Bike-Sharing in NYC

Optimization in Balance between
Demand Capturing and Cost Deployment

Background

Bike-sharing systems have become popular in urban areas, serving both commuters and tourists efficiently. Understanding the demand dynamics is crucial for optimizing station placement and resource allocation.

Insight Derivative

Building upon the insights gleaned from the report "Traffic Prediction in a Bike-Sharing System" by Yexin Li et al., which was conducted in collaboration with researchers from The Hong Kong University of Science and Technology, Microsoft Research, and Shanghai Jiao Tong University, the project seeks to enhance the sustainability and accessibility of urban transportation systems in NYC.

Methods

To address this issue, we propose to forecast future bicycle check-out and check-in numbers based on historical data and the demand based on meteorological conditions. These predictions improve the efficiency of the bike sharing system by optimizing resource utilization. Using a two-week dataset of the NYC bike-sharing system, we preprocess the data, employ predictive modeling techniques to forecast demand and transition of bikes in the coming week, and aim to optimize bike allocation.

Data Description

Datasets containing records of bicycle trips from New York City's bike-sharing system, obtained from previous research efforts. These datasets encompass essential attributes such as station IDs, trip durations, user demographics, and meteorological data like temperature and weather conditions, crucial for demand prediction. With a total of 473,621 trips recorded at 325 sites over a two-week period, these datasets offer comprehensive insights.

	starttime	day	start_hour	year	weekday	end_hour	tripduration	starttime	stoptime	start station id_x	...	end station id	end station name	end station latitude	end station longitude	bikeid	usertype	birth year	gender	cluster label start	cluster label end
0	2014-08-01 00:00:00	1	0	2014	4	0	1142	8/1/2014 0:00	8/1/2014 0:19	470	...	312	Allen St & E Houston St	40.722055	-73.989111	19117	Subscriber	1969	1	1	2
1	2014-08-04 18:49:00	4	18	2014	0	19	1111	8/4/2014 18:49	8/4/2014 19:08	470	...	312	Allen St & E Houston St	40.722055	-73.989111	16536	Subscriber	1982	1	1	2
2	2014-08-07 18:00:00	7	18	2014	3	18	1092	8/7/2014 18:00	8/7/2014 18:18	470	...	312	Allen St & E Houston St	40.722055	-73.989111	15319	Subscriber	1984	2	1	2
3	2014-08-01 12:26:00	1	12	2014	4	12	320	8/1/2014 12:26	8/1/2014 12:32	236	...	312	Allen St & E Houston St	40.722055	-73.989111	19505	Subscriber	1989	1	2	2
4	2014-08-01 19:43:00	1	19	2014	4	19	249	8/1/2014 19:43	8/1/2014 19:48	236	...	312	Allen St & E Houston St	40.722055	-73.989111	19778	Subscriber	1970	1	2	2

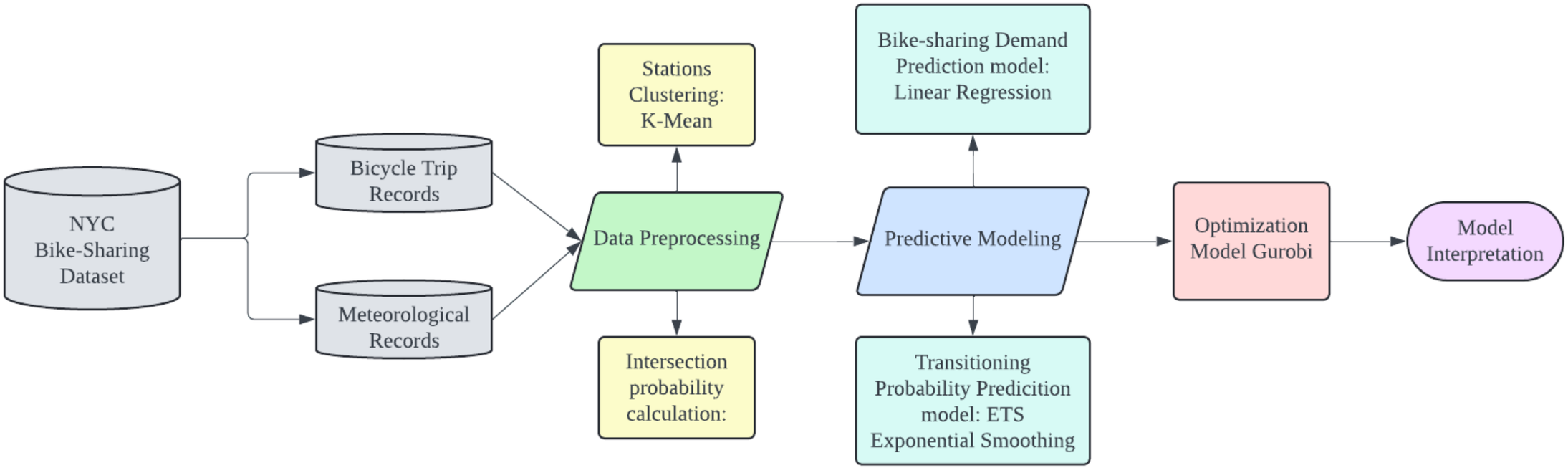
Figure 1: Structures of Dataframes used 1.1: NYC Bike Sharing Transactions Dataframe

	time	Temperature ° F	Wind Speed mph	Weather
0	8-1-2014 0:51 EDT	73.9	3.5	no value
1	8-1-2014 1:51 EDT	73.0	0.0	no value
2	8-1-2014 2:51 EDT	72.0	3.5	no value
3	8-1-2014 3:51 EDT	72.0	3.5	no value
4	8-1-2014 4:51 EDT	71.1	4.6	no value
...	...	...	...	...
444	8-15-2014 19:51 EDT	69.1	9.2	no value
445	8-15-2014 20:51 EDT	69.1	0.0	no value
446	8-15-2014 21:51 EDT	69.1	4.6	no value
447	8-15-2014 22:51 EDT	68.0	0.0	no value
448	8-15-2014 23:51 EDT	69.1	3.5	no value

449 rows × 4 columns

1.2: Meteorology Dataframe

Bike-sharing System Analysis Workflow



Data Preprocessing

Preprocessing involved assigning hour and day labels to trip records, aggregating data, and filling in missing values through interpolation methods for meteorological data. Additionally, the 'weather' column was cleaned and standardized to include only 'sunny,' 'rainy,' and 'haze' categories for consistency.

K-mean Clustering

Previous research highlights high interdependence among nearby bike-sharing stations, leading to volatile traffic dynamics. While, anomalous events like concerts further complicate demand prediction.

Approach: K-means Clustering:

- Aggregated trip records and created an adjacency matrix.
- Employed K-means clustering (k=3) based on spatial proximity and usage patterns.

Rationale and Benefits:

- Focuses on predictable inter-cluster transitions for robustness.
- Promotes consistency in traffic patterns within clusters.
- Enhances prediction accuracy and provides stable city-level measures.

Operational Framework:

- Utilizes a two-level hierarchy: geographical coordinates &. adjacency matrix
- Considers both transition patterns and station locations in a weighted approach.

Advantages:

- Ensures accurate prediction of bike-sharing dynamics.
- Forms a basis for further predictive modeling.
- Determination of optimal cluster count enhances method robustness.

Intersection Probability Estimation

Trips classified into in-cluster/out-cluster and within-hour/out-hour rides.

Assumption: Negligible probability of trips lasting over an hour, with out-of-hour rides returned within the following hour.

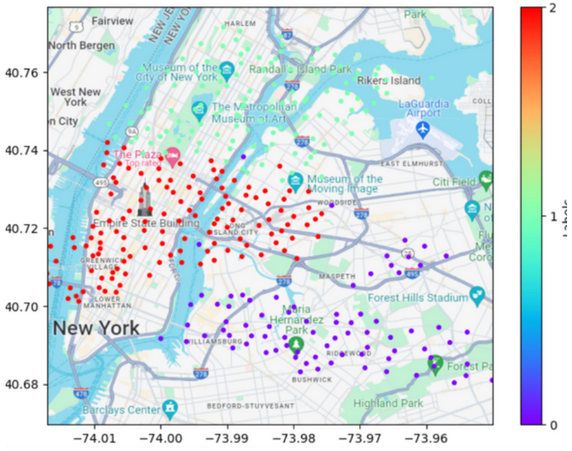


Figure: NYC Map and Station Clusters

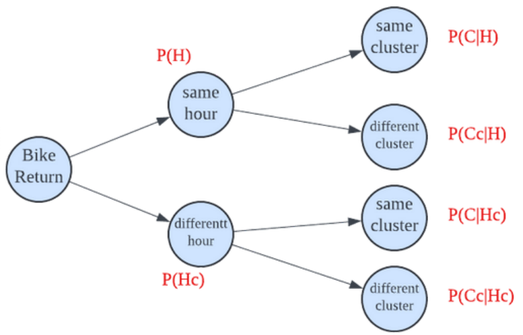


Figure: Tree Diagram

Intersection Probability Estimation (Con't)

Methodology:

- Analysis based on conditional probability and Bayes' rules.
- Data aggregation by 'start cluster label', 'day', and 'hour'.
- Computation of probabilities for in-hour returns and subsequent conditional probabilities.

Bayes' Rule Application:

- Applied Bayes' rules to compute intersection probabilities for event combinations.

Transition Probabilities:

- Used a tree diagram to calculate transition probabilities for various events within and outside clusters.

Application:

- These probabilities served as training data for forecasting trends in the subsequent week.

Methodologies

Predictive Modelling

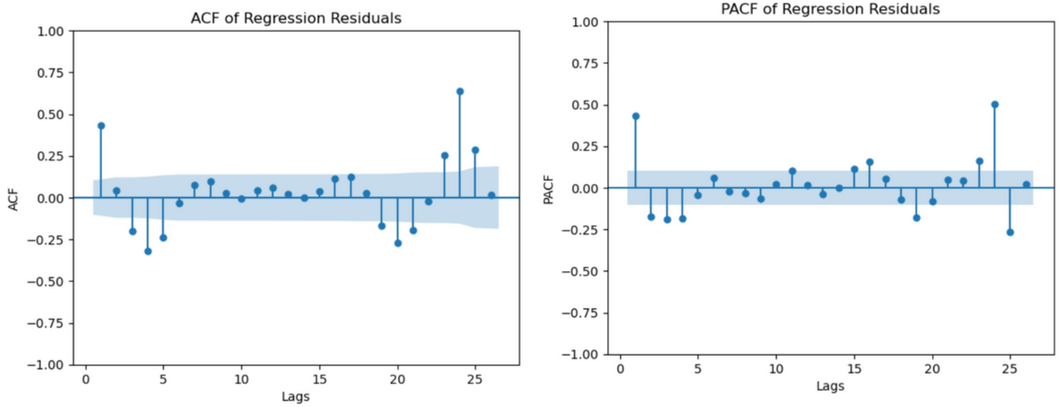
Linear Regression -- Demand Forecasting

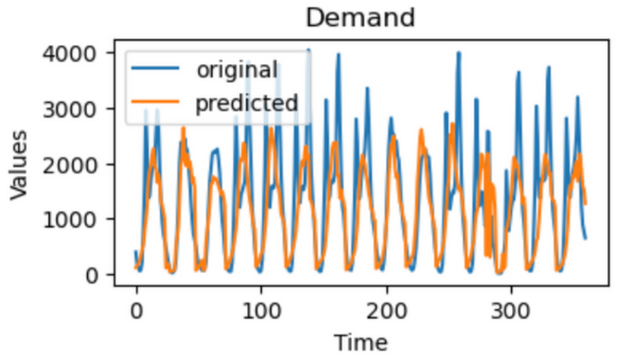
- Considers hour of the day and day of the week for temporal patterns and incorporates meteorological features: weather type, temperature, and wind speed.

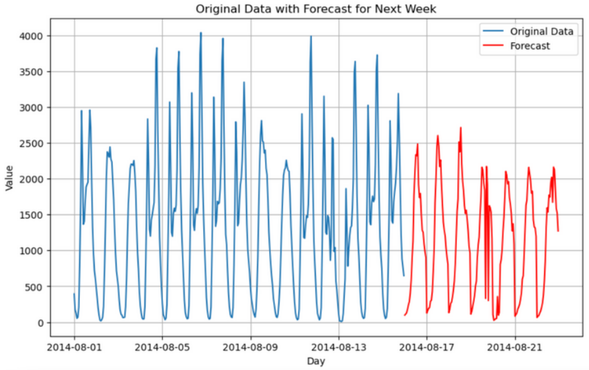
=> Achieves 67.4% variance explained.

=> Utilizes Watson test (Durbin-Watson statistic of 0.781) to assess autocorrelation.

- Persistent seasonality observed in residuals (ACF & PACF), despite attempts to decompose.
- non-stationary residuals.
- Limitations: ARIMA struggles to capture high peak demand accurately.

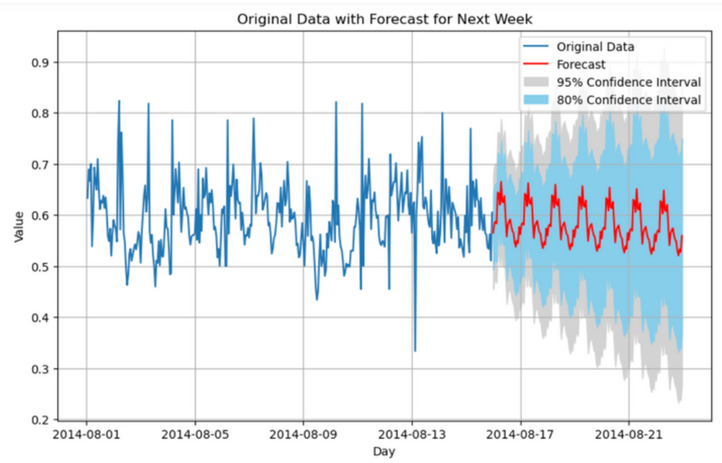






ETS Exponential Smoothing -- Transition Probability Forecasting

- Utilizing ETS Exponential smoothing for predictive modeling to forecast next-week probability of bicycle transitions among clusters and time (check-in/out) .
- Chosen for its capability to capture human behavioral probabilities, past dependencies, and seasonalities, aligning with the nature of bike-sharing systems.
- Application of regularization to enhance model performance, ensuring accurate representation of trends and seasonality.
- Selected model presents good trends and seasonality, making it a suitable choice for the task at hand.



Optimization Model Using Gurobi

Decision Variables

$init\_bike_i$: denotes the number of initial bikes to deploy at Cluster i.
$check\_in_{i,h}$: denotes the number of check-ins to Cluster i in hour h.
$check\_out_{i,h}$: denotes the number of check-outs at Cluster i in hour h.
active_bikei,h: denotes the number of active bikes at Cluster i in hour h.

Constraints

$$1: \ check\_in_{i,0} = \sum_{j=1}^{J} cj\_Ci\_H_0 * check\_out_{j,0}, \ \forall i \neq \forall j$$

$$2: \ check\_in_{i,h} = \sum_{j=1}^{J} cj\_Ci\_H_h * check\_out_{j,h} + \sum_{j=1}^{J} cj\_Ci\_Hc_{h-1} * check\_out_{j,(h-1)}, \ \forall i \neq \forall j \quad (1)$$

$$3: \ check\_in_{i,167} = \sum_{j=1}^{J} (cj\_Ci\_H_{167} + cj\_Ci\_Hc_{167}) * check\_out_{j,167} + \sum_{j=1}^{J} cj\_Ci\_Hc_{166} * check\_out_{j,166}$$

These constraints accurately reflect bikes' inflow based on the outflow patterns observed in the current and preceding hours.

$$active\_bike_{i,h} = init\_bike_i + check\_in_{i,h} - check\_out_{i,h}, \ \boldsymbol{h=0}$$

$$active\_bike_{i,h} = active\_bike_{i,h-1} + check\_in_{i,h} - check\_out_{i,h} \ \forall h > 0 \quad (2)$$

These constraints update active bikes by tracking the number of inflow, outflow, and initial bike deployments.

$$check\_out_{i,h} <= active\_bike_{i,h} \ \forall i \ \forall h,$$

$$check\_out_{i,h} <= demand\_ci_h \ \forall i \ \forall h, \quad (3)$$

Ensuring that bikes checked out do not exceed the active bike or the predicted demand for any period.

$$check\_in_{i,h}, check\_out_{i,h}, init\_bike_i, active\_bike_{i,h} \geq 0 \quad (4)$$

Nonnegativity constraint.

Objective Function

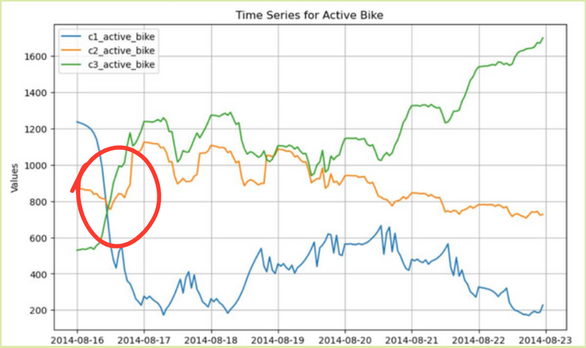
$$min(\sum_{i=1}^{I} \sum_{h=0}^{H} (demand_{i,j} - check\_out_{i,h}) + \lambda * init\_bike_i)) \quad (1)$$

The model's objective seeks to minimize the mismatch between the estimated demand and the number of bikes checked out at each cluster while penalizing attempts to overly increase the number of initial bikes needed, thereby aligning supply with anticipated demand. The optimal $\lambda = 6$ was selected through trials and errors and sensitivity analysis on the magnitude of changes in demand mismatch.

Result & Managerial Insights

Initial allocation strategy for **85.44%** fulfillment of the total demand:

Cluster 1: **1,247** bicycles
Cluster 2: **874** bicycles
Cluster 3: **535** bicycles

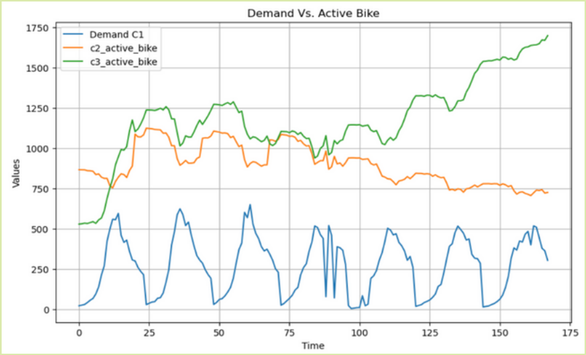


Dynamic Bike & Demand:

Bike redistribution to Clusters 2 and 3 at the beginning of Tuesday led to a temporary shortage in Cluster 1.

Bikes moved to Clusters 2 and 3 tend not to return to Cluster 1

The matching ratio for Cluster 1 is only 43.64%, this can be a result of morning traffic peaks,



Solution:

- reallocating resources (i.e. human intervention by moving bikes from other clusters)
- Innovative programs incentivize users to return bikes to specific areas.

Limitations & Future Implementations

Delve deeper into customer usage patterns
Incorporating meteorological data to refine predictions

Predictive accuracy of the prediction model
Inclusion of Additional Variables such as Regional Variations and Data Collection on Surrounding Infrastructure