Citibike Bikeshare usage demand forecasting

https://www.citibikenyc.com/

A social experiment for predicting future demand of Citibike shared bicycle system in New York city

Sachid Deshmukh [a] | Ann Liu-Ferrara [a] | Ahmed Sajjad [a]

a M.S. in Data Science student CUNY SPS

Highlights

- ❖ Paper focuses on analyzing Citibike data and identifying important patterns to support business viability and overall profitability of Citybike bikeshare system [c]
- ❖ Paper proposes Machine Learning and Timeseries based forecasting methodologies for predicting future usage demand of Citibike bikeshare system in New York city [a,b]
- ❖ Paper proposes various feature engineering techniques for predicting future usage demand of Citibike bikeshare system in New York city
- ❖ Paper evaluates feature importance and proposes key features which contributes towards predicting future usage demand of Citibike bikeshare system in New York city

[a] https://www.citibikenyc.com/ [b] Hyndman & Athanasopoulos. https://www.otexts.org/fpp2/ [c] https://www.citibikenyc.com/system-data

Article Info

Keywords
Bikeshare, Weather, Cycling, Citibike, New York City

Abstract

Bicycling is an activity which yields many benefits: Riders improve their health through exercise, while traffic congestion is reduced if riders move out of cars, with a corresponding reduction in pollution from carbon emissions [5]. In recent years, Bike Sharing has become popular

Article History
Created:
02/16/2020

Revision-1
04/03/2020

in a growing list of cities around the world. The NYC "Citibike" bicycle sharing scheme went live (in midtown and downtown Manhattan) in 2013, and has been expanding ever since, both as measured by daily ridership as well as the expanding geographic footprint incorporating a growing number of "docking stations" [9-10] as the system welcomes riders in Brooklyn, Queens, and northern parts of Manhattan which were not previously served. One problem that many bikeshare systems face is money. An increase in the number of riders who want to use the system necessitates that more bikes be purchased and put into service to accommodate them. Heavy ridership induces wear on the bikes, requiring for more frequent repairs. However, an increase in the number of trips does not necessarily translate to an increase in revenue because riders who are clever can avoid paying surcharges by keeping the length of each trip below a specified limit (either 30 or 45 minutes, depending on user category.) We seek to examine Citibike ridership data [2-3], joined with daily NYC weather data [4], to study the impact of weather on shared bike usage and generate a predictive model which can estimate the number of trips that would be taken on each day [6-8]. The goal is to estimate future demand which would enable the system operator to make expansion plans. Our finding is that ridership exhibits strong seasonality, with correlation to weather-related variables such as daily temperature and precipitation [6-8]. Additionally, ridership is segmented by user type (annual subscribers use the system much more heavily than casual users), gender (there are many more male users than female) and age (a large number of users are clustered in their late 30s). [11-12]

1. Introduction

Since 2013 a shared bicycle system known as Citibike has been available in New York City [1]. The benefits to having such a system include reducing New Yorkers' dependence on automobiles and encouraging public health through the exercise attained by cycling [5]. Additionally, users who would otherwise spend money on public transit may find bicycling more economical – so long as they are aware of Citibike' s pricing constraints. There are currently about 12,000 shared bikes which users can rent from about 750 docking stations located in Manhattan and in western portions of Brooklyn and Queens. A rider can pick up a bike at one station and return it at a different station. The system has been expanding each year, with increases in the number of bicycles available and expansion of the geographic footprint of docking stations. For planning purposes, the system operator needs to project future ridership to make good investments. The available usage data provides a wealth of information which can be mined to seek trends in usage 2-3]. With such intelligence, the company would be better positioned to determine what actions might optimize its revenue stream.

- Because of weather, ridership is expected to be lower during the winter months, and on foul-weather days during the rest of the year, compared to a warm and sunny summer day. Using the weather data, we can seek to model the relationship between bicycle ridership and fair/foul or hot/cold weather [6-8].

- What are the differences in rental patterns between annual members (presumably, residents) vs. casual users (presumably, tourists?) [12]
- Is there any significant relationship between the age and/or gender of the bicycle renter vs. the rental patterns? [12]
- What is the seasonal component of bikeshare pattern exhibited by consumers (weekday vs weekend) [11]

2. Data Collection

We are leveraging two major sources of dataset for this scientific experiment

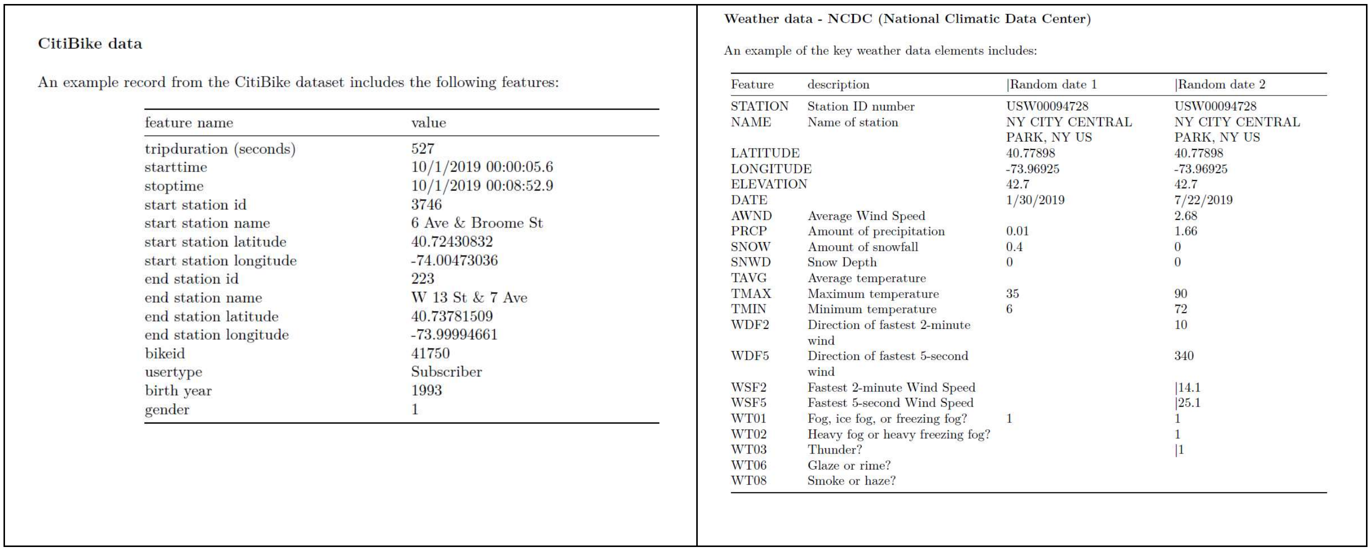
1. **Citibike trip dataset [2-3]**
   Citibike makes a vast amount of data available regarding system usage as well as sales of memberships and short-term passes. For each month since the system's inception, there is a file containing details of (almost) every trip. (Certain "trips" are omitted from the dataset. For example, if a user checks out a bike from a dock but then returns it within one minute, the system drops such a "trip" from the listing, as such "trips" are not interesting.) There are currently 77 monthly data files for the New York City bikeshare system, spanning July 2013 through November 2019. Each file contains a line for every trip. The number of trips per month varies from as few as 200,000 during winter months in the system's early days to more than 2 million trips this past summer. The total number of entries was more than 90 million, resulting in 17GB of data. Because of the computational limitations which this presented, we created samples of 1/1000 and 1/100 of the data. The samples were created deterministically, by subsetting the files on each 1000th (or, 100th) row.

2. **Central Park daily weather data [4]**
   Also we obtained historical weather information for 2013-2019 from the NCDC (National Climatic Data Center) by submitting an online request to https://www.ncdc.noaa.gov/cdo-web/search. Although the weather may vary slightly within New York City, we opted to use just the data associated with the Central Park observations as proxy for the entire city's weather. We believe that the above data provides a reasonable representation of the target population (all Citibike rides) and the citywide weather.

3. Methodology

For Citibike bikeshare usage demand forecasting, we are leveraging Citibike trip dataset [2-3] along with Central Park daily weather data [4]. Goal of the project is to analyze Citibike bikeshare usage data at daily level combined with Central Park weather data and predict future bikeshare usage demand in terms of no of trips per day for next 30 days. Future usage demand is very critical to Citybike system operator for planning increase in the number of bicycles availability and expansion of the geographic footprint of docking stations. This enables data driven decision system for making informed business expansion plans resulting into increased ROI and long-term business profitability and viability of Citibike operations

Dataset Snapshot:



CitiBike data

An example record from the CitiBike dataset includes the following features:

feature name	value
tripduration (seconds)	527
starttime	10/1/2019 00:00:05.6
stoptime	10/1/2019 00:08:52.9
start station id	3746
start station name	6 Ave & Broome St
start station latitude	40.72430832
start station longitude	-74.00473036
end station id	223
end station name	W 13 St & 7 Ave
end station latitude	40.73781509
end station longitude	-73.99994661
bikeid	41750
usertype	Subscriber
birth year	1993
gender	1

Weather data - NCDC (National Climatic Data Center)

An example of the key weather data elements includes:

Feature	description	|Random date 1	|Random date 2
STATION	Station ID number	USW00094728	USW00094728
NAME	Name of station	NY CITY CENTRAL PARK, NY US	NY CITY CENTRAL PARK, NY US
LATITUDE		40.77898	40.77898
LONGITUDE		-73.96925	-73.96925
ELEVATION		42.7	42.7
DATE		1/30/2019	7/22/2019
AWND	Average Wind Speed		2.68
PRCP	Amount of precipitation	0.01	1.66
SNOW	Amount of snowfall	0.4	0
SNWD	Snow Depth	0	0
TAVG	Average temperature		
TMAX	Maximum temperature	35	90
TMIN	Minimum temperature	6	72
WDF2	Direction of fastest 2-minute wind		10
WDF5	Direction of fastest 5-second wind		340
WSF2	Fastest 2-minute Wind Speed		|14.1
WSF5	Fastest 5-second Wind Speed		|25.1
WT01	Fog, ice fog, or freezing fog?	1	1
WT02	Heavy fog or heavy freezing fog?		1
WT03	Thunder?		|1
WT06	Glaze or rime?		
WT08	Smoke or haze?		

Data Cleaning and Imputation

Citibike dataset is huge. Processing big volume data like this demands sophisticated big data platform with distributed data processing capability. To resolve system scaling challenge, we employed stochastic random sampling technique for this analysis [13]. Underlying dataset is selected for scientific experimentation and analysis using uniform random sampling technique reducing the volume of data for faster computation [13]

In addition to that following data cleaning and data imputation techniques are used for data preparation.

1] Data column names changes slightly from month to month data file. Need to add column name standardization transformation for data processing
2] In some months, Citibike specifies dates in YYYY-MM-DD format, while in other months, dates are Specified in MM/DD/YYYY format. Need to add date standardization transformation for data processing
3] Dataset has some records with unusually high recording of bikeshare session duration. (StartTime and EndTime) Need to employ outlier detection techniques to identify such records and data imputation techniques to normalize the data
4] Data Aggregation and join: Aggregate individual Citibike trip data by day, and join to daily weather data

Data Analysis and inference

Citibike users tend to book shorter bikeshare trips, most of them are less than 10 mins

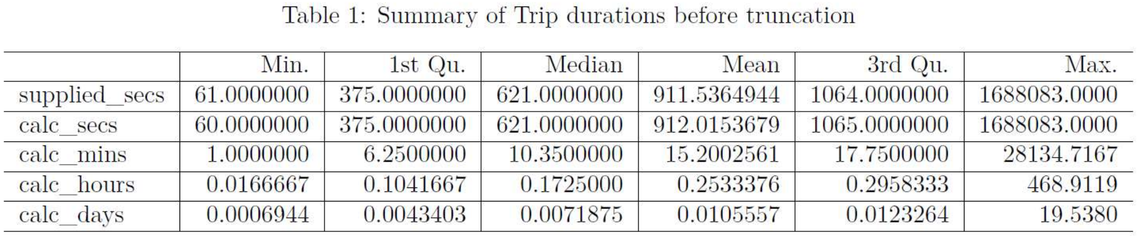


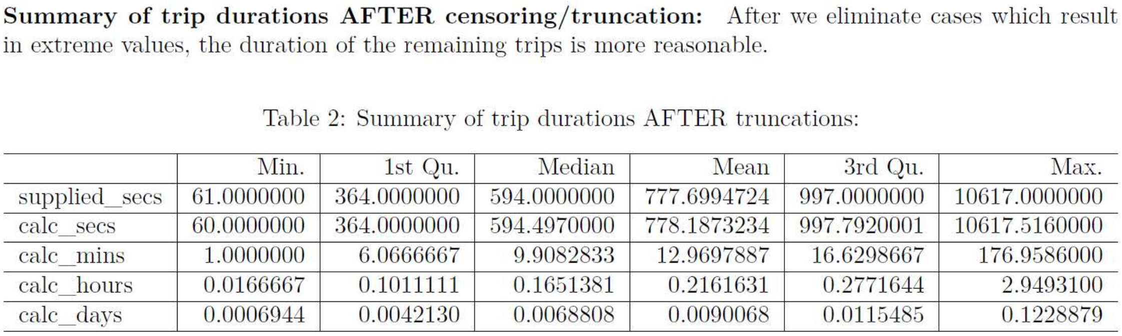
Table 1: Summary of Trip durations before truncation

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
supplied_secs	61.0000000	375.0000000	621.0000000	911.5364944	1064.0000000	1688083.0000
calc_secs	60.0000000	375.0000000	621.0000000	912.0153679	1065.0000000	1688083.0000
calc_mins	1.0000000	6.2500000	10.3500000	15.2002561	17.7500000	28134.7167
calc_hours	0.0166667	0.1041667	0.1725000	0.2533376	0.2958333	468.9119
calc_days	0.0006944	0.0043403	0.0071875	0.0105557	0.0123264	19.5380

The above indicates that the duration of the trips (in seconds) includes values in the millions which likely reflects a trip which failed to be properly closed out.

We removed cases with unreasonable trip_duration values Let’s assume that nobody would rent a bicycle for more than a specified time limit (say, 3 hours), and drop any records which exceed this:

```
## [1] "Removed 157 trips (0.174%) of longer than 3 hours."
## [1] "Remaining number of trips: 90213"
```

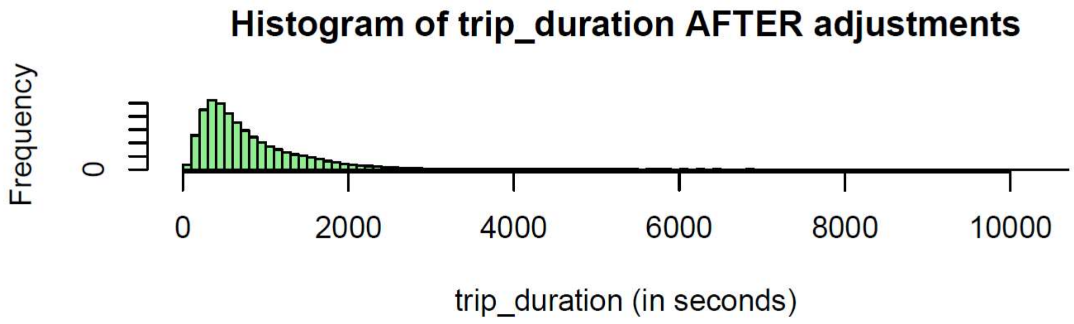


Summary of trip durations AFTER censoring/truncation: After we eliminate cases which result in extreme values, the duration of the remaining trips is more reasonable.

Table 2: Summary of trip durations AFTER truncations:

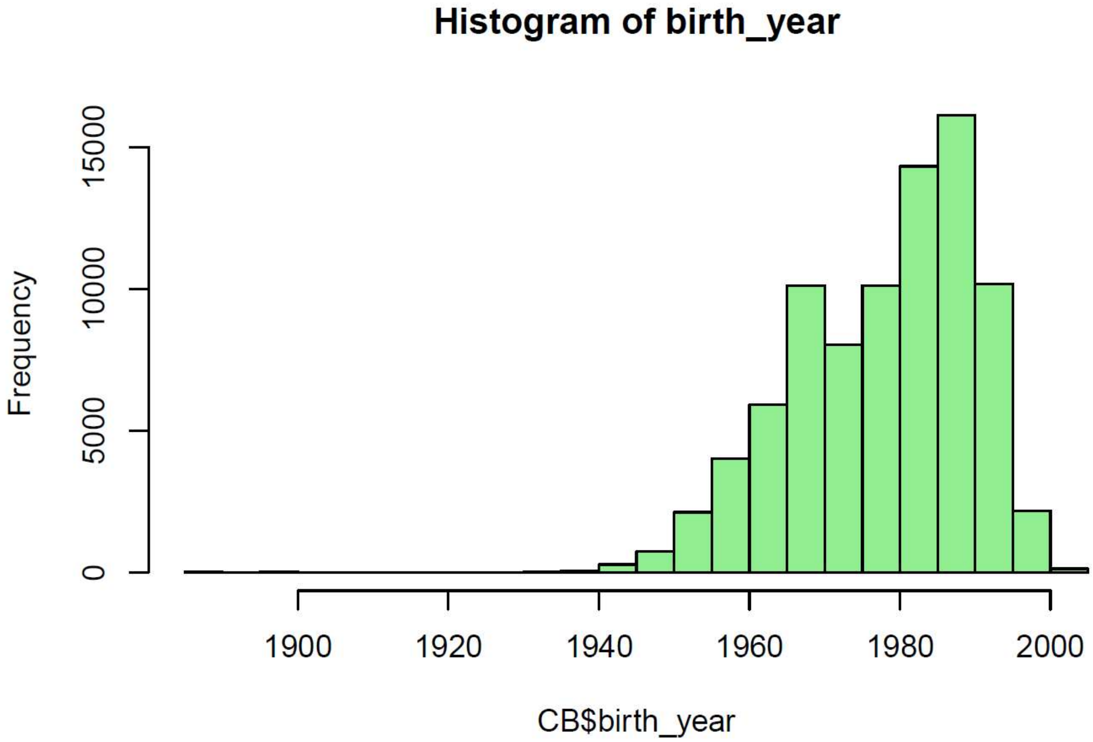
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
supplied_secs	61.0000000	364.0000000	594.0000000	777.6994724	997.0000000	10617.0000000
calc_secs	60.0000000	364.0000000	594.4970000	778.1873234	997.7920001	10617.5160000
calc_mins	1.0000000	6.0666667	9.9082833	12.9697887	16.6298667	176.9586000
calc_hours	0.0166667	0.1011111	0.1651381	0.2161631	0.2771644	2.9493100
calc_days	0.0006944	0.0042130	0.0068808	0.0090068	0.0115485	0.1228879

Trip duration statistics after outlier removal looks more realistic



Most of Citibike users are in their 30s/40s

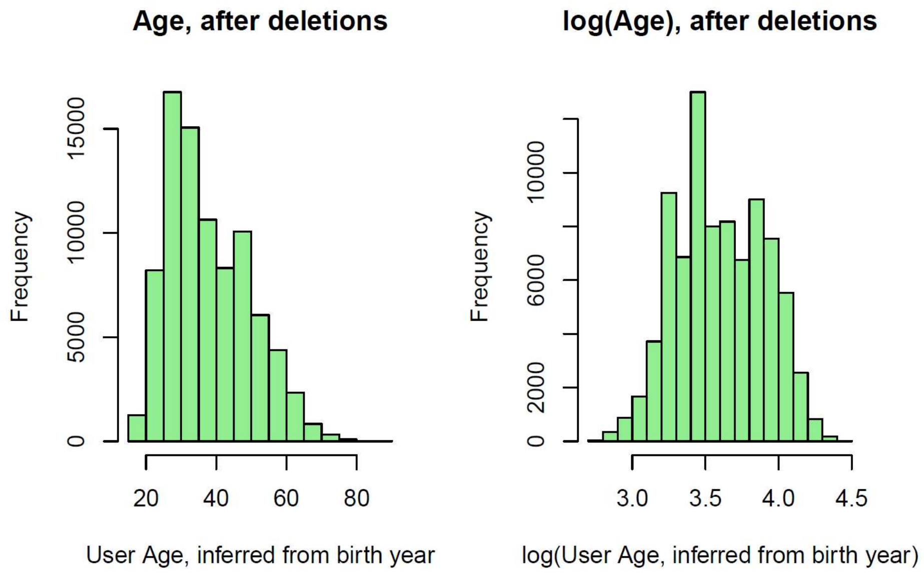
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1885    1969    1981    1978    1988    2003    5828
```



Birth year recorded for some of the users is 1885 which can’t be true

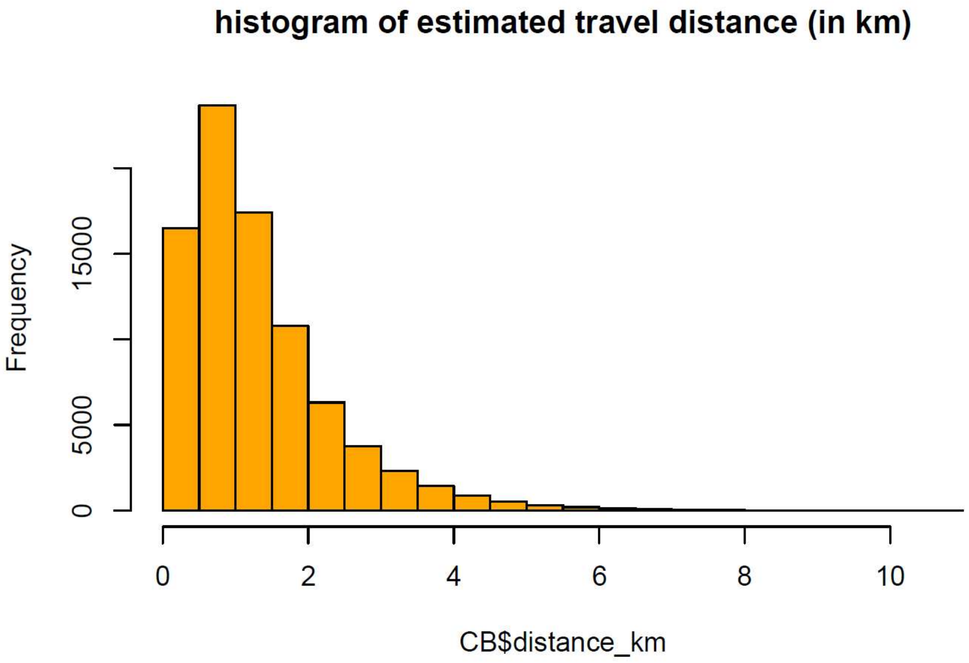
Remove trips associated with very old users (age>90) and remove trips associated with missing birth_year)

```
## [1] "Removed 40 trips (0.044%) of users older than 90 years."
## [1] "Removed 5828 trips (6.449%) of users where age is unknown (birth_year unspecified)."
## [1] "Remaining number of trips: 84345"
```



Citibike consumers mostly use bikeshare service for shorter trips < 2 KM [a]

[a] This is straight-line distance between (longitude, latitude) points – it doesn't incorporate an actual bicycle route.

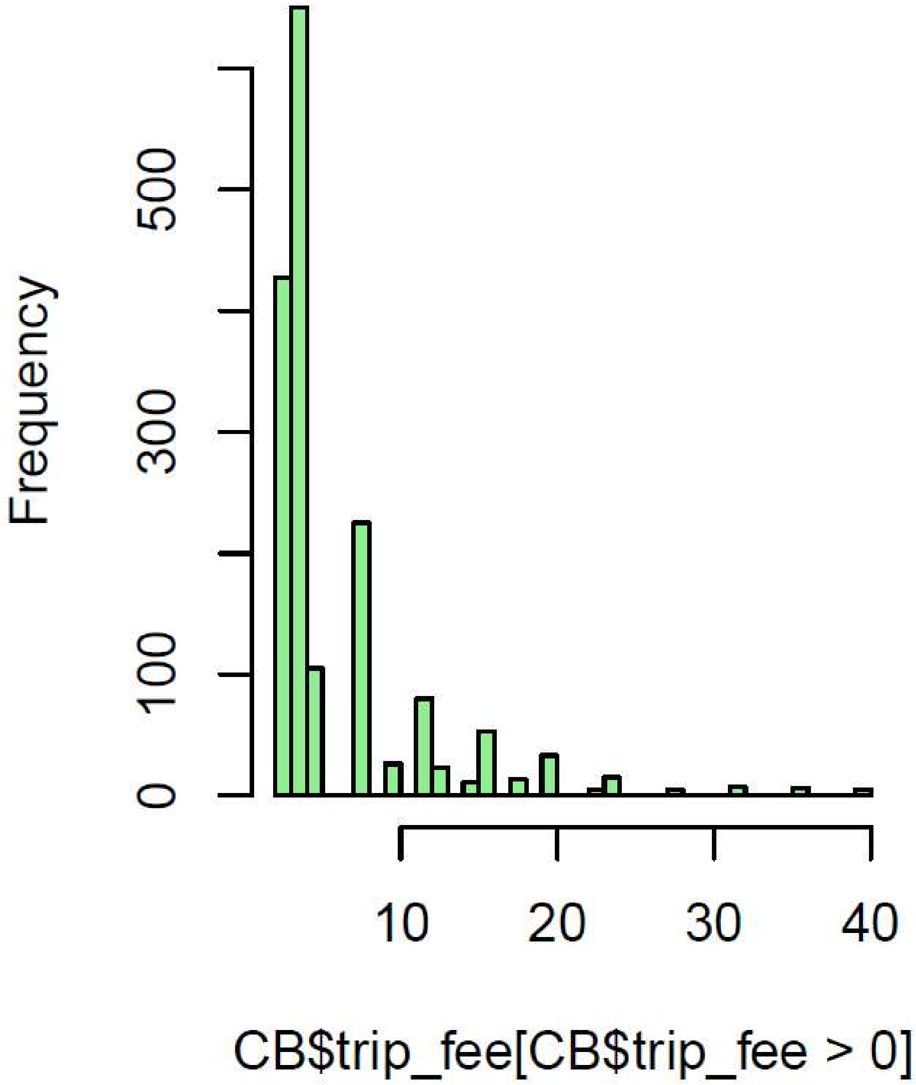


Analyze Trip Fee

Compute usage fee There is a time-based usage fee for rides longer than an initial period:

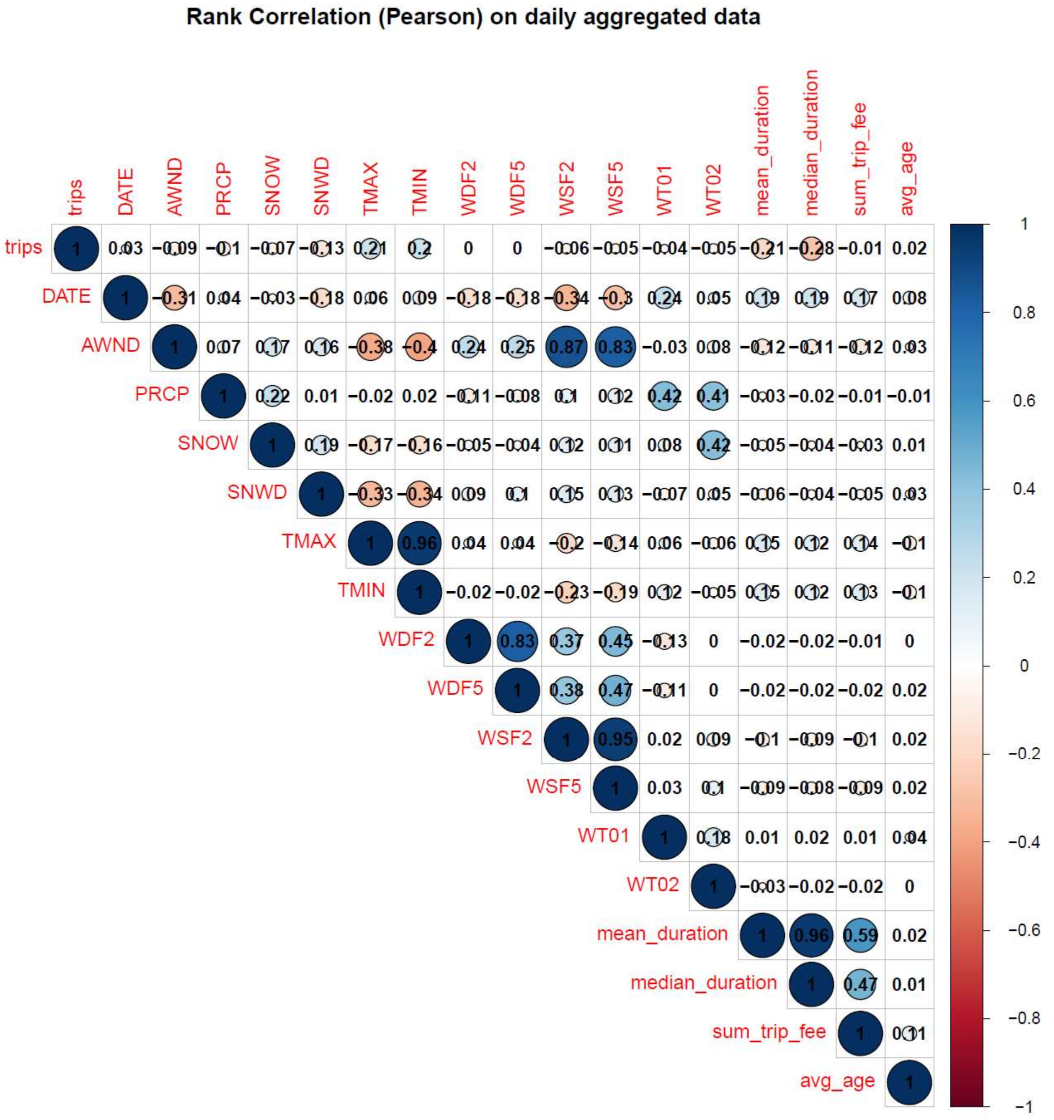
- For user type=Subscriber, the fee is $2.50 per 15 minutes following an initial free 45 minutes per ride.

- For user type=Customer, the fee is $4.00 per 15 minutes following an initial free 30 minutes per ride.
- There are some cases where the user type is not specified (we have relabeled as “UNKNOWN”, and we do note estimate usage fees for such trips.)



Most of Citibike users spent less than 10$ per bikeshare trip

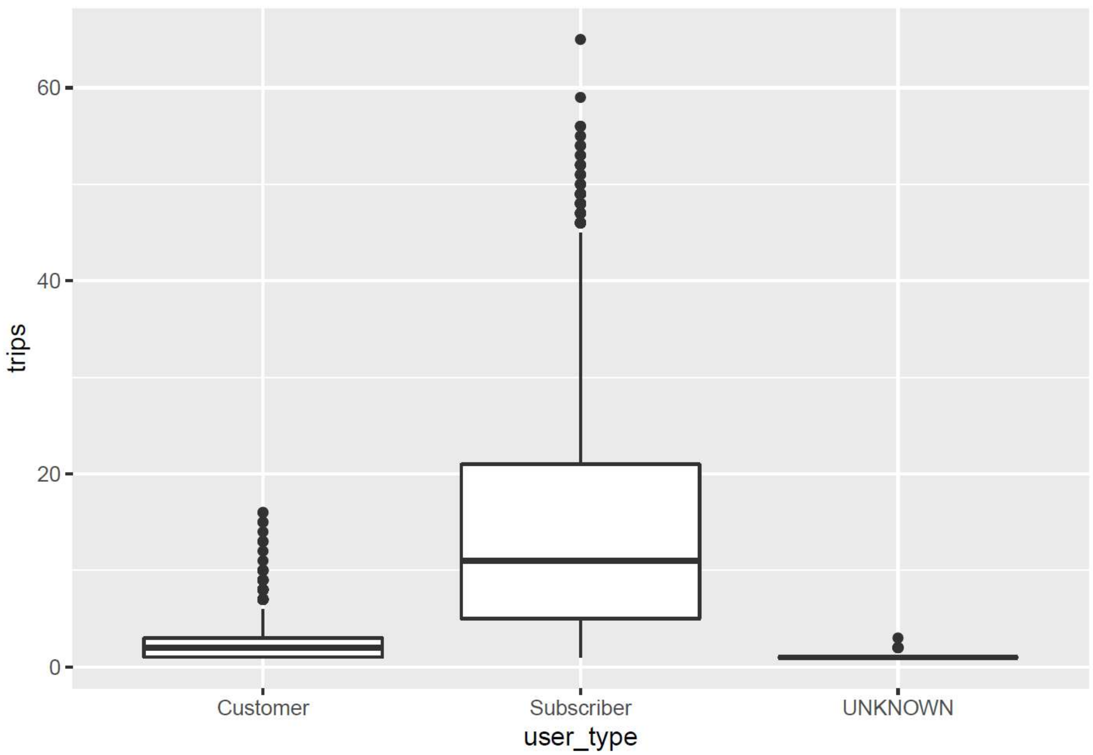
Feature correlation: Correlation between input feature set reveals very interesting relationship between variables, and surfaces out potential problems of multicollinearity [14].



1] Negative correlation between trips and AWND suggests that people tend to take fewer bike trips on a windy day
2] Negative correlation between trips and SNOW suggests that as amount of snowfall increases number of bikeshare trips decreases
3] Negative correlation between trips and SNWD suggests that as amount of snow depth increases number of bikeshare trips decreases
4] TMAX and TMIN has positive correlation with trips. This indicates that on a warm sunny day people tends to take more bikeshare trips

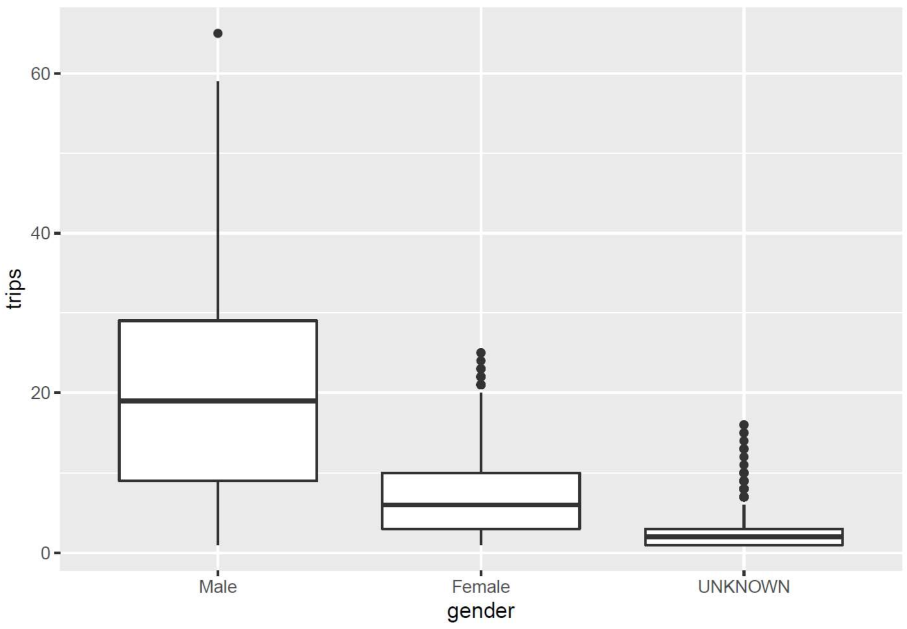
Number of trips by user types

Citibike users who have monthly subscription tends to take a greater number of trips compared to other user types



Number of trips by gender

Male customers tend to take more bikeshare trips compared to female counterpart



4. Model building

We aimed to focus on wide variety of univariate statistical time series models and multivariate machine learning models for Citibike ridership demand forecasting. This experiment also attempts to leverage ensemble-based modelling solution to demand forecast complex business scenario such as Citibike . For model evaluation this paper relies on performance metric as RMSE [17].

Citibike ridership time series and STL timeseries decomposition

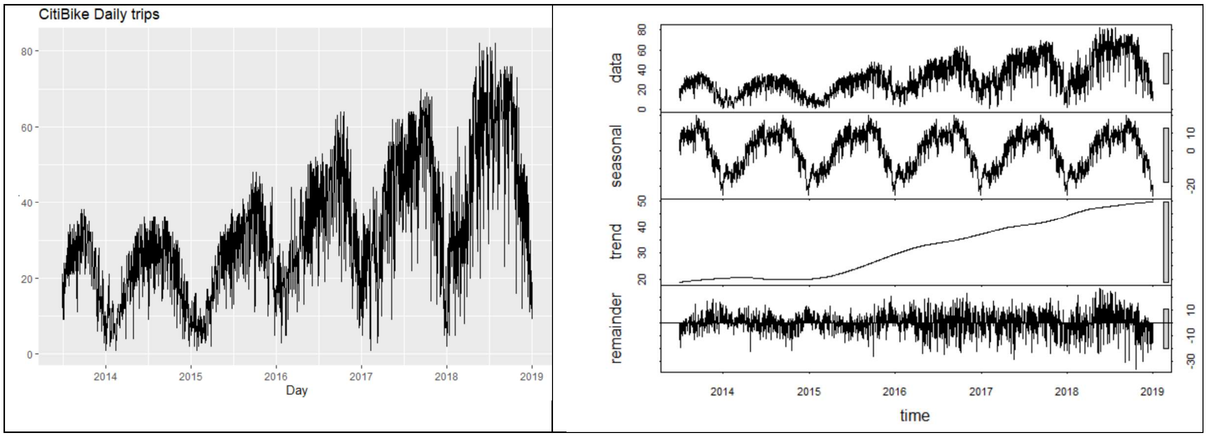


Fig 4.1

Left hand side of Fig 4.1 above shows Citibike ridership timeseries from year 2013 to 2019. There exists a strong yearly seasonality in Citibike ridership pattern where ridership count increases in summer and decreases in winter. Timeseries plot also exhibits a positive trend over the time showing growing Citibike popularity and customer base from year 2013 to 2019

Right hand side of Fig 4.1 shows timeseries decomposition using STL (Seasonality and Trend Decomposition using Loess) [16]. Seasonal decomposition plot further justifies the presence of yearly seasonality and strong positive trend in the Citibike bikeshare ridership data.

Experimental Models [15]

As part of this experimentation, following univariate timeseries models and multivariate ML models are leveraged for demand forecasting Citibike ridership pattern [15]

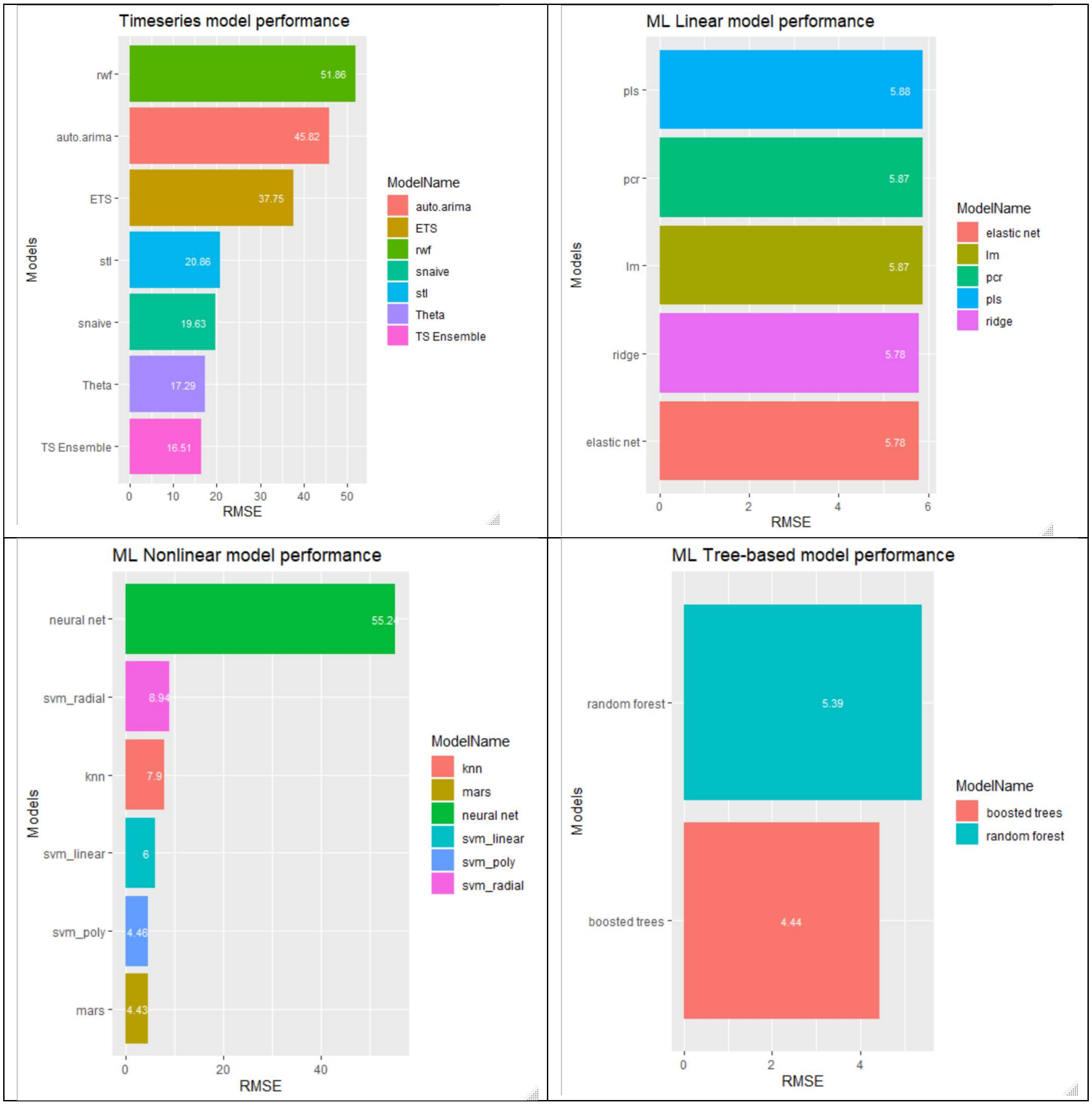


Fig 4.2

As seen in fig 4.2 above, best model in each category along with respective RMSE are as follows

Model Name	**RMSE**	**Model Category**
TS Ensemble [a]	16.51	Time Series
Elastic Net [b]	5.78	Linear Models
MARS	4.63	Nonlinear model
Boosted Trees	4.44	Tree-based models

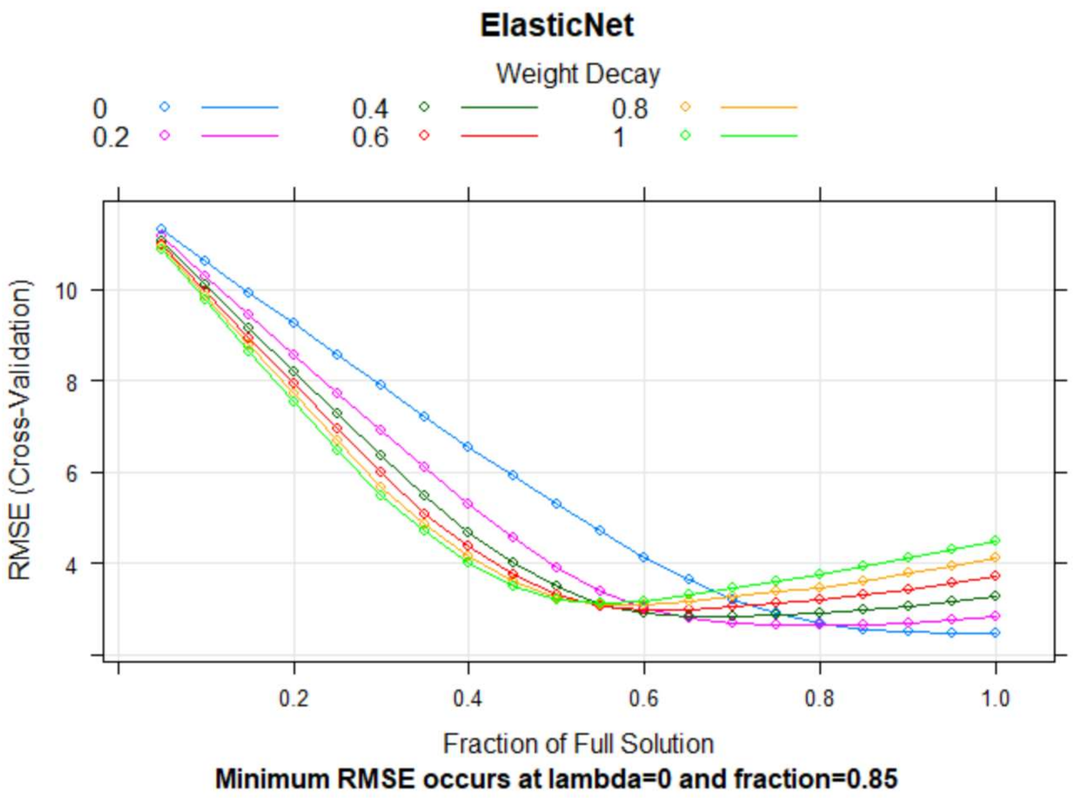


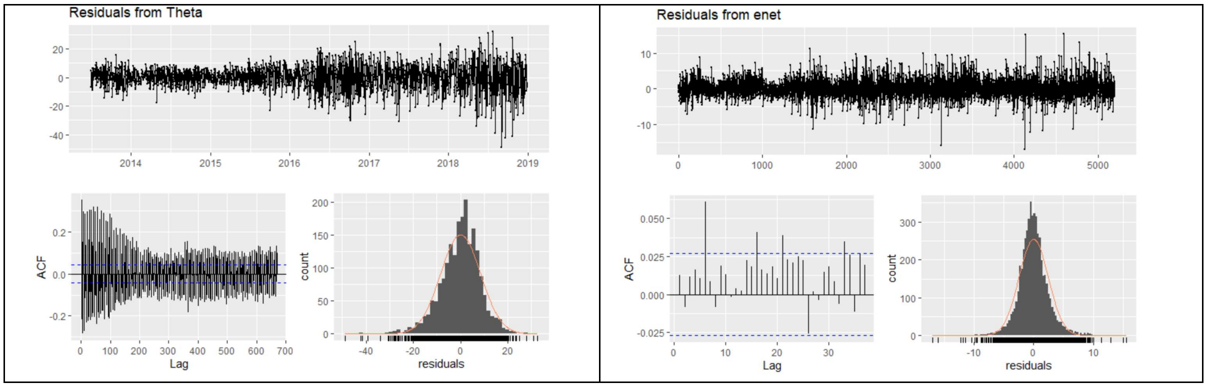
Fig 4.3

a. TS Ensemble – Ensemble of top three models as per RMSE (Theta, snaive and STL)

b. Best RMSE for elastic net occurs at lambda = 0 indicating no penalty for the model

Fig 4.3 shows that min RMSE is obtained with lambda = 0 indicating no penalty for elastic net model [15]

Best performing model residual analysis [15]



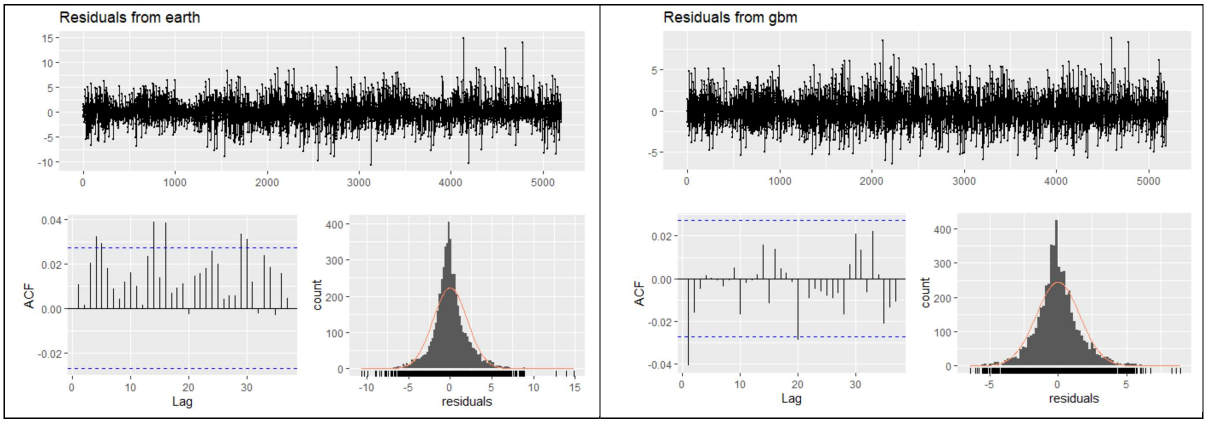
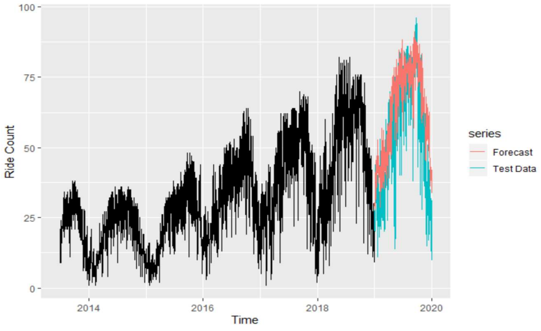
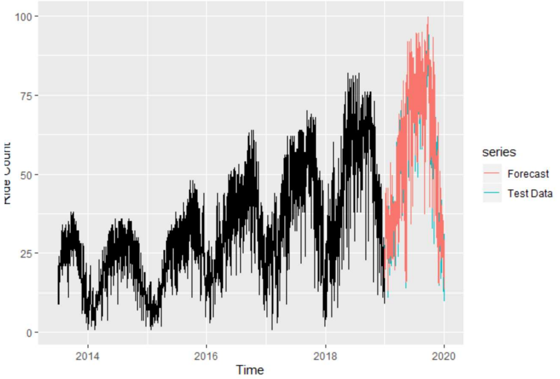


Fig 4.6

Figure 4.6 above shows residual analysis of best performing models in each model category. Top portion of fig 4.6 shows that residuals of time series (theta) and linear (enet) models are not normally distributed and ACF plots for both those categories shows strong correlation amongst the residuals. Bottom portion of fig 4.6 shows that residuals of nonlinear (MARS/earth) and tree-based (gbm) models are much better. We can see for both those categories residuals are normally distributed and ACF plots doesn't show significant correlation amongst residuals. From the residual plots above and performance metrics of the models (RMSE) [17] we can conclude that Non-linear models and Tree-based models are best suited for demand forecasting Citibike ridership pattern

Best performing model actuals vs forecast overlay [15]

Actuals vs Forecast - theta		**Actuals vs Forecast – enet**
		
Actuals vs Forecast - mars		**Actuals vs Forecast - gbm**

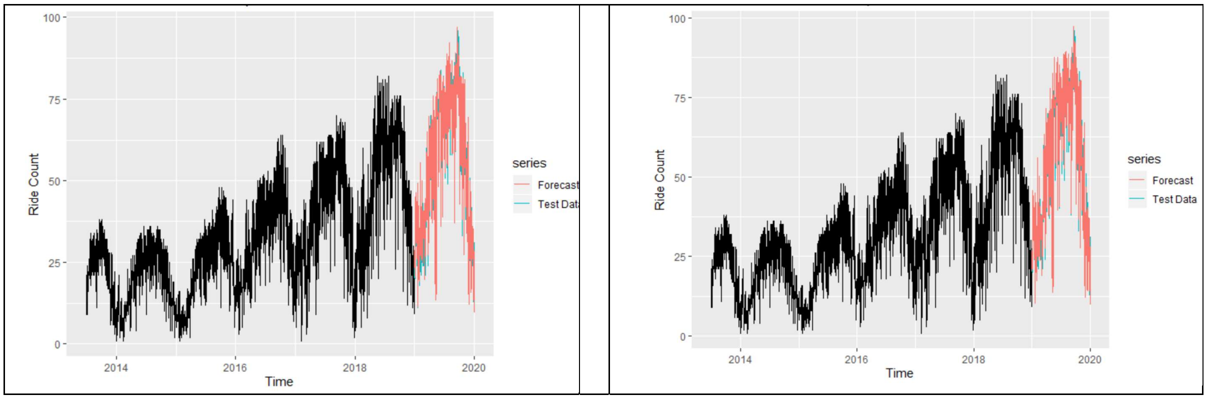


Fig 4.7

Figure 4.7 above shows Citibike ridership actuals vs forecast overlay for year 2019 along with historical Citibike ridership data from year 2013 to 2018. Top portion of fig 4.7 shows actuals vs forecast overlay for timeseries (theta) and linear (enet) model. From the actual vs forecast overlay plot above we can see that forecast generated using both these model categories deviates a lot from actual Citibike ridership pattern observed for year 2019. Bottom portion of fig 4.7 shows actuals vs forecast overlay for nonlinear (MARS) and tree-based (gbm) model. From the actual vs forecast overlay plot above we can see that forecast generated using both these model categories is much better and depicts the true actual Citibike ridership pattern observed for year 2019.

Overall model performance [15]

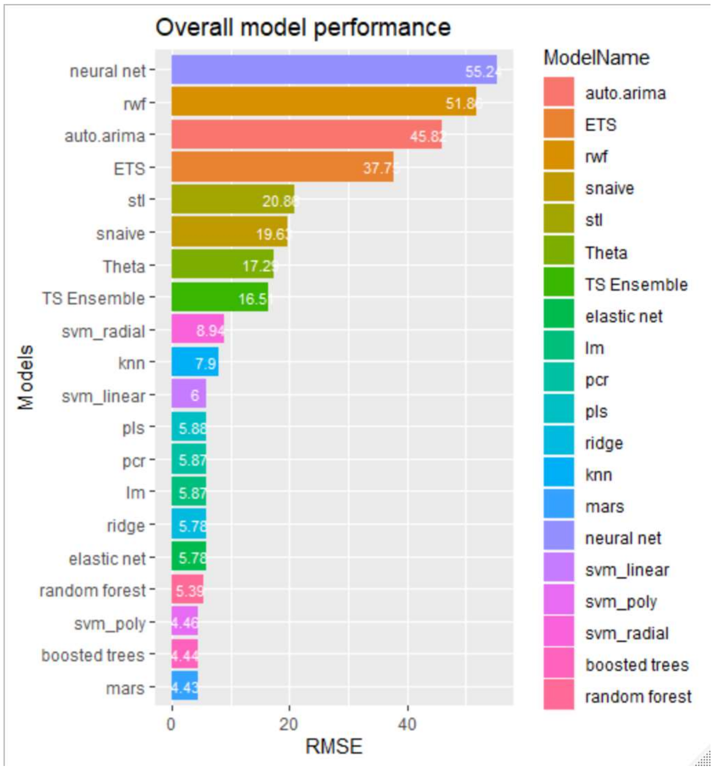


Fig 4.8

Fig 4.8 above shows overall model performance of all models leveraged in this experiment for demand forecasting Citibike ridership pattern. Top three performing model as per RMSE are MARS, Boosted Trees and SVM_Poly

Variable Importance – Winning model [15]

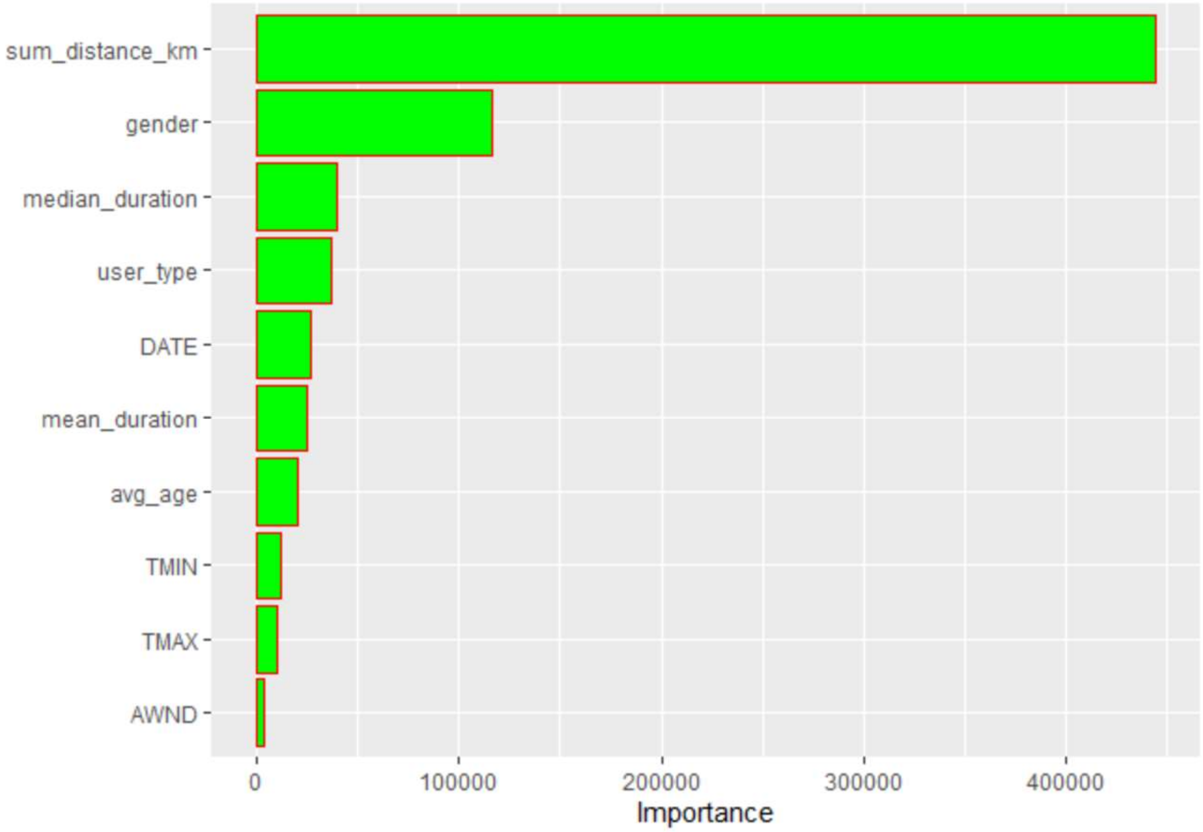


Fig 4.9

Fig 4.9 above shows top 10 important features which influence demand forecasting of Citibike ridership pattern significantly.

Model Ensembles – Stacked top three winning models [15]

One of the goals of this scientific experiment is to evaluate model ensemble (stack model) performance in Citibike ridership demand forecasting. For this analysis we selected top three winning models as shown in fig 4.8 above (MARS, Boosted Trees and SVM_Poly)

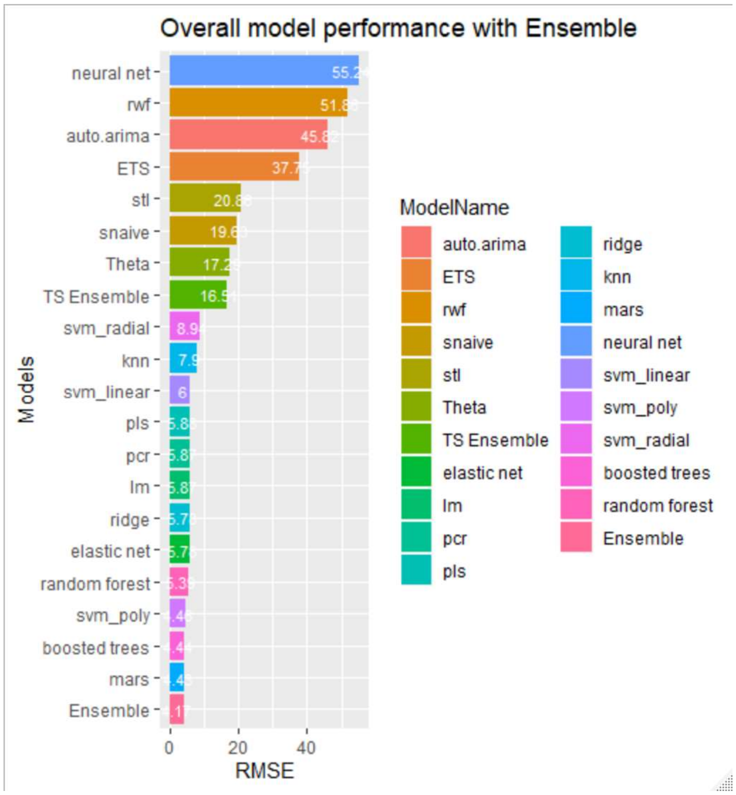


Fig 4.10

Fig 4.10 shows that when ensemble of top 3 best performing models is leveraged for demand forecasting Citibike ridership pattern, it yields best results amongst all the models in terms of RMSE.

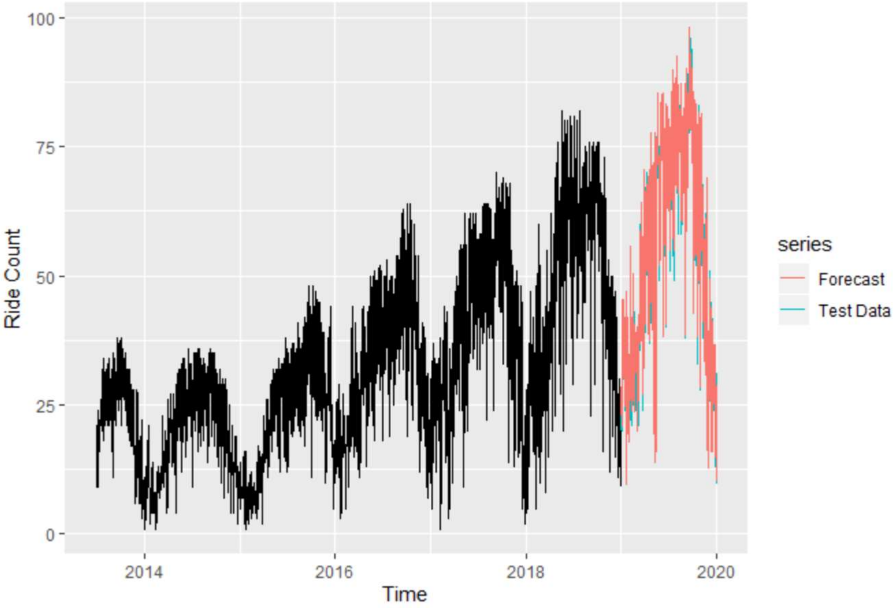


Fig 4.11

Fig 4.11 shows Citibike ridership demand forecast generated using ensemble model overlaid with actual Citibike ridership usage observed for year 2019. From fig 4.11 conclusion can be derived that ensemble model is perfectly suited for demand forecasting scenario as complex as Citibike ridership operation

5. Conclusion

The Citybike ridership analysis determined in this paper have many implications. Analysis reveals interesting scientific facts which are statistically significant. Citibike ridership data shows a strong annual seasonality and increasing trend. Ridership count increases in summer and decreases in winter. Increasing trend in the ridership count from 2013 to 2019 shows increasing popularity of Citybike ridership service over the period. It is established that most Citibike riders are young in their 30s and 40s. Citibike customers prefers to use ride share service for making short distance trips < 2 KM. It is established with statistical significance that weather has a significant impact on Citibike ridership pattern. Good weather has positive correlation with Citibike ridership count.

For predicting future Citibike ridership count, statistical timeseries models alone are not sufficient. Based on RMSE and residual analysis, statistical timeseries models alone doesn't do a good job of forecasting Citibike rideshare count. This is because statistical timeseries models don't leverage the rich feature set available for model optimization e.g. weather data. Based on RMSE, Linear models do better job of forecasting Citibike ridership count than timeseries models. This is because Linear models are multivariate and can leverage rich input features for model optimization.

Nonlinear models further improve the RMSE and outperforms Linear models. The fact that Nonlinear models are better suited for Citibike ridership dataset suggests that there exists a nonlinear relationship between input features and predictor variables

The best model performance is derived from MARS models. MARS model outperforms all other models and provides best performance in terms of RMSE. Residual analysis of MARS model looks more appropriate where residuals are normally distributed and ACF plot shows no correlation between residuals. The variable importance of MARS models reveals important information about what factors significantly impact the Citibike ridership pattern.

From the MARS model variable importance plot, we can conclude that following factors are statistically significant in influencing Citibike ridership future demand

- Distance between station
- Gender
- User Type (customers/ Subscribers)
- Trip Duration
- Day of week
- Avg age of users
- Temperature
- Wind Speed

These facts are key input for Citibike system operator for planning increase in the number of bicycles availability and expansion of the geographic footprint of docking stations.
It is imperative that for a complex business operation like Citibike which involves many factors (Internal and External) one model alone won't be sufficient for ridership forecasting. When we tried Ensemble of top three models (Boosted Trees, MARS and SVM Poly) we got the best performance in terms of RMSE (7 % RMSE improvement). We can conclude that the best modelling strategy for predicting future ridership pattern for Citibike rideshare service is to use model ensembles. Winning model ensembles established in this paper will enable Citibike system operator to make informed business expansion plans resulting into increased ROI and long-term business profitability and viability of Citibike operations

6. References

1. Citibike website: https://www.citibikenyc.com/
2. Citibike data description: https://www.citibikenyc.com/system-data
3. Citibike trip data repository: https://s3.amazonaws.com/tripdata/index.html
4. National Climatic Data Center (NCDC) weather data: https://www.ncdc.noaa.gov/cdo-web/search
5. **https://en.wikipedia.org/wiki/Cycling**
6. Westland et al. examined consumer behavior in bike sharing in Beijing using a deep-learning model incorporating weather and air quality, time-series of demand, and geographical location; later adding customer segmentation.
7. [@Faghih-Imani_Eluru_2018] An et al. examine weather and cycling in New York City and find that weather impacts cycling rates more than topography, infrastructure, land use mix, calendar events, and peaks. They do so by exploring a series of interaction effects, which each capture the extent to which two characteristics occurring simultaneously exert a combinatorial effect on cycling ridership – e.g., how is cycling impacted when it is both wet and a weekend day or humid day in the hilliest parts of the cycling network?
8. In the 1990s, Nankervis examined the effect of weather and climate on university student bicycle commuting patterns in Melbourne, Australia by examining counts of parked bicycles at local universities and correlating with the weather for each day, finding that the deterrent effect of bad weather on commuting was less than commonly believed (though still significant.) [@Nankervis_1999]
9. [@Westland_Mou_Yin_2019] Jia et al. performed a retrospective study of dock less bike sharing in Shanghai to determine whether introduction of such program increased cycling. Their methodology was to survey people in various neighborhoods where the areas were selected by sampling, and the individuals were selected by interviewing individuals on the street.
10. [@Jia_Ding_Gebel_Chen_Zhang_Ma_Fu_2019] Jia and Fu further examined whether dock less bicycle-sharing programs promote changes in travel mode in commuting and non-commuting trips, as well as the association between change in travel mode and potential correlates, as part of the same Shanghai study.
11. [@DellAmico_Iori_Novellani_Subramanian_2018] Zhou analyzed massive bike-sharing data in Chicago, constructing a bike flow similarity graph and using a

fast-greedy algorithm to detect spatial communities of biking flows. He examined the questions 1. How do bike flow patterns vary because of time, weekday or weekend, and user groups? 2. Given the flow patterns, what was the spatiotemporal distribution of the over-demand for bikes and docks in 2013 and 2014? [@Zhou_2015]

12. Hosfordetal surveyed participants in Vancouver, Canada and determined that public bicycle share programs are not used equally by all segments of the population. In many cities, program members tend to be male, Caucasian, employed, and have higher educations and incomes compared to the general population. Further, their study determined that the majority of bicycle share trips replace trips previously made by walking or public transit, indicating that bicycle share appeals to people who already use active and sustainable modes of transportation
13. https://research-methodology.net/sampling-in-primary-data-collection/random-sampling/
14. https://en.wikipedia.org/wiki/Multicollinearity
15. https://www.rpubs.com/sachid/Data698_CapstoneProject
16. https://www.statsmodels.org/dev/examples/notebooks/generated/stl_decomposition.html
17. https://en.wikipedia.org/wiki/Root-mean-square_deviation