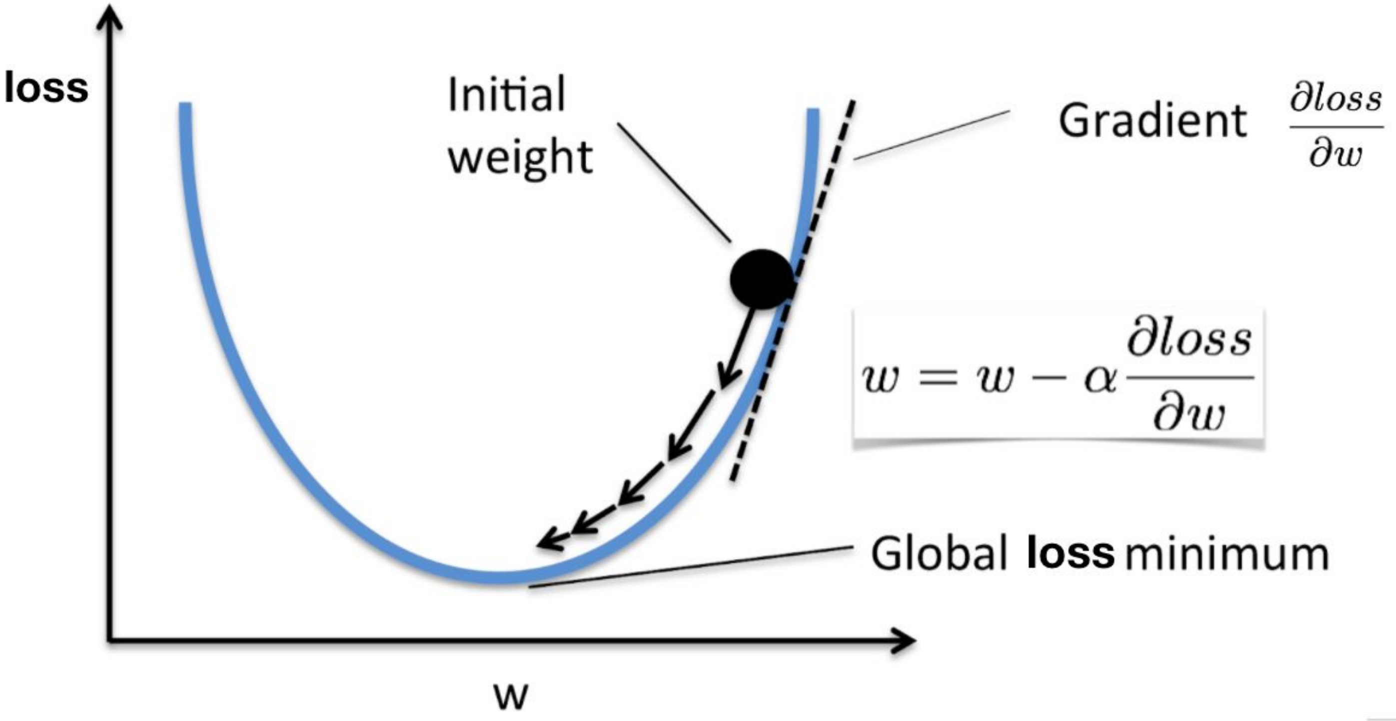

Figure 3: Gradient Descent [15]

LSTM: Gated Cells

- LSTMs solve the exploding and vanishing gradient problems
- Use Gated Cells
  - Update
  - Relevance
  - Forget
  - Output

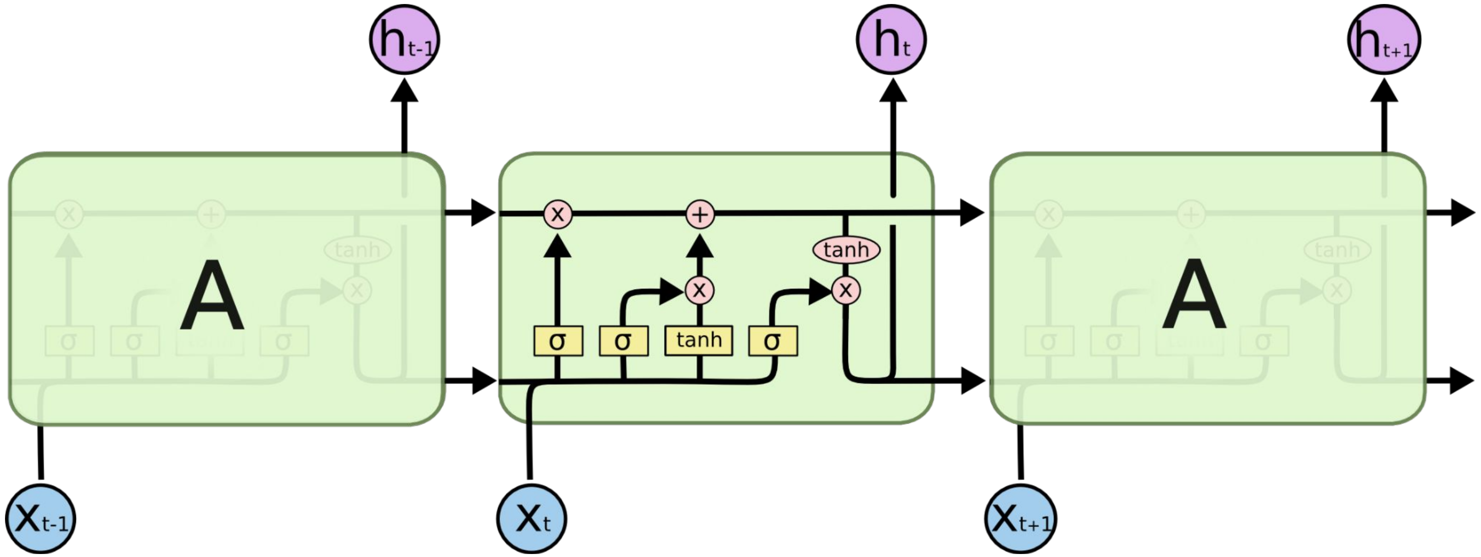

Figure 4: Model of LSTM Cell Gates [14]

Data Sets

- New York City Citi Bike bike-sharing demand data [16]
- National Oceanic and Atmospheric Administration (NOAA) meteorological data [17]
- From May 1st, 2019 to July 31st, 2019

Data Pre-processing Methods

- Hourly bike demand determined for each Citi Bike station
- Time-lagged factors
- Citi Bike and NOAA datasets combined
- Feature engineering
  - Cyclical temporal features
  - Weekday or weekend
  - a.m. or p.m. peak periods

Table 2: Pre-processed Data Set	
Feature Type	**Features**
Hourly bike demand	num_trips, start_datetime, start_station_id, start_station_longitude, start_station_latitude
Time-lagged factors	num_trips_1hr, num_trips_2hr, num_trips_3hr, num_trips_4hr, num_trips_5hr, num_trips_6hr, num_trips_24hr, num_trips_48hr, num_trips_week
Temporal factors	day_of_week, day_of_month, month, hour, is_weekend, hour_sin, hour_cos, day_of_week_sin, day_of_week_cos, day_of_month_sin, day_of_month_cos, month_sin, month_cos, is_am_peak, is_pm_peak

Evaluation Metrics

- RMSE used as evaluation metric during training
- Final models compared on RMSE, MAE, R^2
- MAPE not used due to high rate of records with station-level hourly bike demand of 0

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(x_i - \widehat{x_i}\right)^2}$$

$$MAE = \frac{1}{n}\Sigma_{i=1}^{n}\left|x_i - \widehat{x_i}\right|$$

$$R^2 = \frac{\Sigma_{i=1}^{n}\left(\hat{x}_i - \bar{x}\right)^2}{\Sigma_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

XGBoost: Sensitivity Analyses

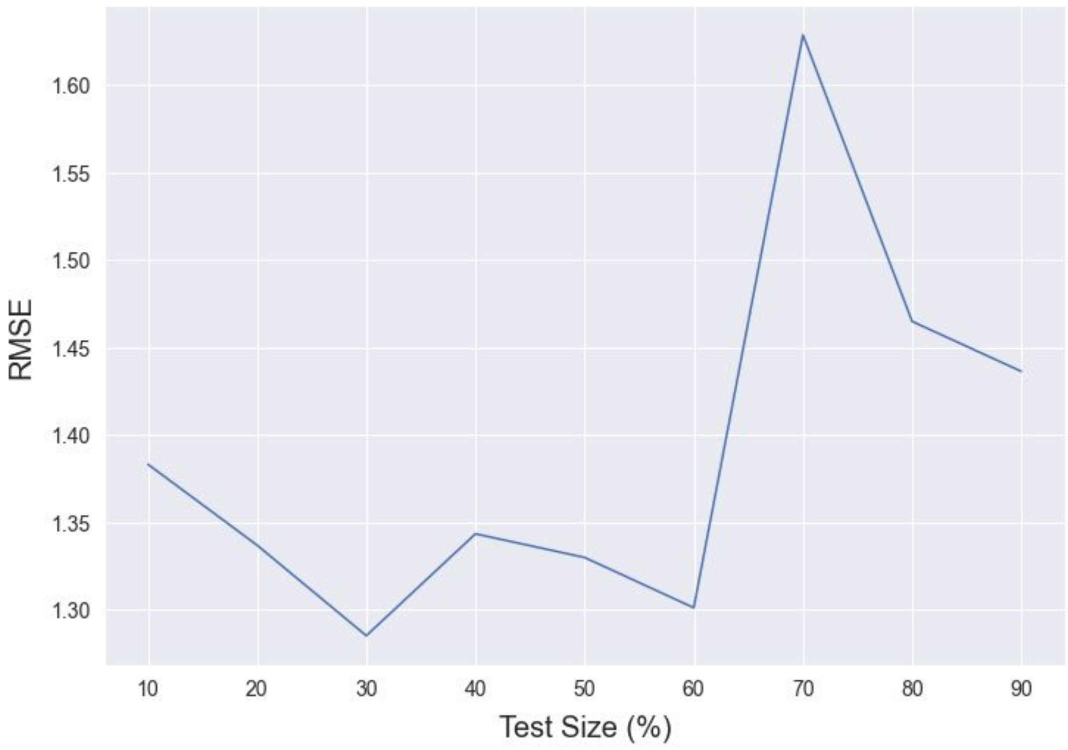

Figure 5: Test Size vs. RMSE

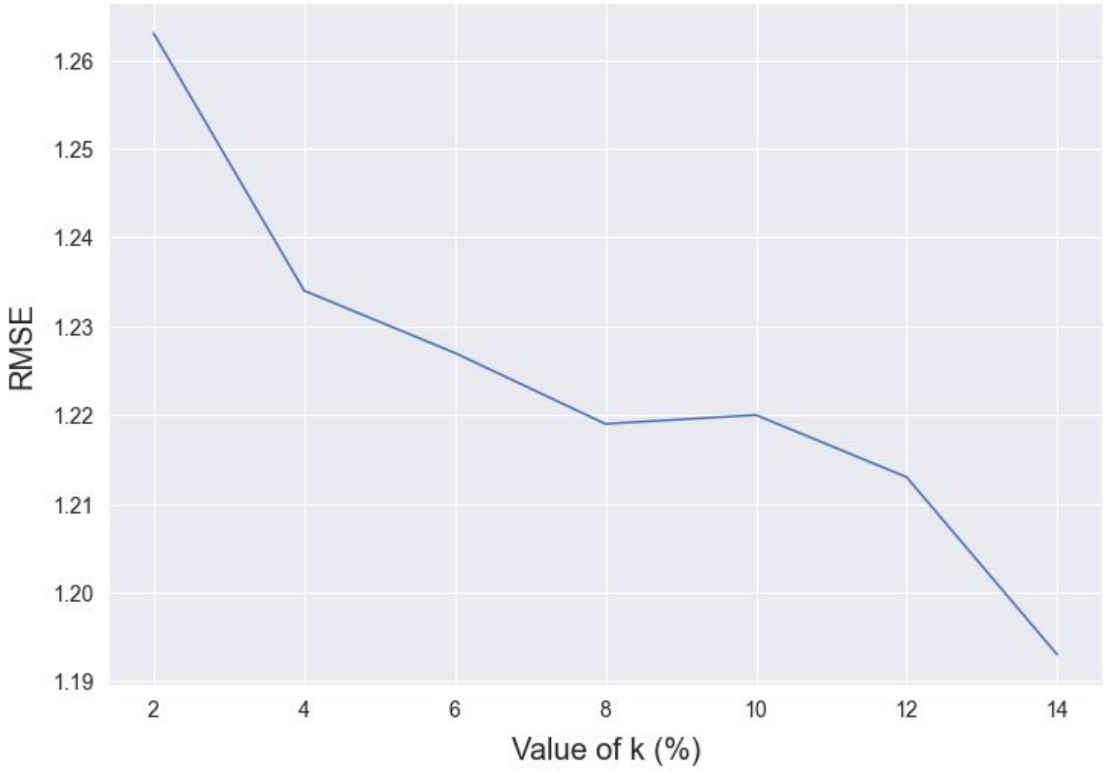

Figure 6: Value of k vs. RMSE

XGBoost: Hyperparameter Selection

Table 3: XGBoost Model Training			
Hyperparameter	**Description**	**Values Tested**	**(Value Chosen: RMSE on Validation Data Set**
max_depth	Maximum depth of each tree	{3, 6, 9}	(6: 1.297)
min_child_weight	Minimum sum of instance weight required in each child	{1, 3, 5, 7}	(7: 1.297)
n_estimators	Number of gradient boosted trees	{10, 25, 50, 100, 200, 400, 500}	(500: 1.190)
learning_rate	Step size shrinkage used in each boosting step	{0.0001, 0.001, 0.01, 0.1, 0.2, 0.3}	(0.01: 1.190)
subsample	Subsample ratio of training instances	{0.5, 0.6, 0.7, 0.8, 0.9, 1.0}	(0.8: 1.184)
colsample_bytree	Subsample ratio of columns when constructing each tree	{0.5, 0.6, 0.7, 0.8, 0.9, 1.0}	(0.5: 1.170)

LSTM: Model Structure

- Input layer: 1,272 neurons corresponding to the 24 features at each of the 53 bike stations
- Output layer: 53 neurons corresponding to the 53 bike stations

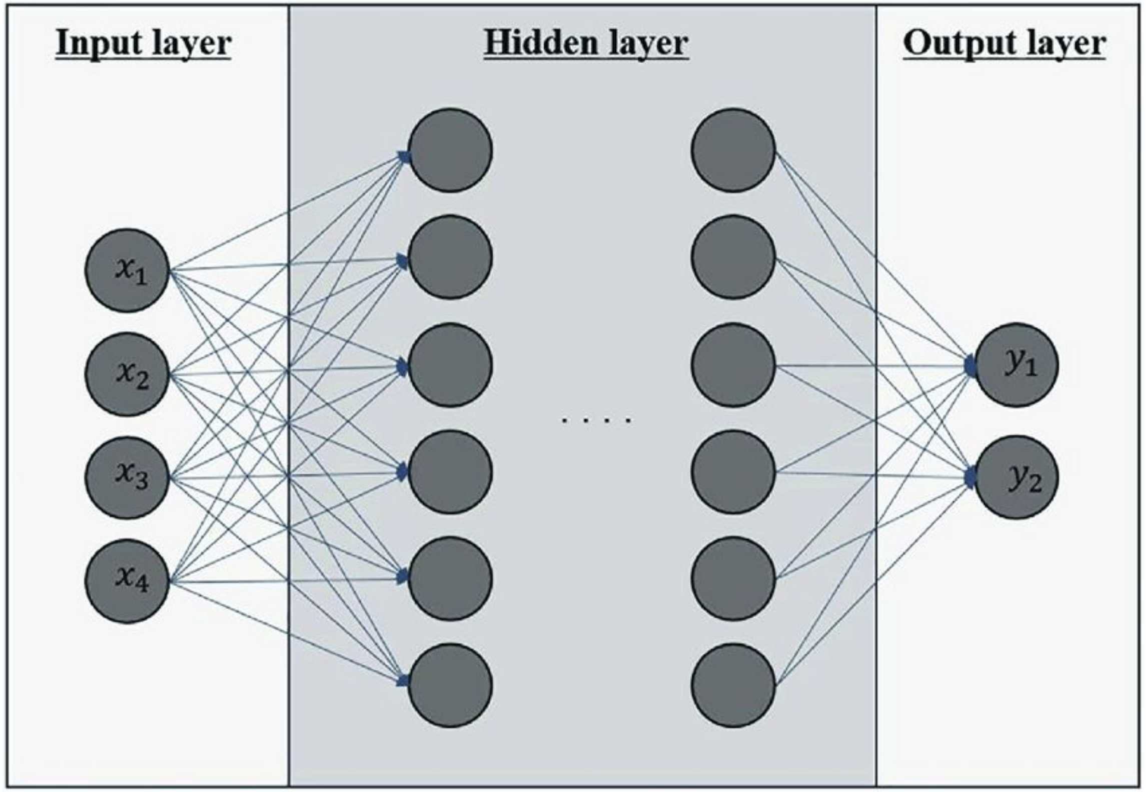

Figure 7: Neural Network Structure [18]

LSTM: Input Data Structure

- (samples x timesteps x features)
- Samples and timesteps dependent on number of timesteps chosen
- 1,272 features corresponding to the 24 input variables for the 53 bike stations

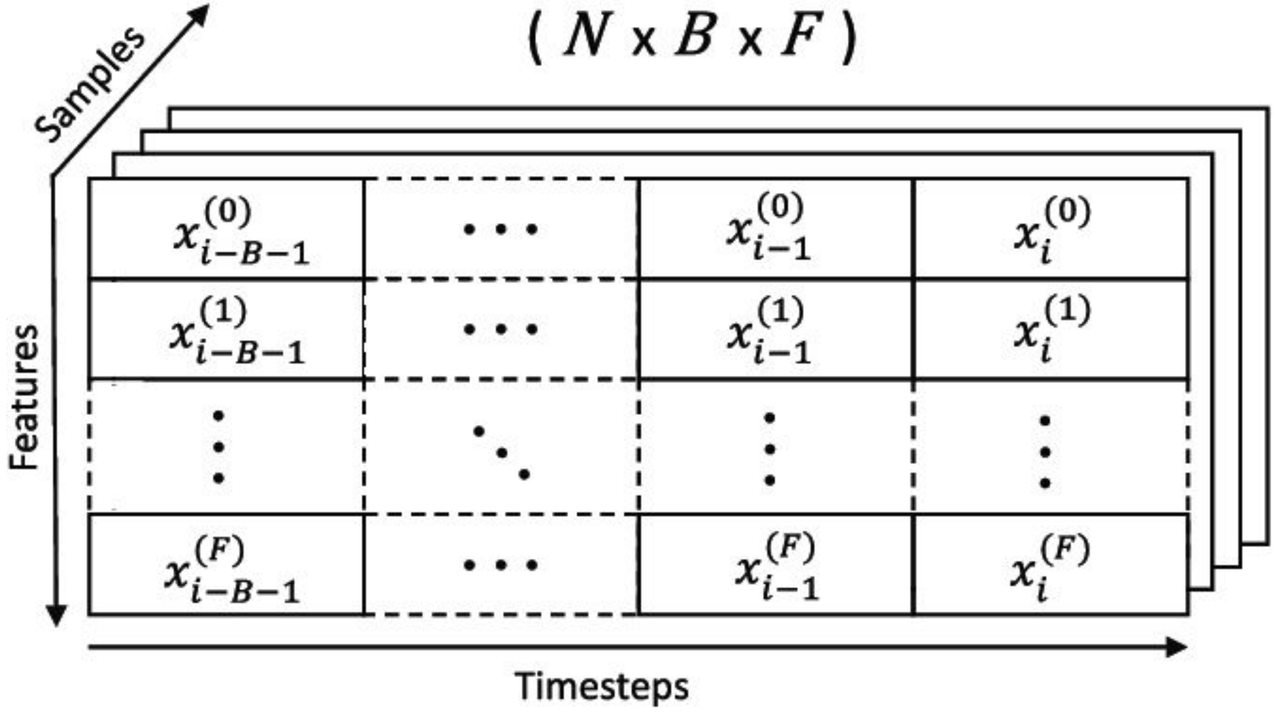

Figure 8: LSTM Input Data Structure [19]

LSTM: Sensitivity Analyses

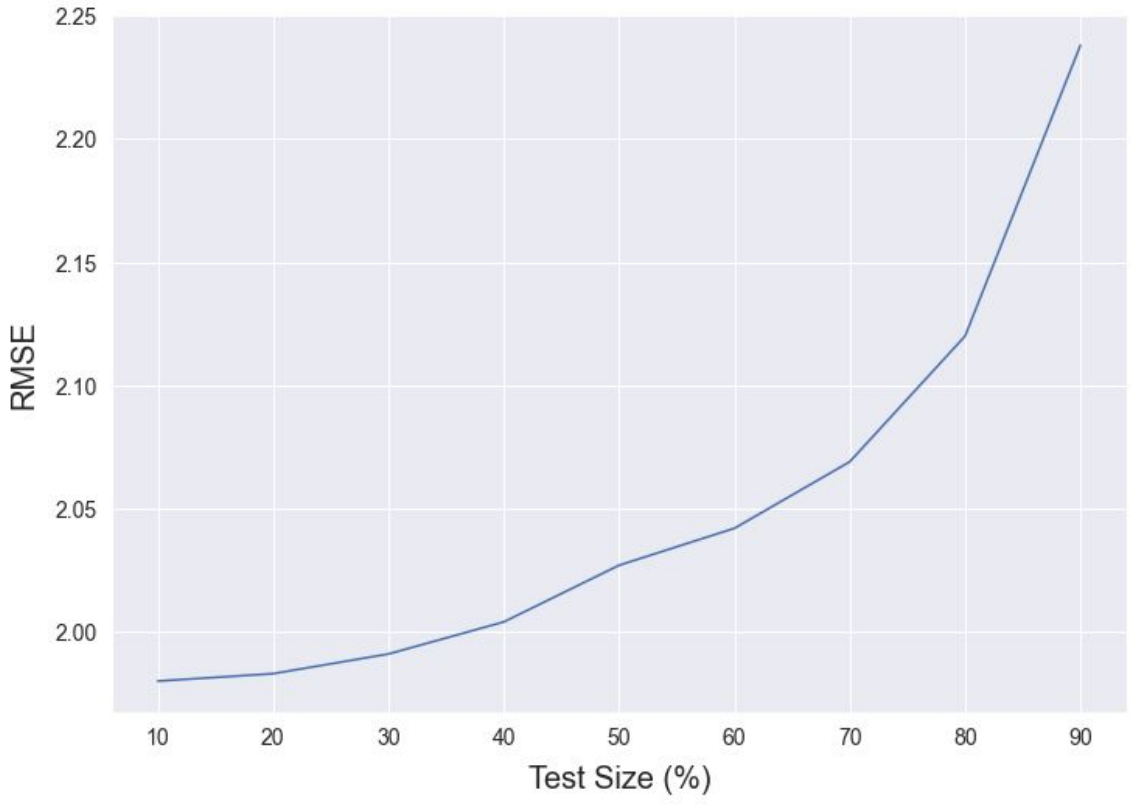

Figure 9: Test Size vs. RMSE

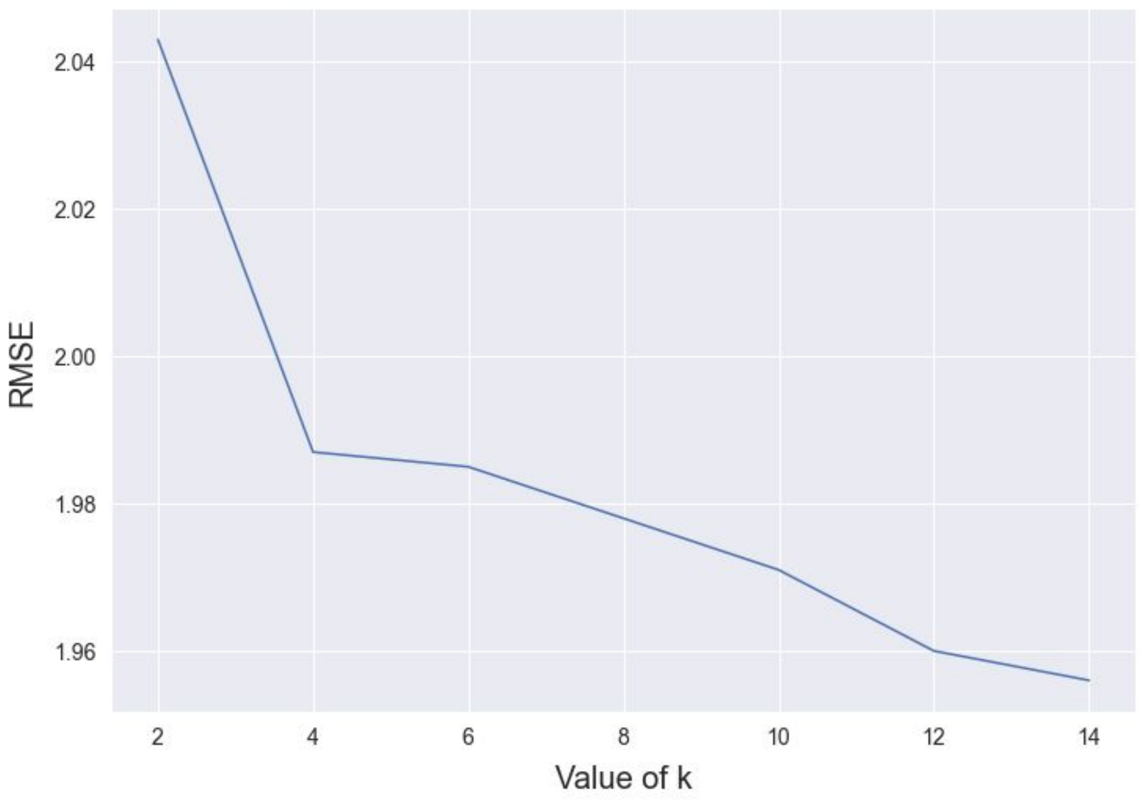

Figure 10: Value of k vs. RMSE

LSTM Training: Parameter Selection

Table 4: LSTM Model Training			
Parameter	**Description**	**(Values Tested: RMSE on Validation Data Set)**	**Value Chosen**
Data Scaler	Standardization or Normalization of data	{(RobustScaler: 3.019), (StandardScaler: 2.775), (MinMaxScaler: 2.777), (PowerTransformer: 5.005), (Unscaled: 2.390)}	Unscaled
Timesteps	Number of timesteps in each sample	{(1: 2.034), (3: 2.024), (6: 2.017), (10: 2.013), (24: 2.002), (48: 1.997), (72: 1.997), (168: 2.050), (336: 2.040)}	72
Hidden Nodes	Number of hidden nodes	{(53: 2.332), (250: 2.112), (500: 2.033), (850: 1.997), (1037: 1.990), (1272: 1.976), (1484: 1.972)}	1484
Layers	Number of LSTM layers	{(1: 1.979), (2: 1.975), (3: 2.089)}	1
Unidirectional vs Bidirectional Layer	Unidirectional or Bidirectional LSTM layers	{(Unidirectional: 1.975), (Bidirectional: 1.958)}	Bidirectional
Activation Function	Mathematical function that transforms inputs to outputs	{(sigmoid: 1.954), (tanh: 1.967)}	sigmoid
Optimizer	Controls how network weights are updated during training	{(adam: 1.955), (Adadelta: 2.564), (SGD: 2.217)}	adam
Number of Epochs	Number of times the model processes the input data	{(5: 2.128), (10: 2.129), (20: 2.129), (50: 2.129), (100: 2.131)}	0.5

Results and Comparison

- XGBoost and LSTM models used to predict bike demand on the test data set
- XGBoost outperforms LSTM on RMSE, MAE, and R^2

Table 5: XGBoost and LSTM Model Results			
Model	**RMSE**	**MAE**	**R-Squared**
XGBoost	1.31	0.52	0.75
LSTM	2.33	1.06	0.26

Results and Comparison: Actual vs Predicted Demand

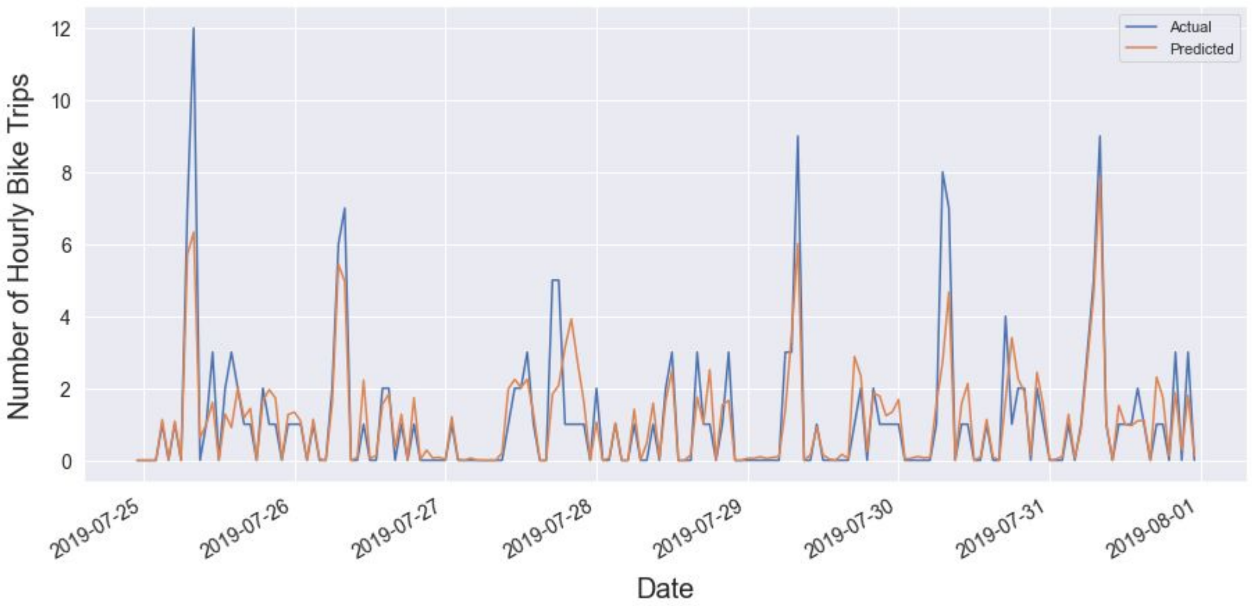

Figure 7: XGBoost Actual vs. Predicted Bike Demand

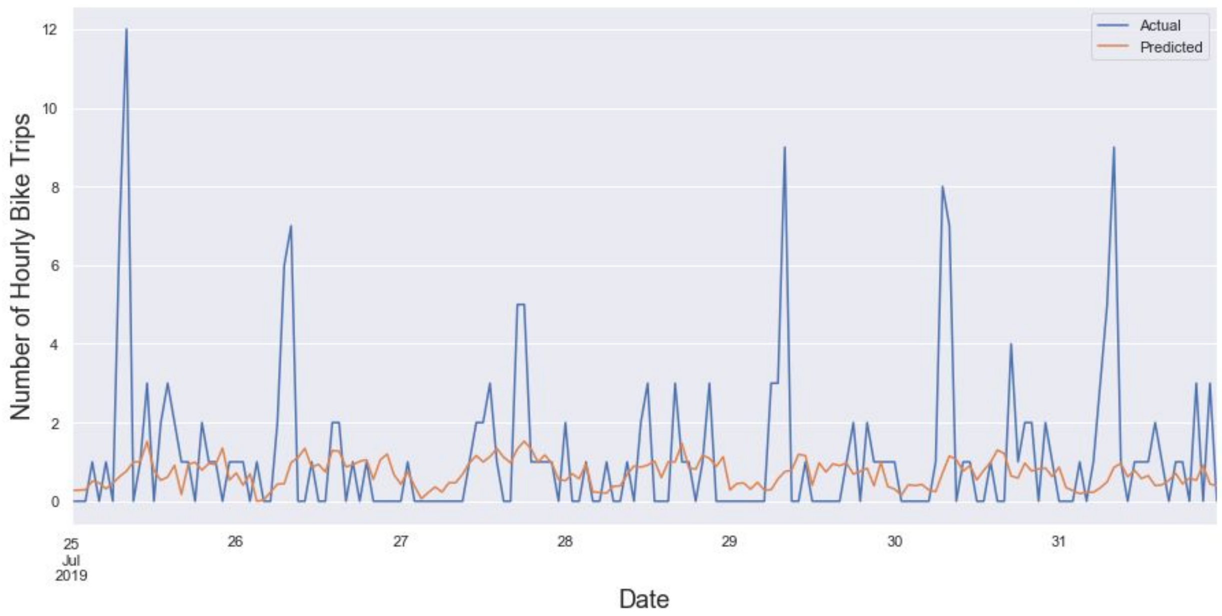

Figure 8: XGBoost Actual vs. Predicted Bike Demand

Results and Comparison: Absolute Error Distribution

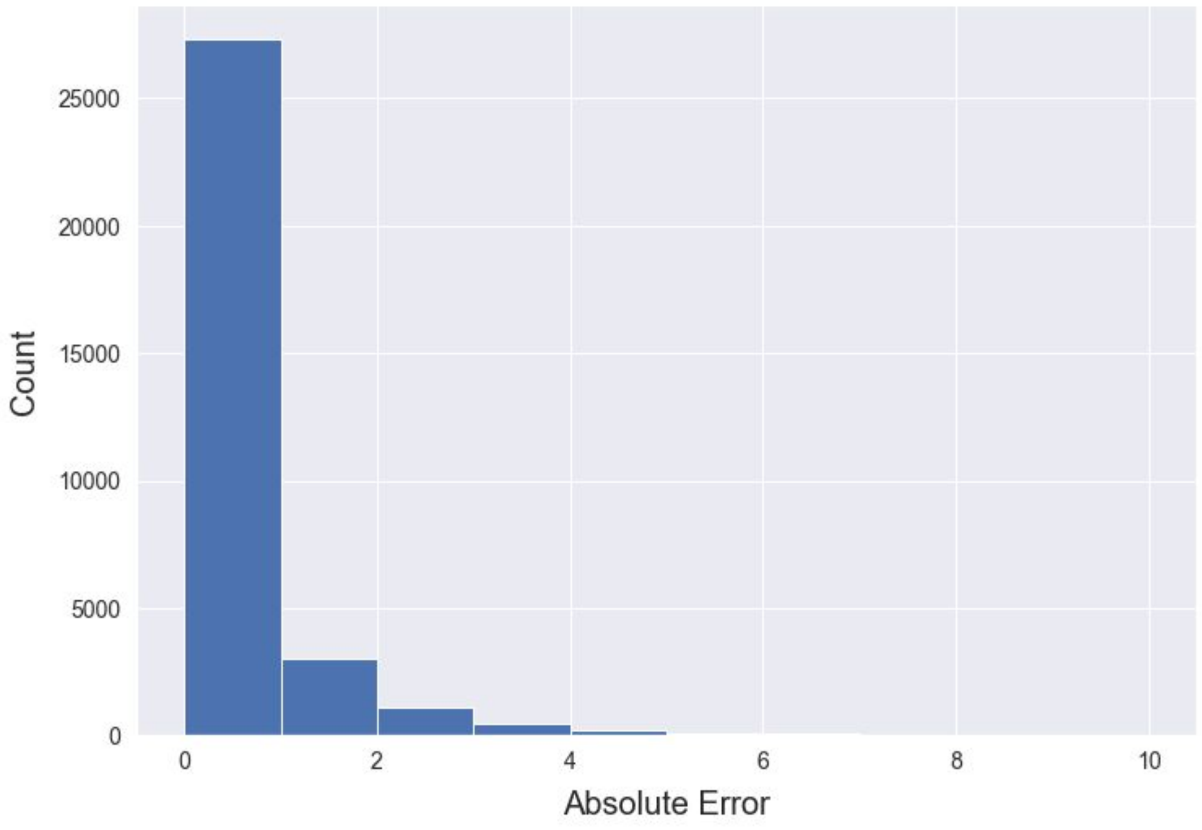

Figure 9: Histogram of XGBoost Absolute Error

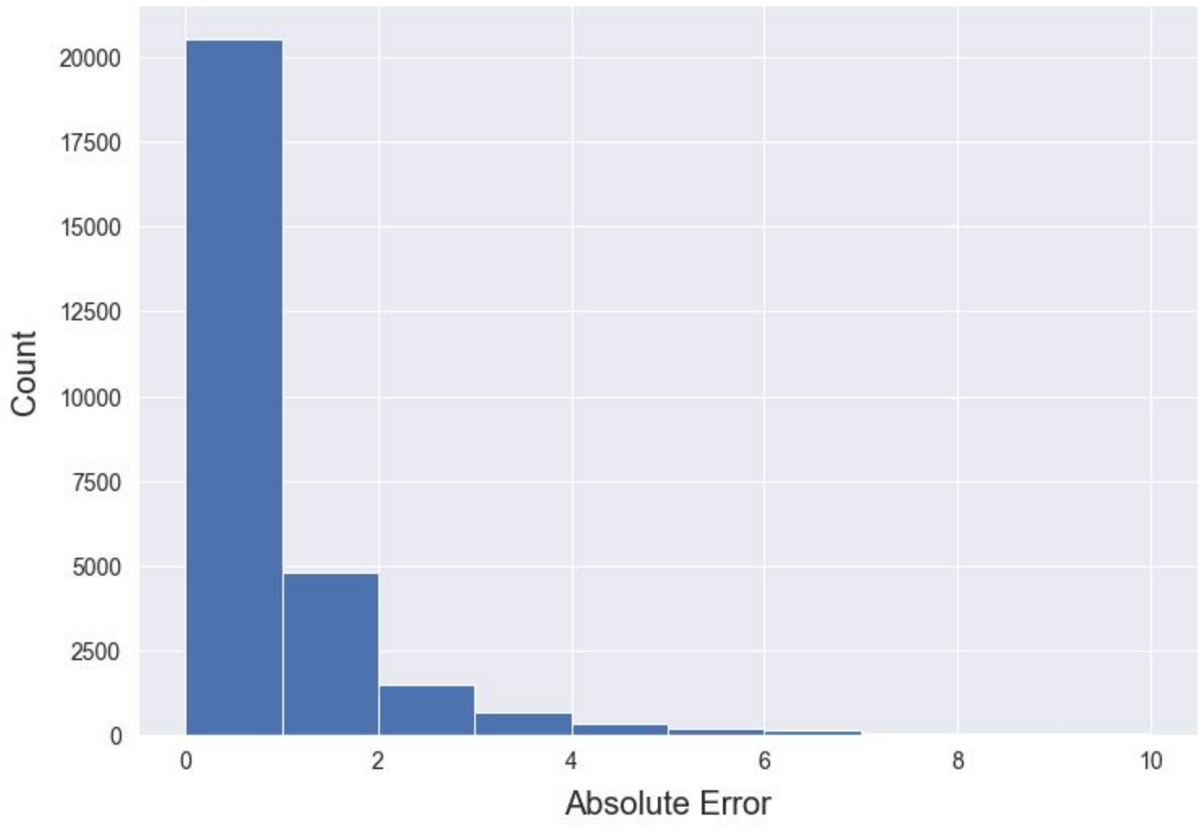

Figure 10: Histogram of LTM Absolute Error

Limitations and Future Research Directions

- Limitations
    - Data sets
    - Models predict one hour into the future
    - Models do not account for spatial dependencies
- Future Research Directions
    - Data mining techniques to improve data sets, especially for large, irregular events
    - Multi-step time forecasting
    - GCNN to model spatial dependencies

Photo Credit: Tupungato, istockphoto.com

References

[1] United Nations, Department of Economic and Social Affairs, Population Division (2019). *World Population Prospects 2019: Highlights. ST/ESA/SER.A/423.* [Online]. Available: https://www.un.org/development/desa/en/news/population/world-population-prospects-2019.html.

[2] United Nations, Department of Economic and Social Affairs, Population Division (2019). World Urbanization Prospects: The 2018 Revision (ST/ESA/SER.A/420). New York: United Nations. [Online]. Available: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html.

[3] P. Demaio, "Bike-sharing: History, Impacts, Models of Provision, and Future," Journal of Public Transportation, 12 (4): 41-56, 2009. [Online]. Available: http://doi.org/10.5038/2375-0901.12.4.3.

[4] United Nations, Department of Economic and Social Affairs, Population Division (2019). Bike Sharing Systems. [Online]. Available: https://sustainabledevelopment.un.org/content/documents/4803Bike%20Sharing%20UN%20DESA.pdf.

[5] Zhang, Yongping and Mi, Zhifu, "Environmental Benefits of bike sharing: A big data-based analysis," *Applied Energy.,* vol. 220, June 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261918304392.

[6] Alliance for Biking & Walking, "Bicycling and Walking in the United States: 2010 Benchmarking Report," in Transportation Research Board Weekly, Washington, DC, 2010, [Online]. Available: https://trid.trb.org/view/914503.

References

[7] Ashqar et al, “Modeling Bike Availability in a Bike-Sharing System Using Machine Learning”, Workshops at the 29th AAAI Conference on Artificial Intelligence, 2015 [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.720.4385&rep=rep1&type=pdf

[8] S. -H. Choi and M. -K. Han, "The Empirical Evaluation of Models Predicting Bike Sharing Demand," 2020 International Conference on Information and Communication Technology Convergence (ICTC), 2020, pp. 1560-1562, doi: 10.1109/ICTC49870.2020.9289176 [Online]. Available: https://ieeexplore.ieee.org/document/9289176

[9] Y. Yang, A. Heppenstall, A. Turner, and A. Comber, “Using Graph Structural Information about Flows to Enhance Short-Term Demand Prediction,” Computers, Environment, and Urban Systems, 2020, vol. 83. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0198971520302544

[10] L. Lin, W. Li, and S. Peeta, “Predicting Station-Level Bike-Sharing Demand Using Graph Convolutional Neural Network,” Cornell University Electrical Engineering and Systems Science, 2020. [Online]. Available: https://arxiv.org/abs/2004.08723

[11] D. Singhvi, S. Singhvi, P. Frazier, S. Henderson, E. Mahony, D. Shmoys, D. Woodard, “Predicting Bike Usage for New York City’s Bike Sharing System,” Workshops at the 29th AAAI Conference on Artificial Intelligence, 2015. [Online]. Available: https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/viewPaper/10115

[12] Sathishkumar V E, Jangwoo Park, Yongyun Cho, “Using Data Mining Techniques for Bike Sharing Demand Prediction in Metropolitan City,” Computer Communications, vol. 153 pp. 353-366, 2020. [Online]. Available: https://arxiv.org/abs/2004.08723

References

[13] A. Rogozhnikov, “Gradient Boosting Explained,” 2021. [Online]. Available: https://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html

[14] C. Olah, “Understanding LSTM Networks,” 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[15] S. Kim, “ML/DL for Everyone with PyTorch, Lecture 3: Gradient Descent,” 2021. [Online]. Available: https://docs.google.com/presentation/d/1CF-vEPzMSkVKnePO__AewGfzmui2aGMr8HhI9diD2t8/edit#slide=id.g27be483e1c_0_18

[16] Citi Bike, “Citi Bike System Data,” 2021. [Online]. Available: https://www.citibikenyc.com/system-data

[17] National Oceanic and Atmospheric Administration, National Centers for Environmental Information, “Data Access,” 2021. [Online]. Available: https://www.ncdc.noaa.gov/data-access

[18] Pyo, Sujin & Lee, Jaewook & Cha, Mincheol & Jang, Huisu, “Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the Korean stock markets,“ PLOS ONE. 12. e0188107. 10.1371/journal.pone.0188107, 2017.

[19] Sala, Simone & Amendola, Alfonso & Leva, S. & Mussetta, Marco & Niccolai, Alessandro & Ogliari, Emanuele. “Comparison of Data-Driven Techniques for Nowcasting Applied to an Industrial-Scale Photovoltaic Plant,” Energies. 12. 10.3390/en12234520, 2019.

Thank you! Any questions?

Photo Credit: Citi Bike, citibikenyc.com