

Analyzing & Predicting the pattern of usage of CTA Transportation

By Python Demons Group

Faezehossadat Khademi, Github : Faezeh1900 , fkhade2@uic.edu

Vipul Dhariwal, Github : VipulDhariwal, vdhari3@uic.edu

Piyush Agrawal, 'Github : agrawalpiyush', pagra7@uic.edu

Kasturi Joshi, 'Github: kasturijoshi06', kjoshi27@uic.edu

Akshat Pancholi, 'Github : Loanchip', apanch21@uic.edu



Introduction

CTA is a mass transit in Chicago providing :

- Bus services and
- Train services

But due to the rise in population and movement of commuters

What can CTA do to accommodate the demands of all its commuters and increase the ridership?

What we have done

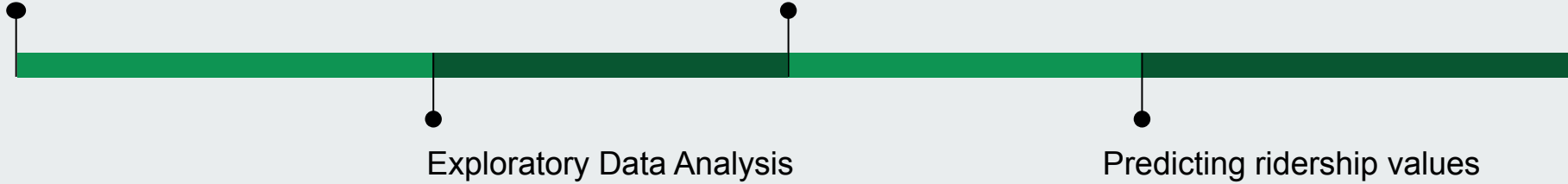
- Explored how the variation in the commuter travel, location and other factors affects the CTA.
- Looked at ways to optimize the CTA transportation and make it prepared for the changes in Commuters' usage of its services.
- Predicted the major changes in ridership to help CTA plan its transportation services efficiently

Outline

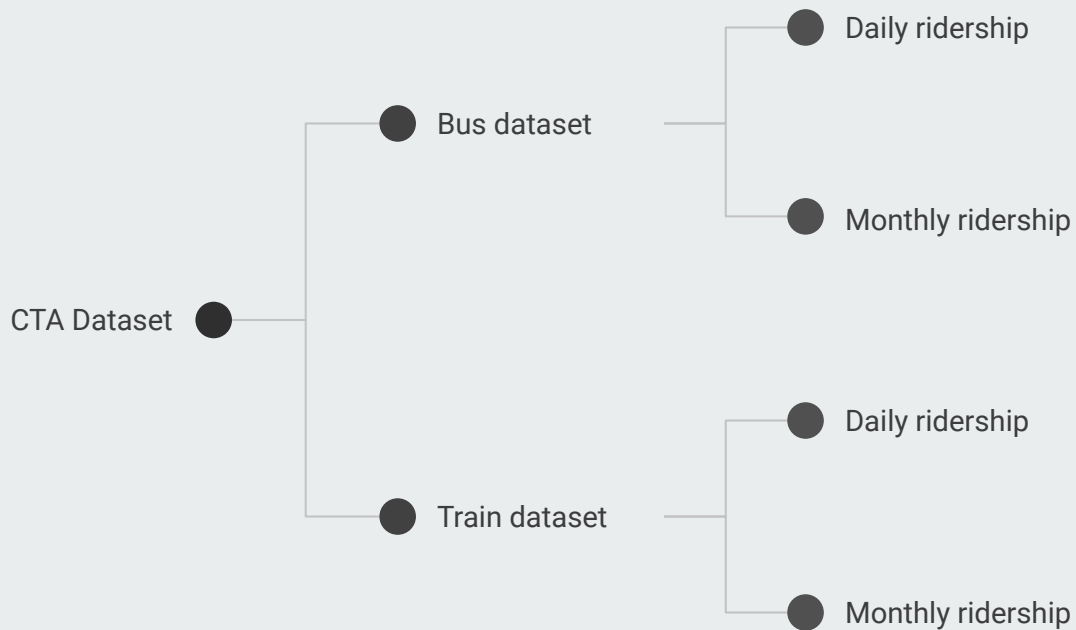


Data Cleaning

Insights and Visualizations



Dataset



Data cleaning



- Separated Year and Month from Date and added columns for them
- Dropped all the NaN & double values
- Converted String into Int in of month total column
- Adjusted all the values for the leap year
- Dropped special routes which did not operate every month of all the years (for Some part of the exploration)
- Dropped Unnecessary Columns
- Separated columns with multiple entries
- Added new column of days by extracting the days
- Grouped the dataset on the basis of 4 features such as route, day, year

Exploratory Data Analysis



Granularity

- The data is grained in monthly fashion for both the bus as well as train datasets
- Data is divided for each bus and each train.

Exploratory Data Analysis



Structure

- Tabular form
- The data is in the form of CSV files
- Bus dataset columns : bus route, route name, average ridership for weekdays, saturday and sundays
- Train dataset columns: Train station, month, average weekday, saturday and sunday ridership

Exploratory Data Analysis

Temporality

- The data we have collected ranges from 1st January 2001 to 1st December 2019

	route	routename	Month_Beginning	Avg_Weekday_Rides	Avg_Saturday_Rides
0	1	Indiana/Hyde Park	01/01/2001	6982.6	0.0
1	2	Hyde Park Express	01/01/2001	1000.0	0.0
2	3	King Drive	01/01/2001	21406.5	13210.7
3	4	Cottage Grove	01/01/2001	22432.2	17994.0
4	6	Jackson Park Express	01/01/2001	18443.0	13088.2
...
31435	172	U. of Chicago/Kenwood	12/01/2019	1435.9	396.8
31436	192	U. of Chicago Hospitals Express	12/01/2019	606.3	0.0
31437	201	Central/Ridge	12/01/2019	1902.8	936.4
31438	206	Evanston Circulator	12/01/2019	507.5	0.0
31439	1001	South Loop Event Route	12/01/2019	104.1	3091.2

Exploratory Data Analysis



Scope

- We have complete data with respect to scope.
- We filtered some of the columns from both the datasets
- Coverage remains the same after filtering

Exploratory Data Analysis



Faithfulness

- Though our data is accurate and gives the idea about ridership, it does not capture the reality completely
- The ridership is counted while boarding, not unboarding. Assumption: Passenger travelling rides the entire journey of the bus/ train
- For trains, count is increased at the station entries. Reality is not captured when a person changes the line internally within a station

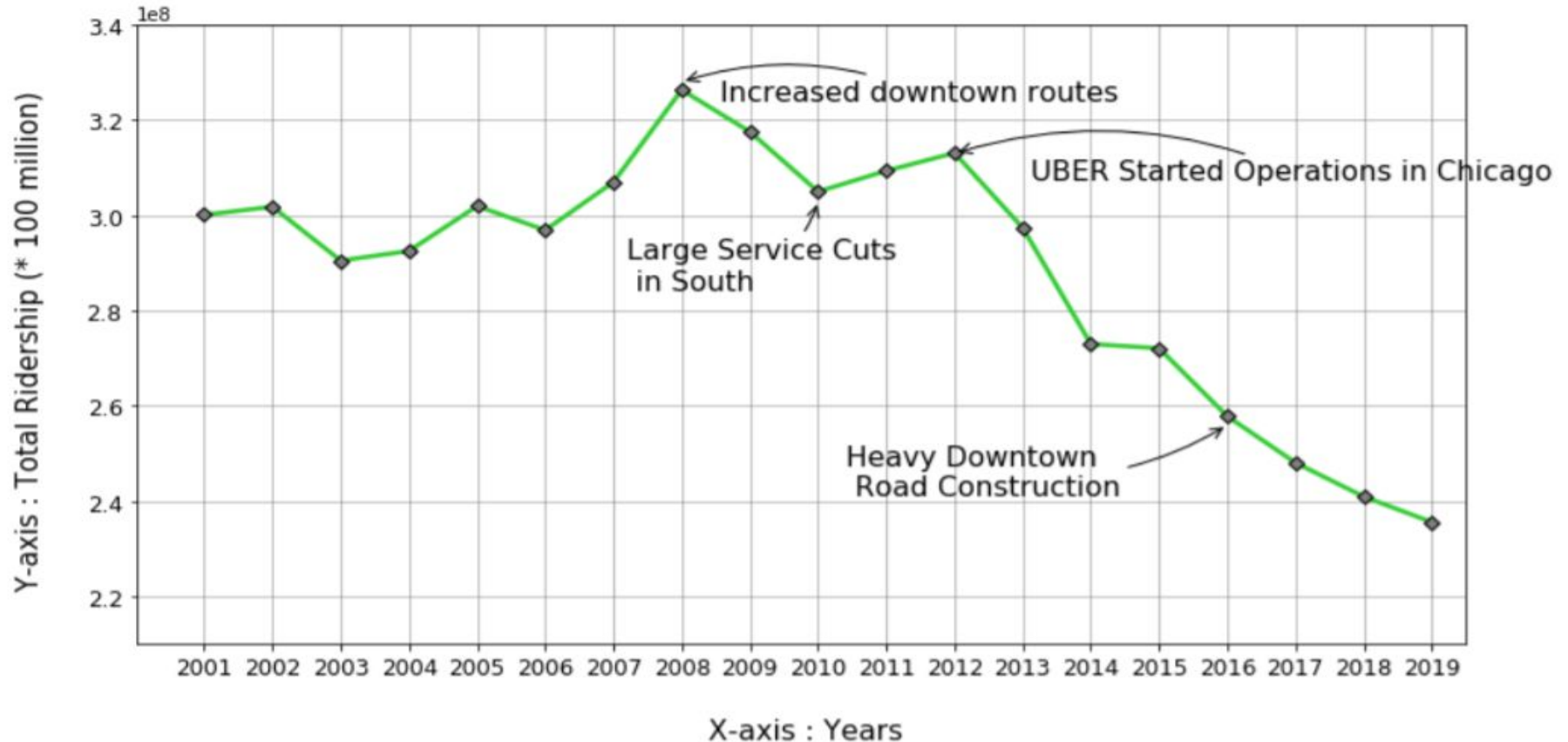
Insights



BUS

- Drastic change (Increase and Decrease in ridership) in bus routes ['169', '1', '51', '2', '108', '171', '11', '28', '120', '100']
- Increase in ridership in the routes from Downtown routes in 2007 - gain in overall ridership
- Stopping services for less popular routes in South region in 2009 led to a significant drop in overall ridership
- Ridership declined steadily since 2012 due to the emergence of cheaper Uber operations like UberX and Uber Share
- Heavy road construction in Downtown - high decrease in ridership in 2016

The Total Ridership of BUS from 2001-2019



Insights

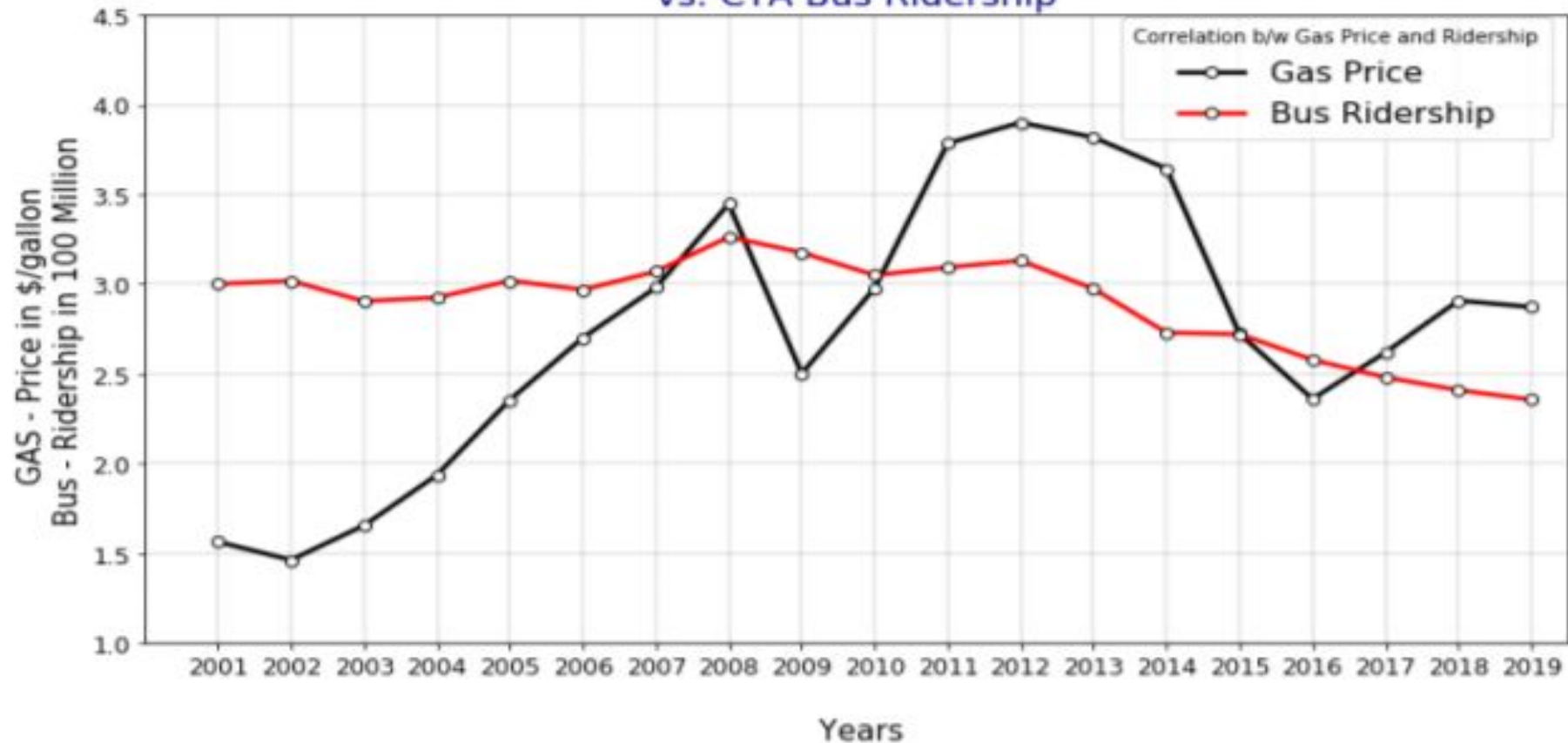


Hypothesis

- ➔ In many of the CTA annual reports, it is stated that the bus ridership declined at some points due to decrease in GAS prices.
- We tested the correlation between the GAS price and total ridership.
- It turned out to be a mere **0.095**.
- Ridership is not correlated to GAS prices

This hypothesis is taken from the official CTA annual reports : <https://www.transitchicago.com/ridership/>

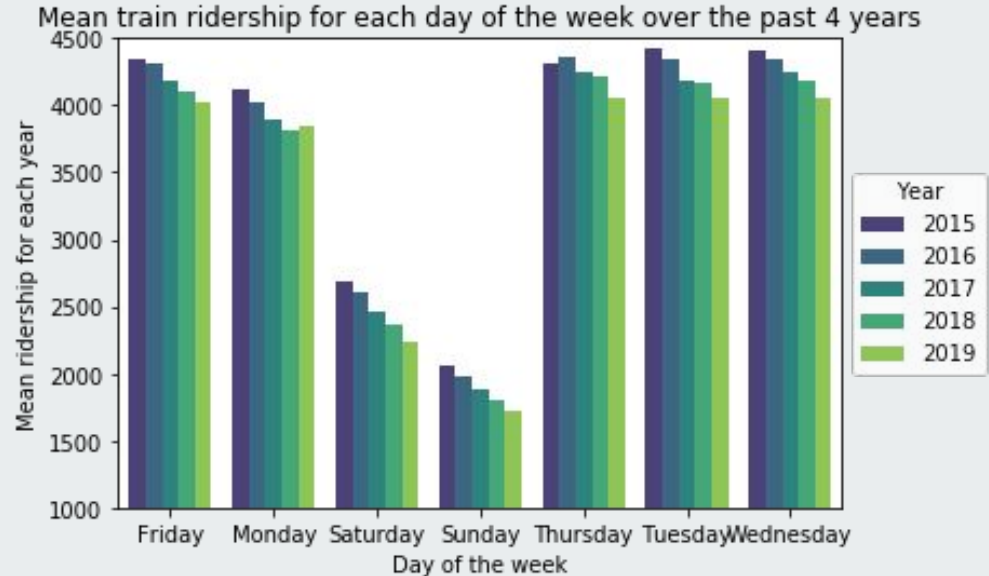
Correlation Between Gas Prices in Chicago vs. CTA Bus Ridership



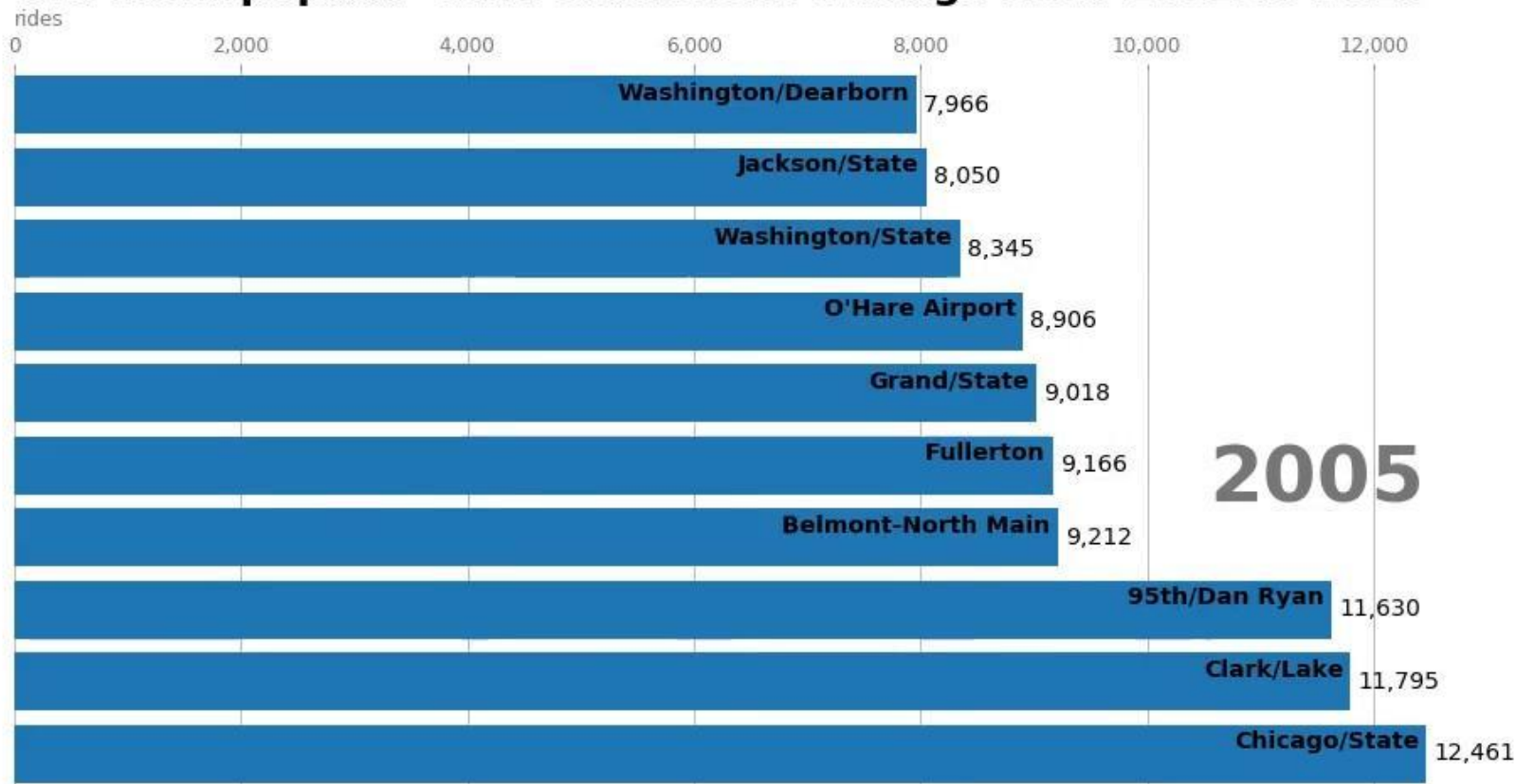
Insights

Train

- The CTA train ridership has plummeted not just on weekends due to office holidays but also on weekdays over the past 4 years



The most popular train stations in Chicago from 2001 to 2019



Popular Train Station race

Insights



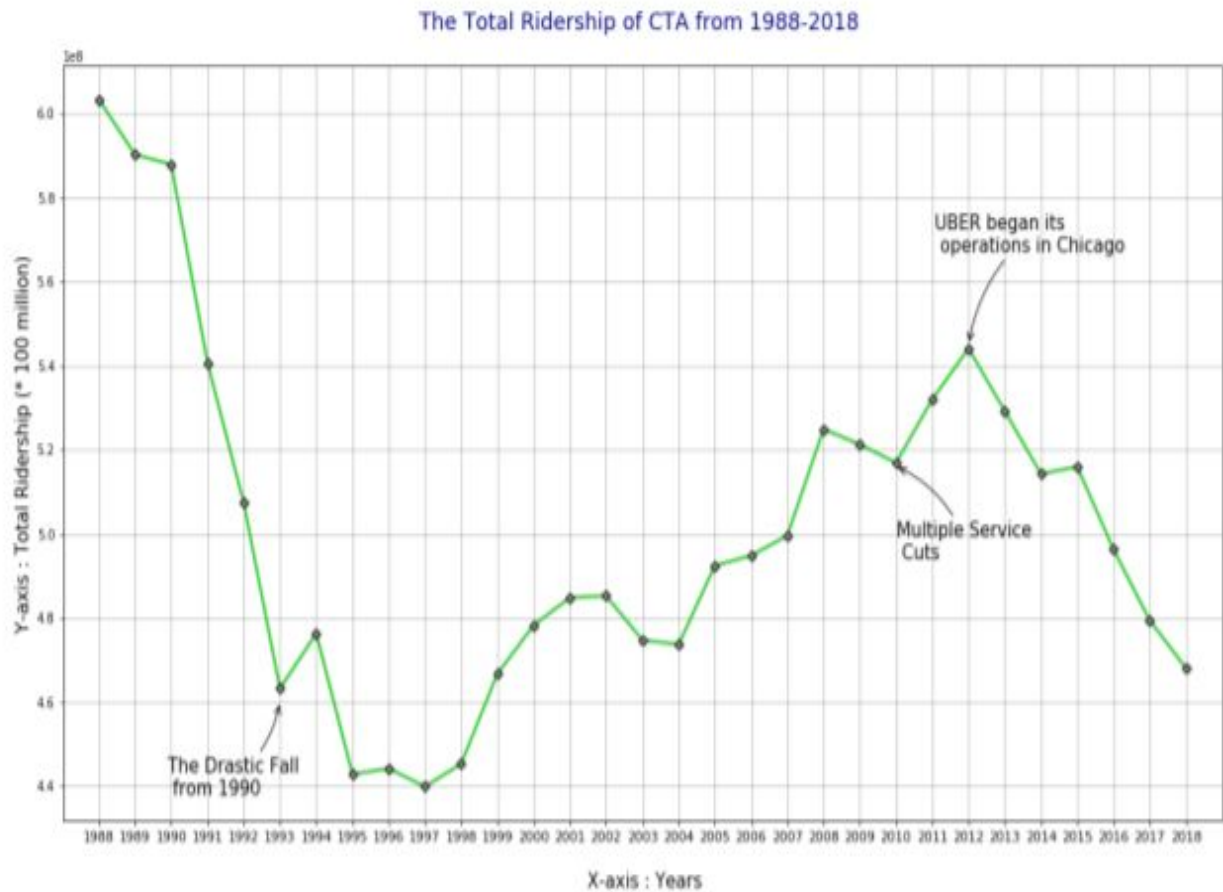
From the train station race

1. The southern train station like 95th/Dan Ryan which was the most popular in 2001 is nowhere in the top 10 in 2019
 - Less popular. Ridership in the 21st century grew in the downtown area - business boom.
 - Population of Chicago has moved from south to the north.
2. Stations like Clark/Lake and Lake/state have the highest ridership since the past 10 years.
 - Boom in the business sector in the loop.
 - All the 8 train lines run on the stations in the loop area. Whereas for the non-loop stations, only maximum 3 lines pass through a single station.
 - Opportunity for commuters to move from one line to the other at the same station.



Total CTA ridership (Train and Bus)

We merged the individual datasets provided by CTA to get a complete picture of the variation in overall Ridership in the CTA services



Predicting the ridership of buses using Recurrent Neural Networks

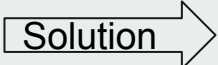


- Used Long short-term memory model (LSTM) to predict the bus ridership for the year 2019
- This ridership is a discrete integer value
- Why LSTM?
 - One of the best models for extracting patterns in long sequences
 - The gated architecture of has the ability to manipulate its memory state thus, they are ideal for time-series analysis
 - Tried support vector regression, linear regression and vector autoregression but they gave a bad accuracy as the features are multivariate time-series features
 - Huge dataset of 60,000 rows

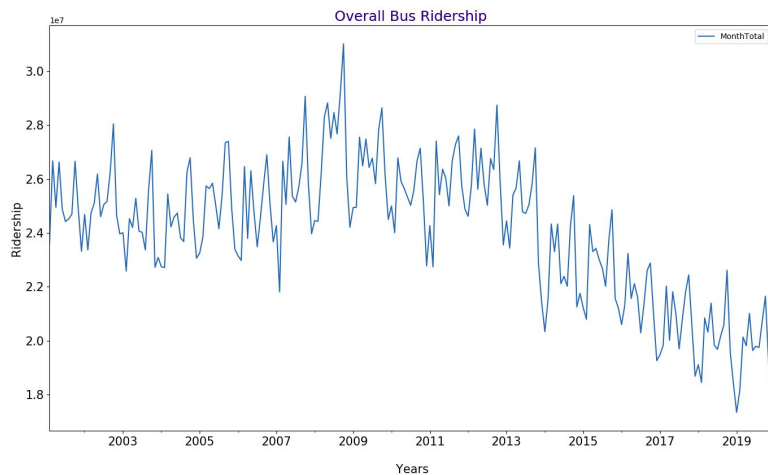
Bus Ridership Predictions



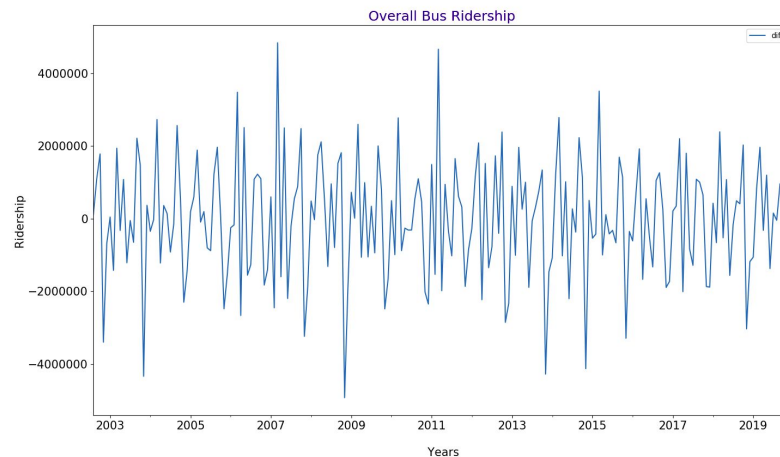
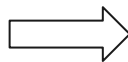
Pre-processing for LSTM:

- Dropped the routes varied drastically- to reduce the noise.
- Unstable data  Took difference of each month by next month. Obtained stability

Pre-processing for LSTM



Unstable data



Stable data

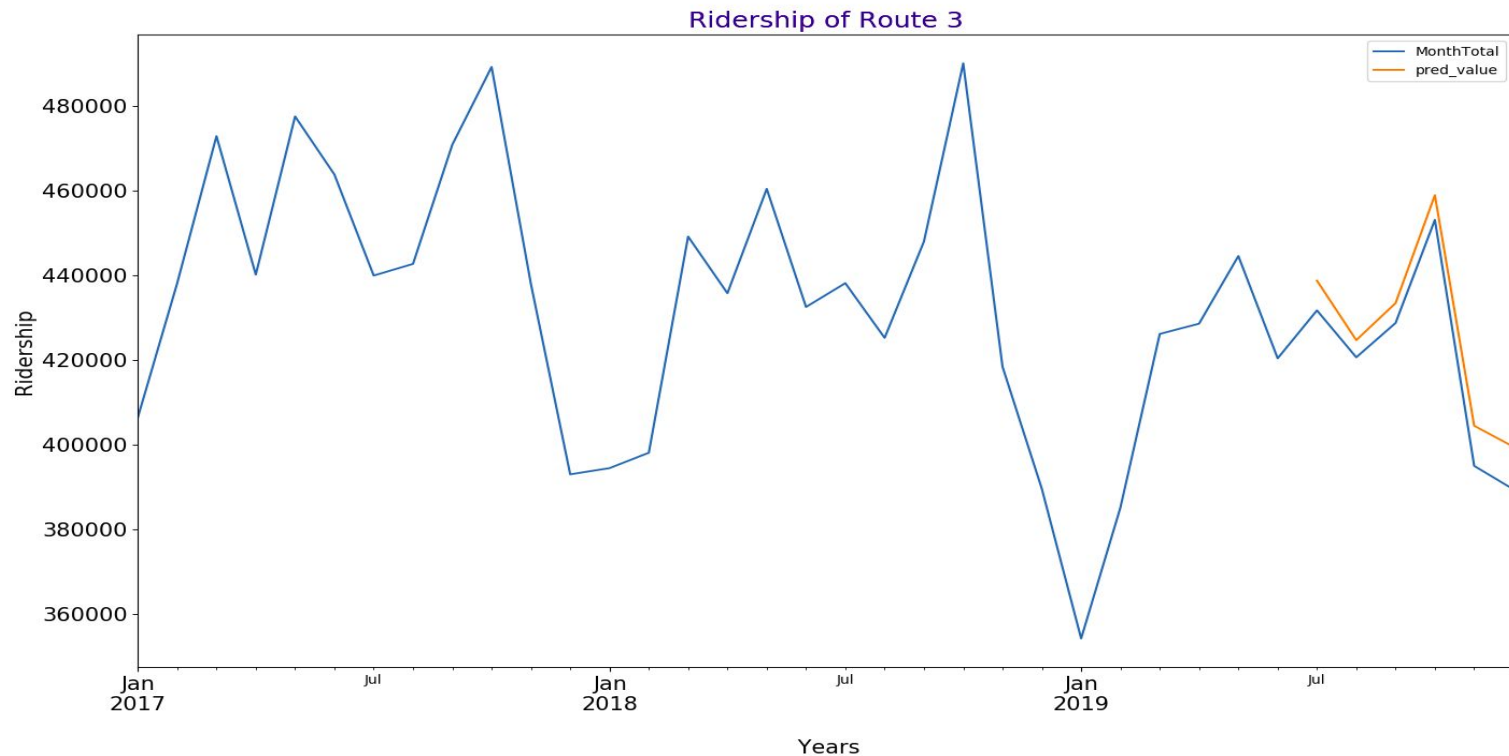
Bus Ridership Prediction- LSTM



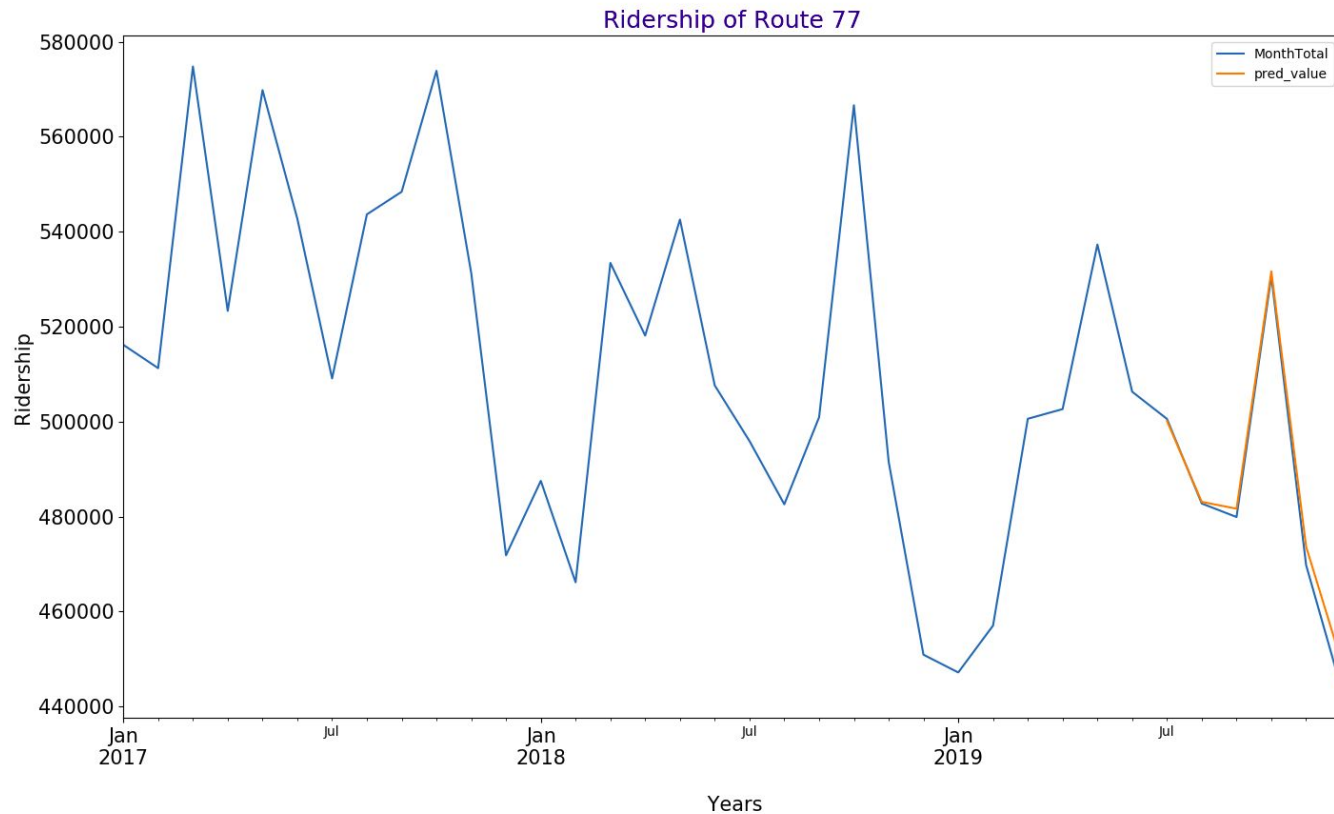
Parameter tuning:

1. Added 18 lags which are essentially difference of the difference
2. Checked our regression coefficient - good
3. But, set the threshold at .65 in order to predict the ridership value for routes which have higher confidence.
4. Discarded the prediction for routes which have accuracy less than 80% (Important - CTA keeps changing the routes of buses. This adds noise)

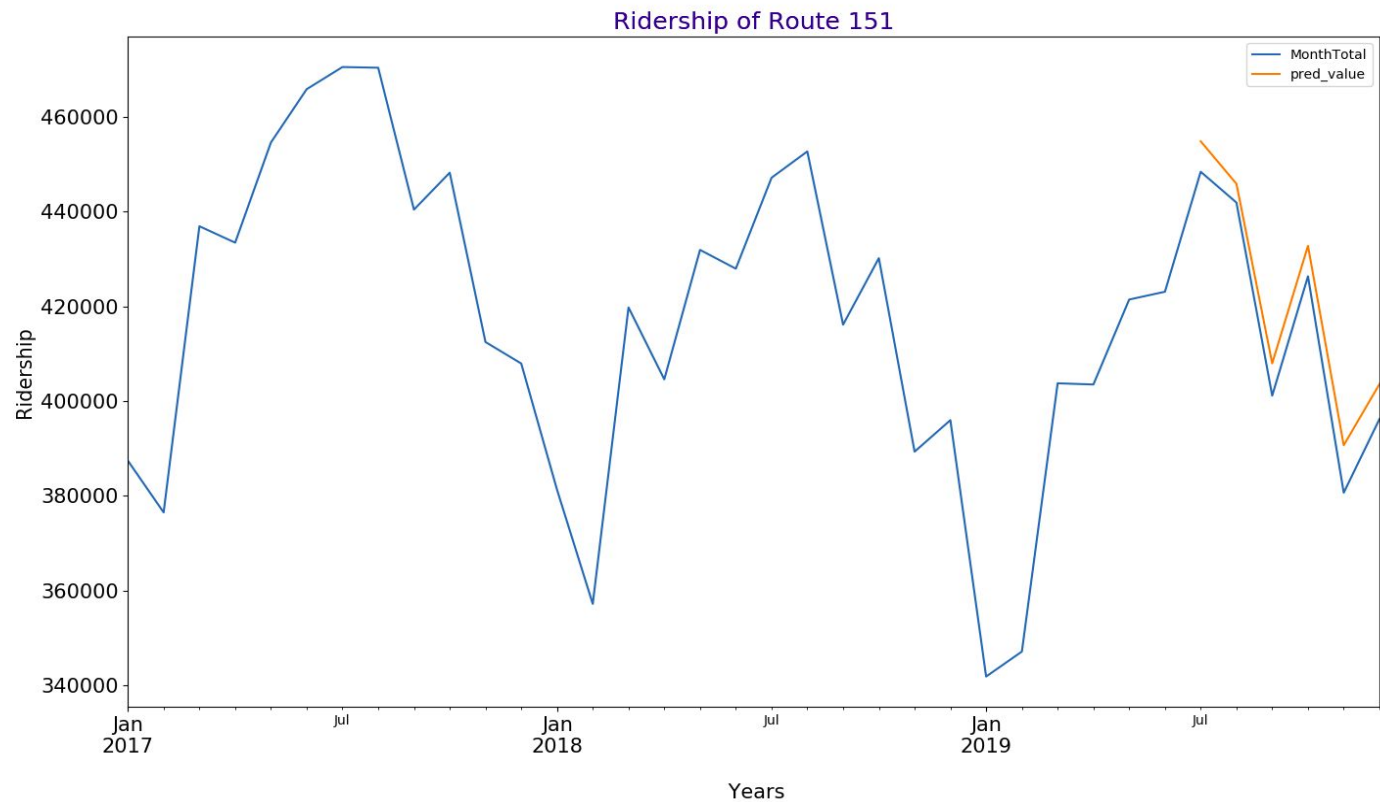
Graphs of ridership prediction for buses - Route 3



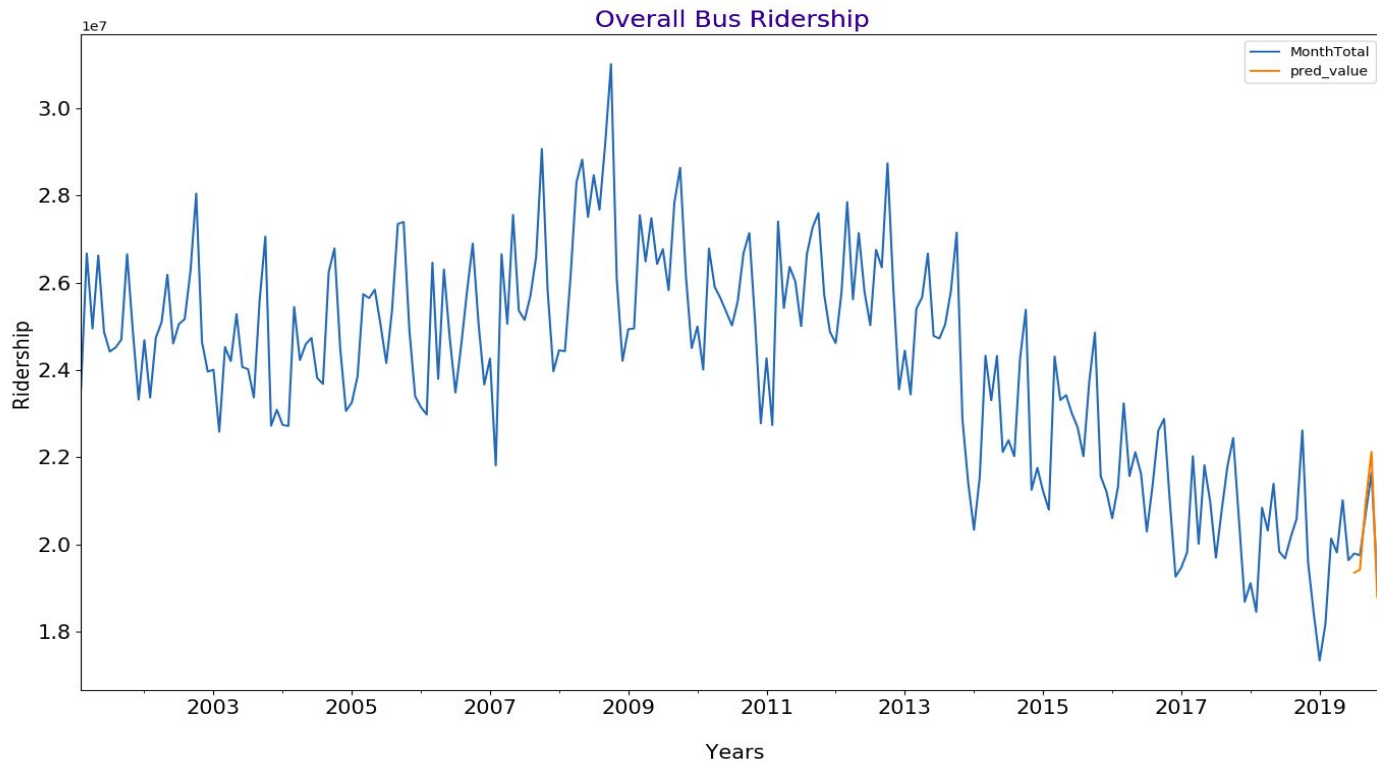
Graphs of ridership prediction for buses - Route 77



Graphs of ridership prediction for buses - Route 151



Predicting the total bus ridership for the year 2019



Classifying predicted ridership values into categories



After predicting the bus ridership for each of the buses, we classified the ridership into 5 categories to suggest optimization of bus services to CTA based on these categories. These categories were divided based on normal distribution of the change in ridership values:

1. Almost same (0.9 - 1.1% change)
2. Slight increase (1.1-1.2 % change)
3. Heavy increase (> 1.2 % change)
4. Slight decrease (0.8-0.9% change)
5. Heavy decrease (< 0.8 % change)

Results

After predicting the ridership values and classifying them into categories, we saved these results in a CSV file. The final file look like this -

	Month	route	Predicted_Ridership_2019	Ridership_2018	Classification	Accuracy_of_Prediction
0	July	3	438196	438230	Almost Same	98.518480
1	August	3	426260	425307	Almost Same	98.681042
2	September	3	434943	448073	Almost Same	98.573122
3	Ocotber	3	459463	490167	Almost Same	98.616081
4	November	3	405580	418467	Almost Same	97.337633
...
235	August	204	28596	26747	Almost Same	95.344055
236	September	204	29164	29765	Almost Same	94.841127
237	Ocotber	204	28086	31949	Slight Decrease	93.904481
238	November	204	27880	32453	Slight Decrease	94.008516
239	December	204	19136	28978	Heavy Decrease	88.814186

Predicting the ridership of buses and trains using ARIMA



- ARIMA: Auto-Regressive Integrated Moving Average
- ARIMA allows us to train and tune the model based on it's 3 hyperparameters (p, d, q)
- These hyperparameters tune the seasonality and the trend factor into the model.
- Hyperparameters:
 - p - controls the auto-regressive aspect of the value trend
 - d - controls the differencing factor between successive values
 - q - controls the weightage of the moving average

Predicting the ridership of buses and trains using ARIMA

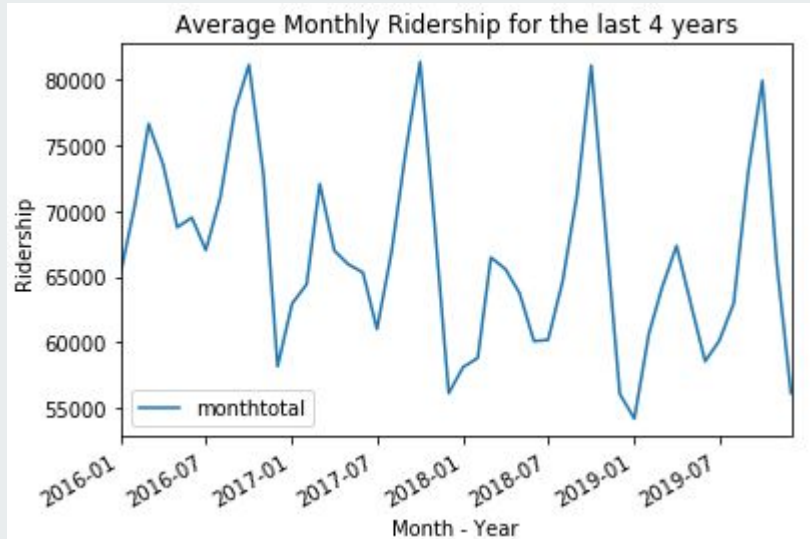


Fig. Shows the yearly trend in ridership for trains

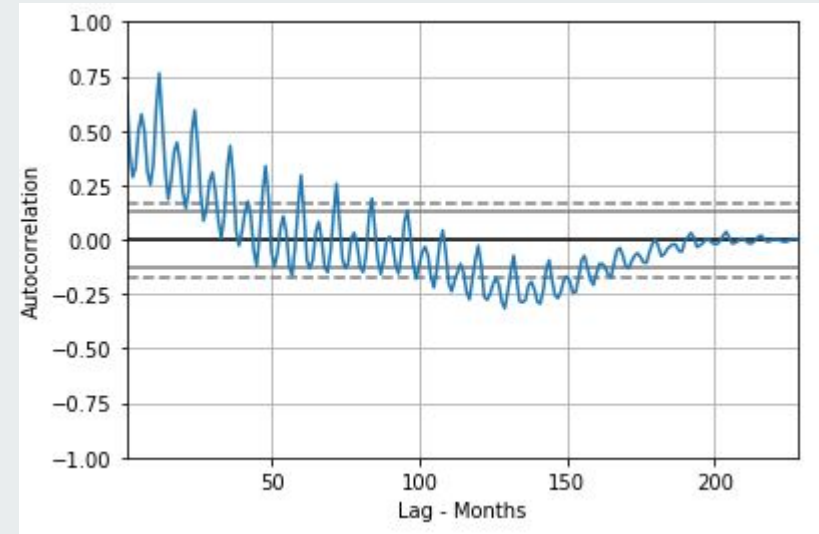
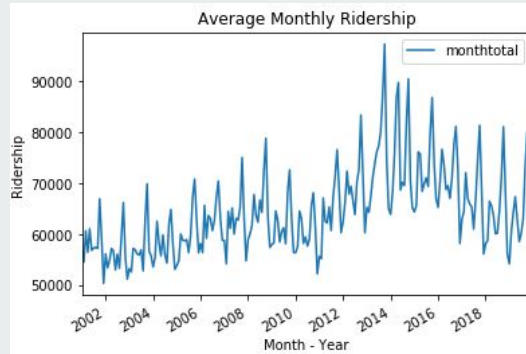


Fig. Auto-correlation between ridership of successive years

Predicting the ridership of buses and trains using ARIMA

- Optimal Hyper-parameters: $(p, d, q) = (8, 1, 1)$
- This means the model performs optimally when consider previous 8 values into our prediction, with single differencing and normal weightage to the moving average.
- This can be confirmed by looking at the chart. It does not mean that we only take the previous 8 years of values, but that that taking 8 years of data is sufficient for accurate predictions.



Predicting the ridership of buses and trains using ARIMA

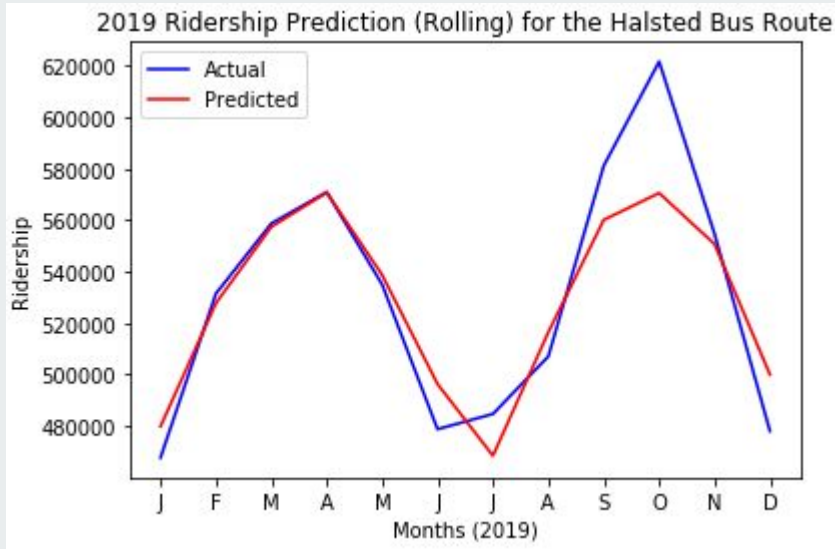


Fig. Rolling predictions for bus dataset

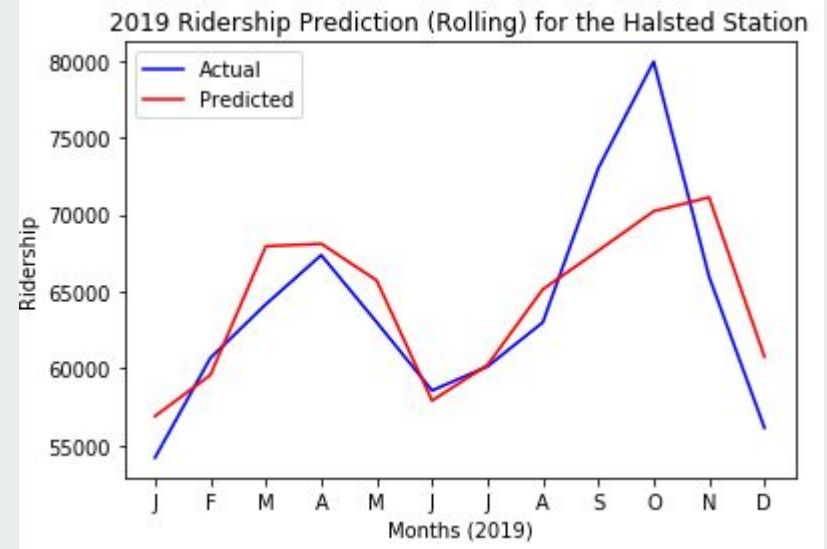


Fig. Rolling predictions for trains dataset

Predicting the ridership of buses and trains using ARIMA



- Attained an accuracy of ~95.6% for the buses dataset and an accuracy of ~95.2% for the trains dataset.
- Rolling predictions predict values for a given month, and then add this prediction to the dataset to predict the value for the next month. Hence, the error might increase the further into the future we try to predict.
- Using these rolling predictions and comparing them to previous ridership values, we can classify which routes and train stations are going to see an increase or decrease in ridership in the following year.
 - Jefferson Park, LaSalle/Van Buren, Pulaski-Forest Park, Halsted/63rd ← Predicted Decrease in Ridership

Evaluation



- We could predict the ridership values with very high accuracy.
- Our model performed well on the dataset
- Achieved an accuracy of **93.6%** for bus ridership prediction using **LSTM**
- Attained an accuracy of **~95.6%** for the buses dataset and an accuracy of **~95.2%** for the trains dataset using **ARIMA**
- This analysis helps CTA in allocating the optimal amount of resources for each route and station to minimize their costs.

Conclusion



- We analyzed the CTA datasets for buses and trains and found out interesting trends and travel patterns
- Predicted the ridership for buses and trains with ~95% accuracy
- Based on the ridership values, we classified the buses and trains into categories according to change in ridership that is likely to arrive in the future

Objective:

The categories predicting the change in ridership for each bus and train route can:

- Help CTA effectively optimize their bus and train rides and infrastructure to accommodate the demands of the commuters and manage their finances effectively.
- They can increase or decrease the buses running on routes with heavy increase and decrease in ridership respectively.



Thank You !