IBM Specialized Models: Time Series and Survival Analysis Final Assignment

By

Ahmed Shahriar Sakib

Introduction

The objective of this project is to perform an exploratory data analysis on London Bike Sharing Dataset and predict number of bike sharing using Deep Neural Network. This is a multivariate time-series problem.

Data Summary

Source: https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset

This dataset contains 17,414 rows and 9 columns. Data was recorded between 2015-01-31 to 2017-01-31

Data Description:

Variables

- "timestamp" - timestamp field for grouping the data
- "cnt" - the count of a new bike shares (**Target variable for this project**)
- "t1" - real temperature in C
- "t2" - temperature in C "feels like"
- "hum" - humidity in percentage
- "windspeed" - wind speed in km/h
- "isholiday" - boolean field - 1 holiday / 0 non holiday
- "isweekend" - boolean field - 1 if the day is weekend
- "season" - category field meteorological seasons:
    - 0-spring ;
    - 1-summer;
    - 2-fall;
    - 3-winter.
- "weathercode" - category of the weather
    - 1 = Clear ; mostly clear but have some values with haze/fog/patches of fog/ fog in vicinity
    - 2 = scattered clouds / few clouds
    - 3 = Broken clouds
    - 4 = Cloudy
    - 7 = Rain/ light Rain shower/ Light rain

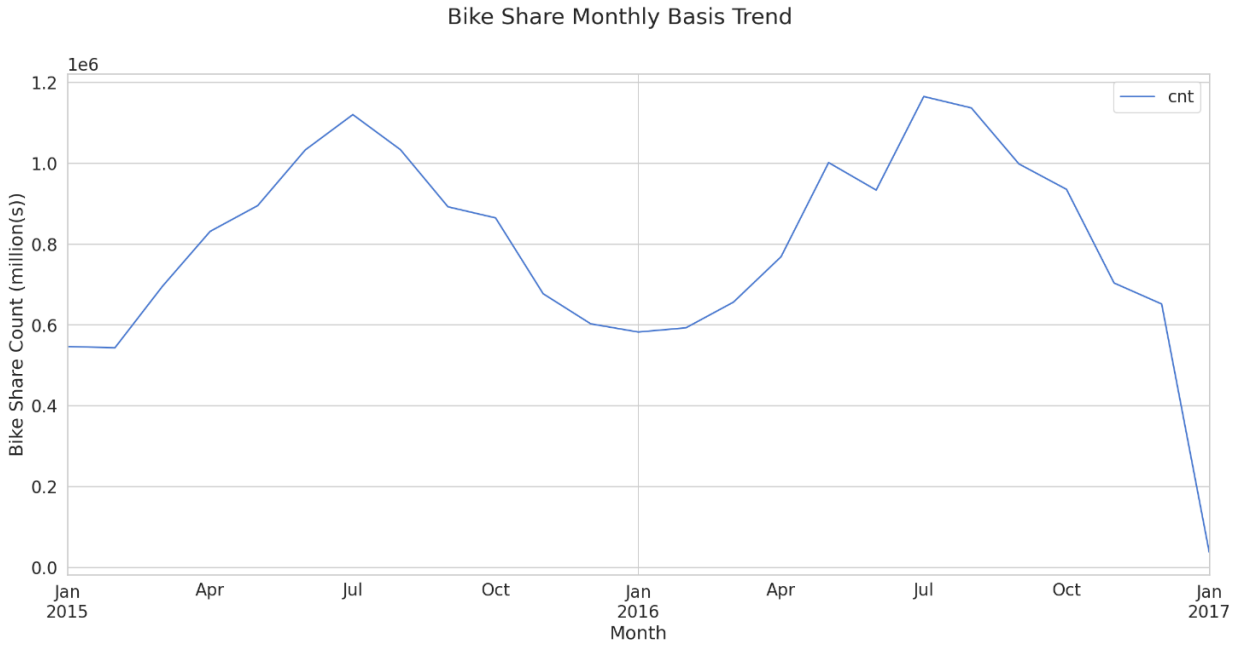
- 10 = rain with thunderstorm
- 26 = snowfall
- 94 = Freezing Fog

Exploratory Data Analysis

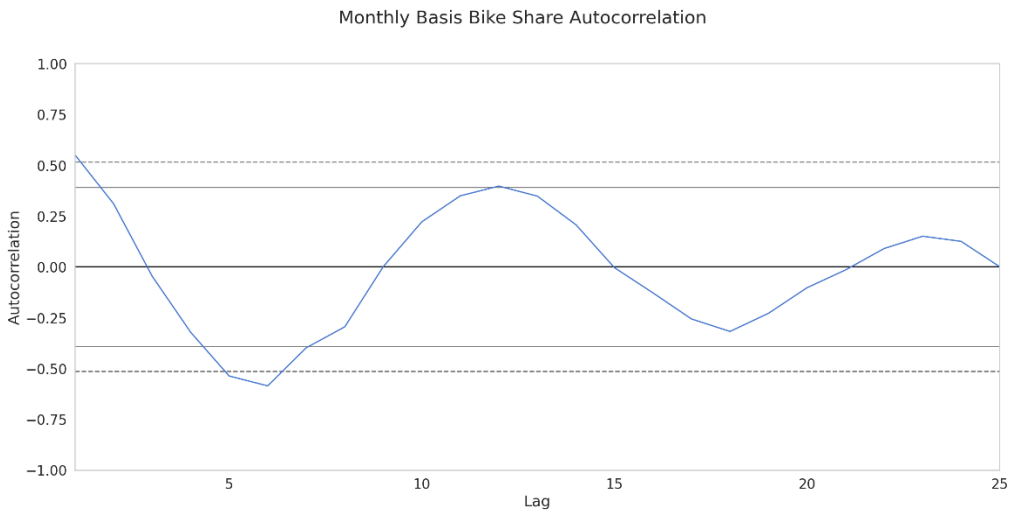
Columns	count	mean	std	min	25%	50%	75%	max
cnt	17414	1143.10	1085.11	0.0	257.0	844.0	1671.75	7860.0
t1	17414	12.49	5.57	-1.5	8.0	12.5	16.00	34.0
t2	17414	11.520	6.61	-6.0	6.0	12.5	16.00	34.0
hum	17414	72.32	14.31	20.5	63.0	74.5	83.00	100.0
wind_speed	17414	15.91	7.89	0.0	10.0	15.0	20.50	56.5
weather_code	17414	2.72	2.34	1.0	1.0	2.0	3.00	26.0
is_holiday	17414	0.02	0.14	0.0	0.0	0.0	0.00	1.0
is_weekend	17414	0.28	0.45	0.0	0.0	0.0	1.00	1.0
season	17414	1.49	1.11	0.0	0.0	1.0	2.00	3.0

- Four new columns were added - 'hour', 'month', 'day_of_week', and 'day_of_month'.

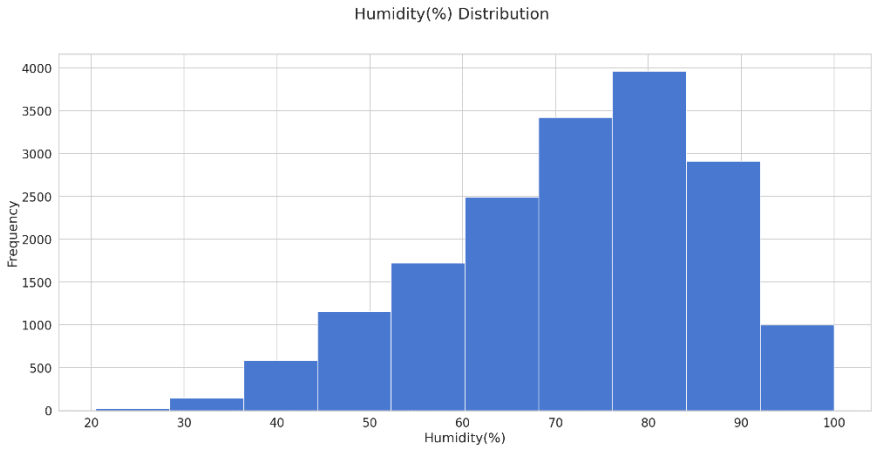
Bike Share Trend ('cnt')



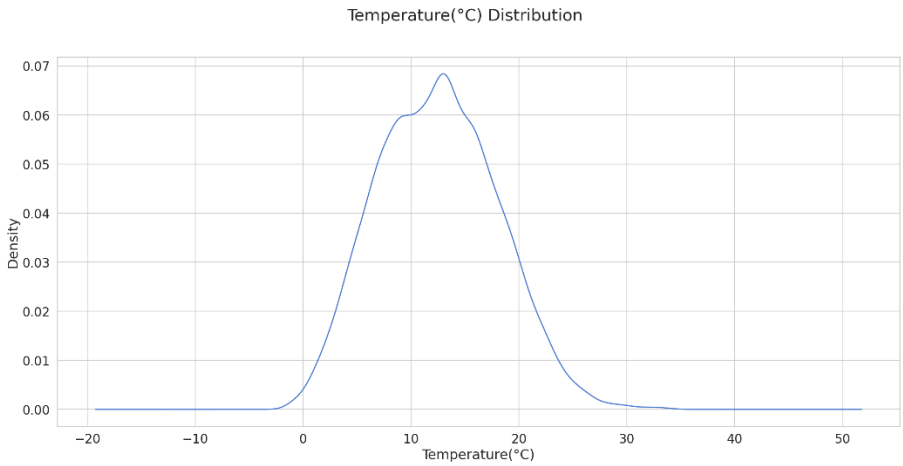
- From the graph above it is clear that it has a strong seasonality component. Summer months are good for bike sharing business



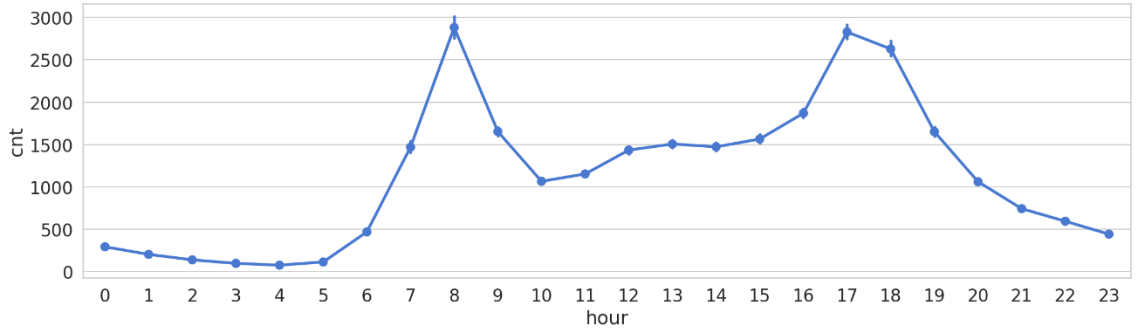
- If month is considered as the lag unit, then, there can be seen cycles of strong negative and positive correlation. Correlation values above the dotted lines are statistically significant

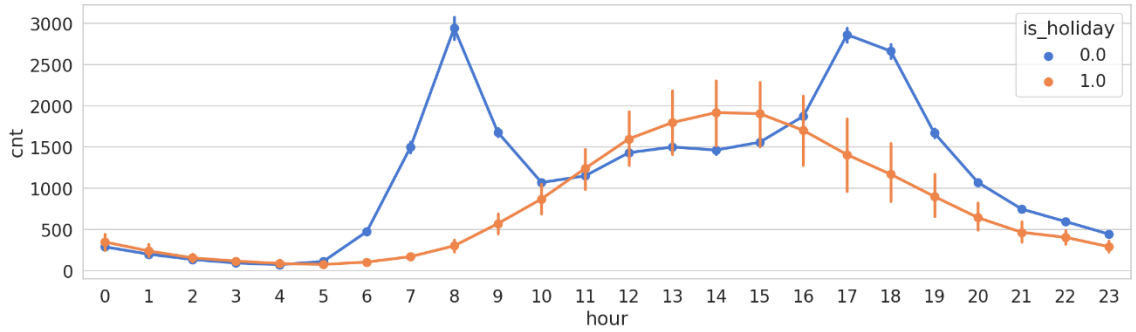


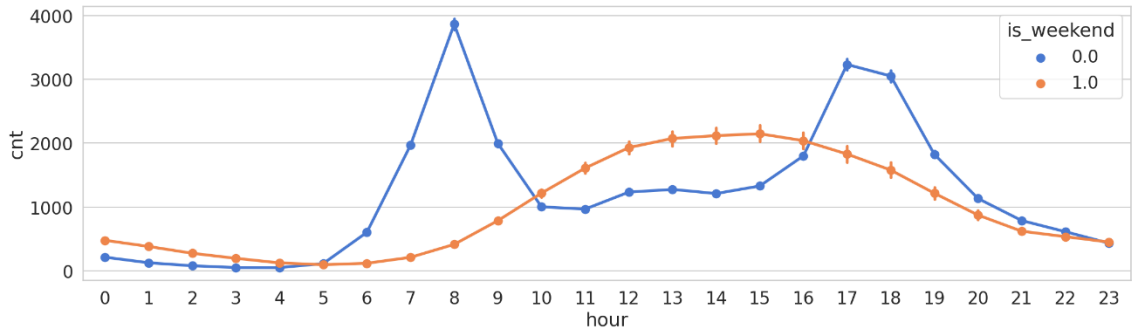
- The humidity mostly falls around 80%, and temperature mostly around 10-18(°C)

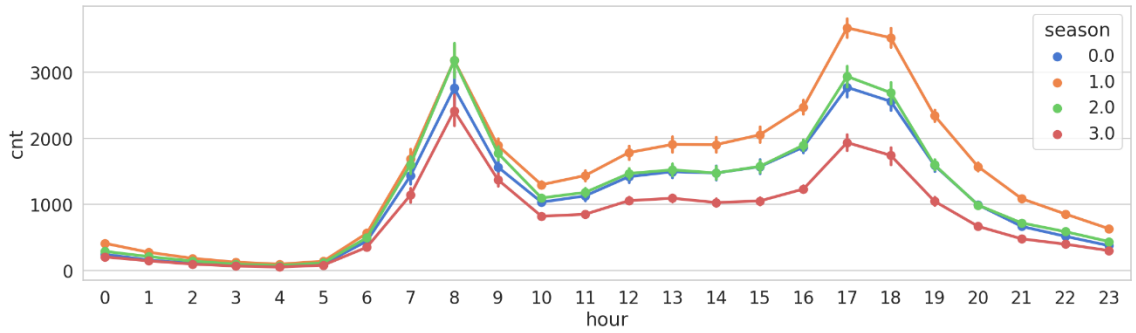


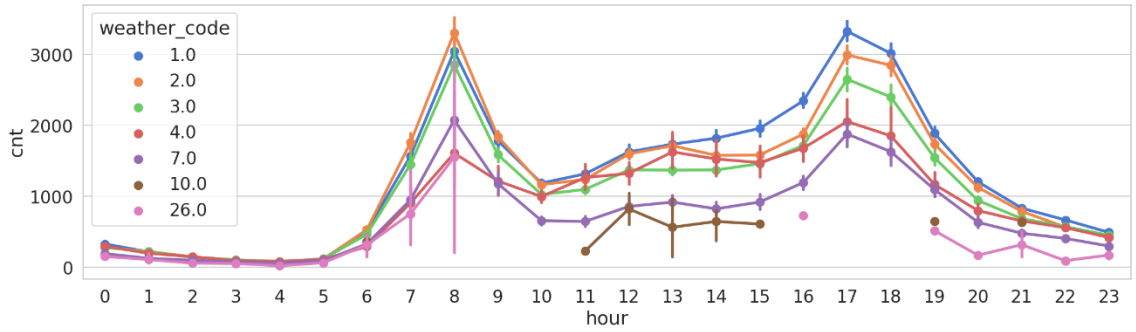
Bike Sharing Trend Over Time (Hourly Basis)











Bike Sharing (hourly basis) based on -

1. Day Type –

- The number of Bike shares differs a lot on an hourly basis based on the day type - workday or weekend or holiday.
- During the workday, people mostly use bike share bikes in two times - during the morning and late afternoon (work in between)
- On the other hand, on weekends (or holidays) early to late afternoon hours seem to be the busiest

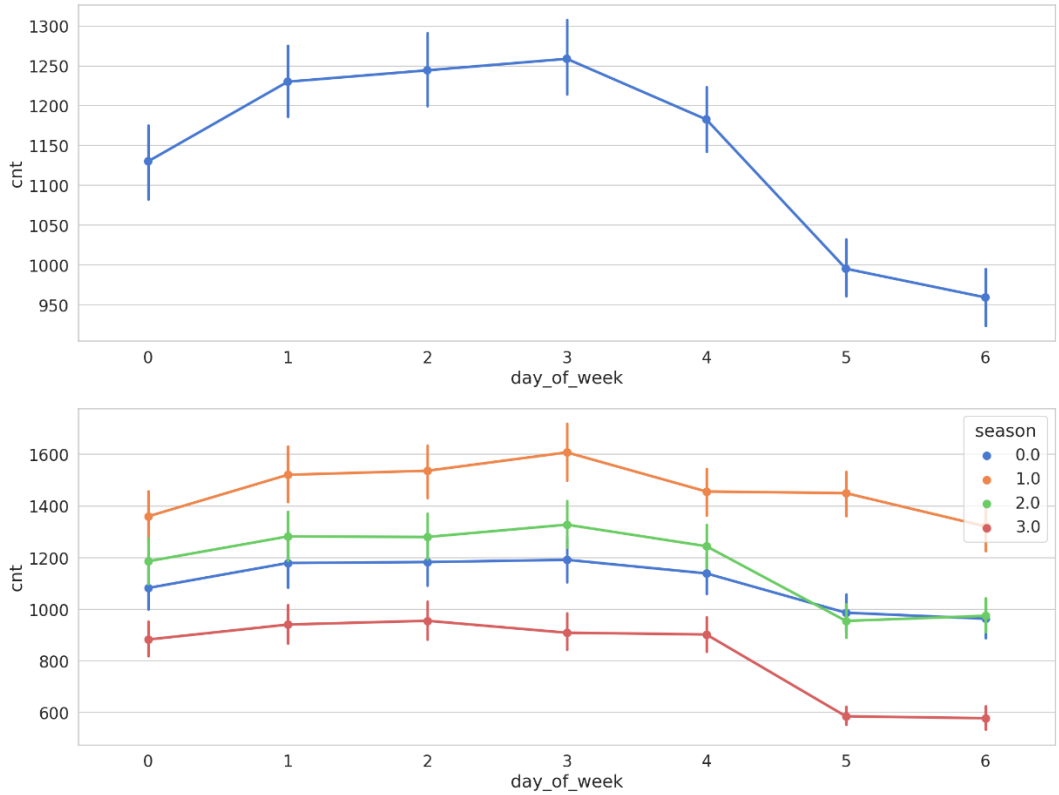
2. Season Type –

- In terms of the ride-share summer season is the busiest, then fall, spring and winter

3. Weather condition –

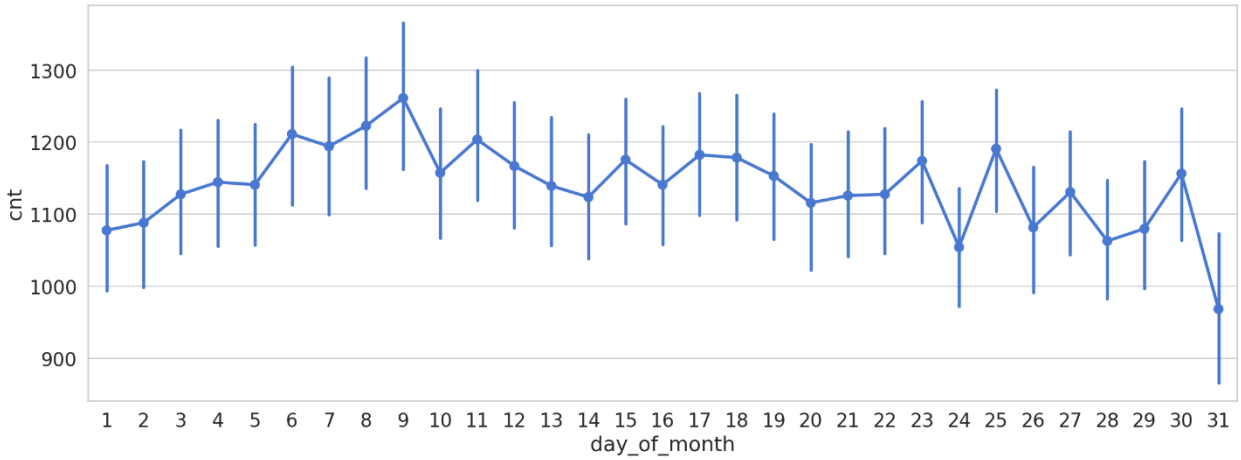
- Bike share is higher when the sky is clear or has few clouds, its lowest when there is rain with a thunderstorm or snowfall
- People don't prefer bike share when there is rain

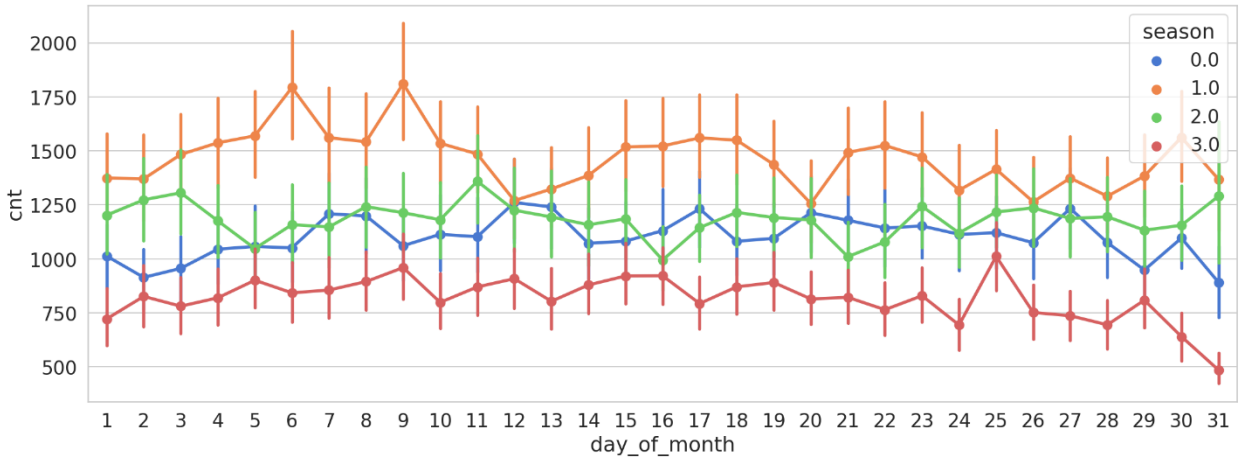
Bike Share Trend Over Time (Weekly Basis)



- Number of Bike share drops on weekends (Saturday and Sunday)

Bike Share Trend Over Time (Month)





- There is a similar pattern between summer and winter

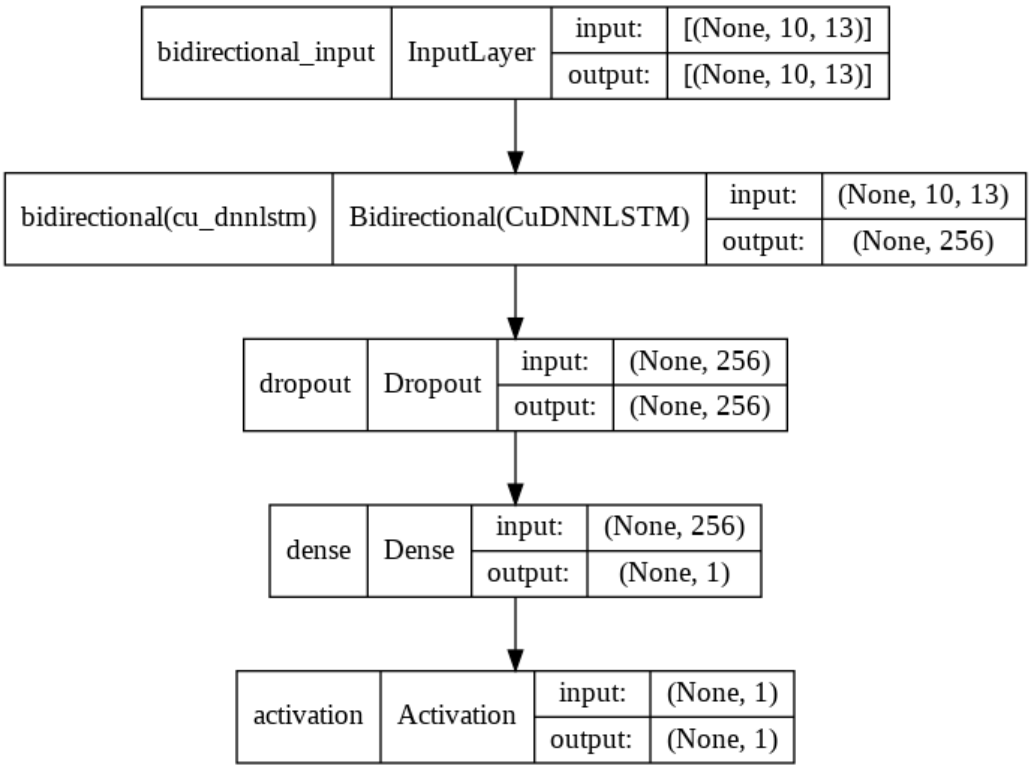
Data Preparation

- Data is divided into train and test set with 90:10 ratio
- Scaled numerical data using robust scaler from sci-kit-learn package which can handle outliers
- Prepare the data for forecasting (Sequence mapping) with 10 timesteps

Modeling

Vanilla Bi-LSTM

Architecture

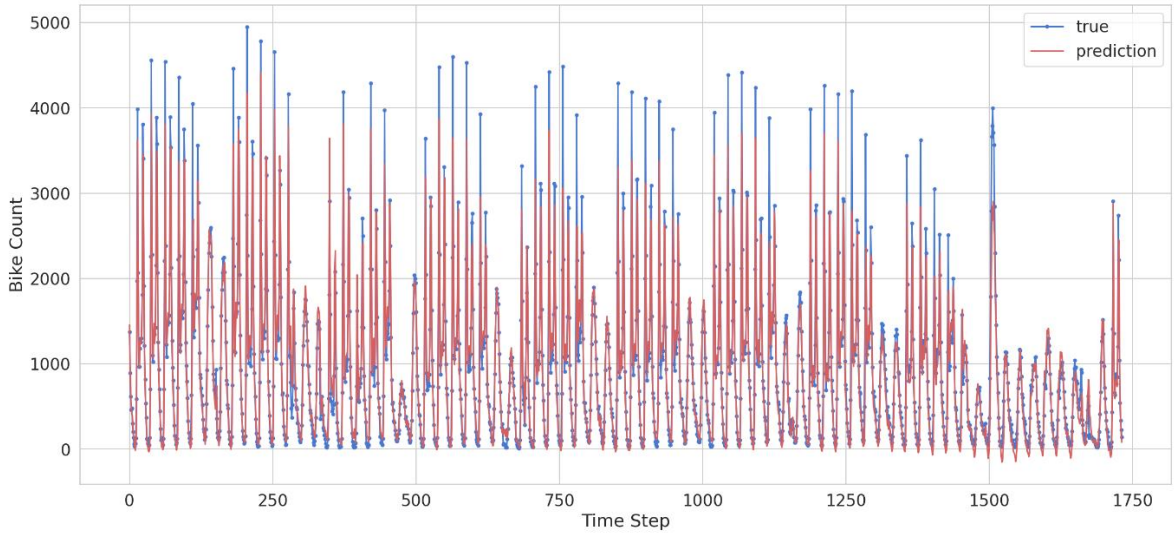


Result:

Evaluation loss **0.0274**

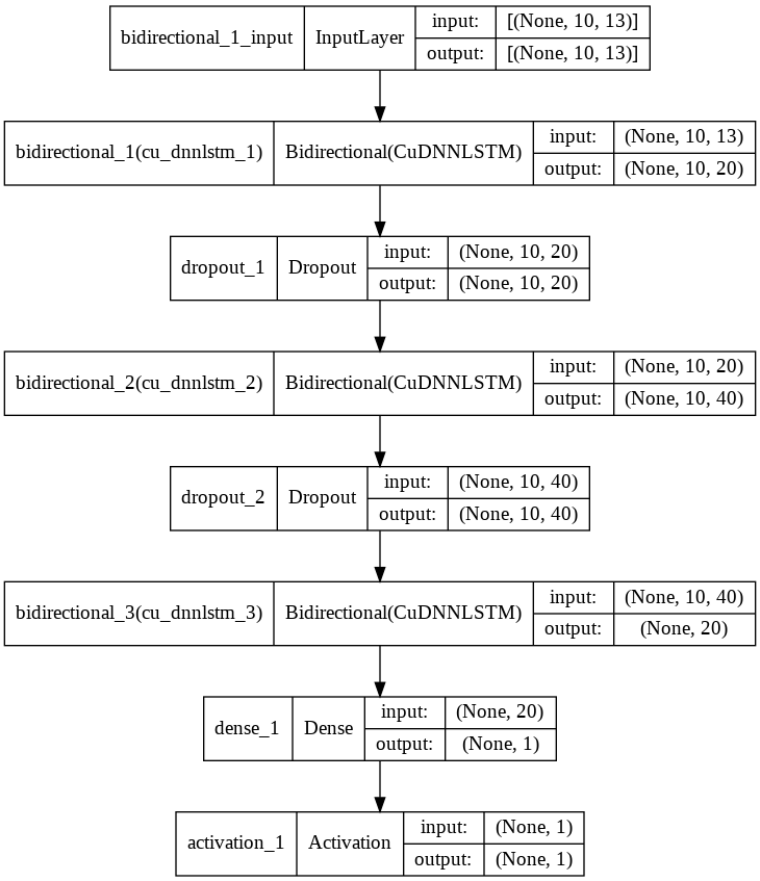
RMSE score: **240.362**

r2 score: **0.9329**



Stacked Bi-LSTM 1

Architecture

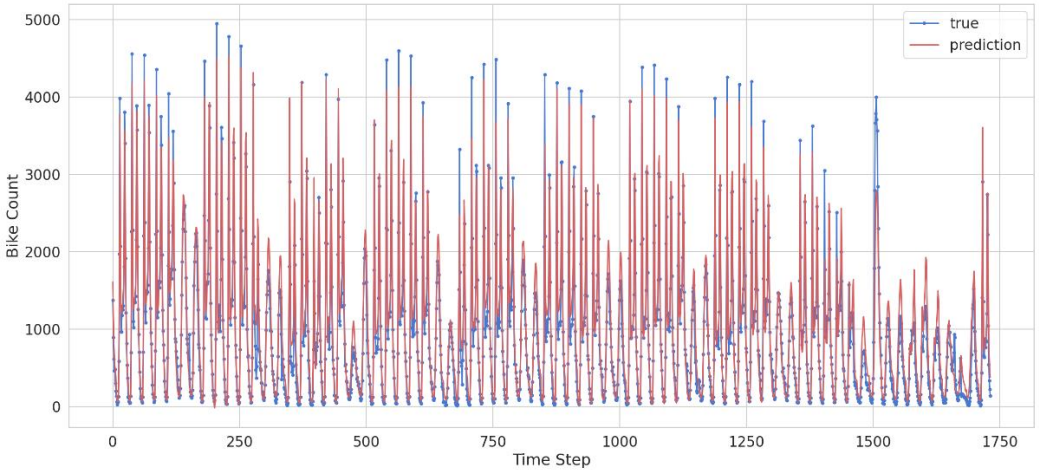


Result:

Evaluation loss **0.0336**

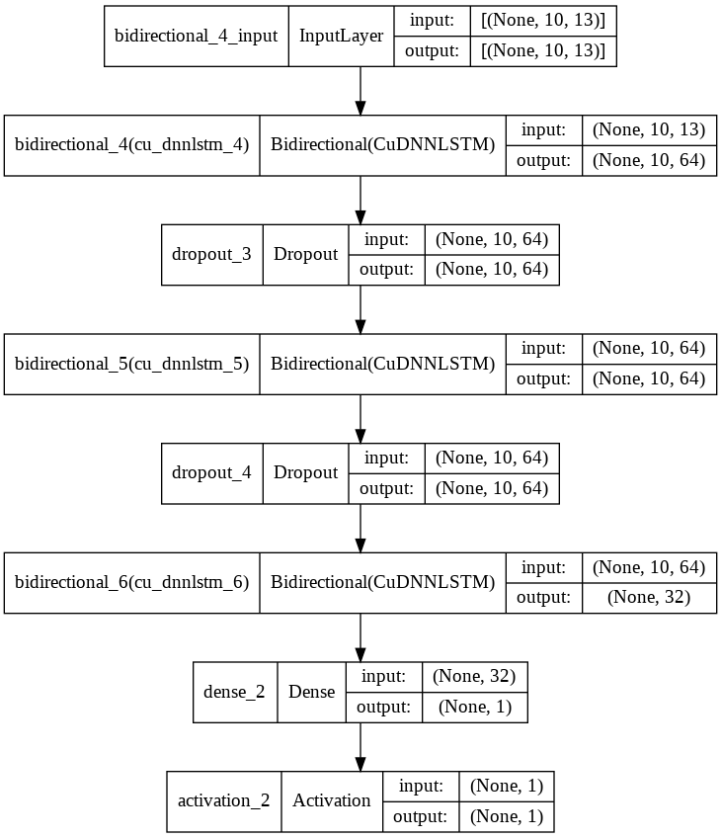
RMSE score: **266.189**

r2 score: **0.9177**



Stacked Bi-LSTM 2

Architecture

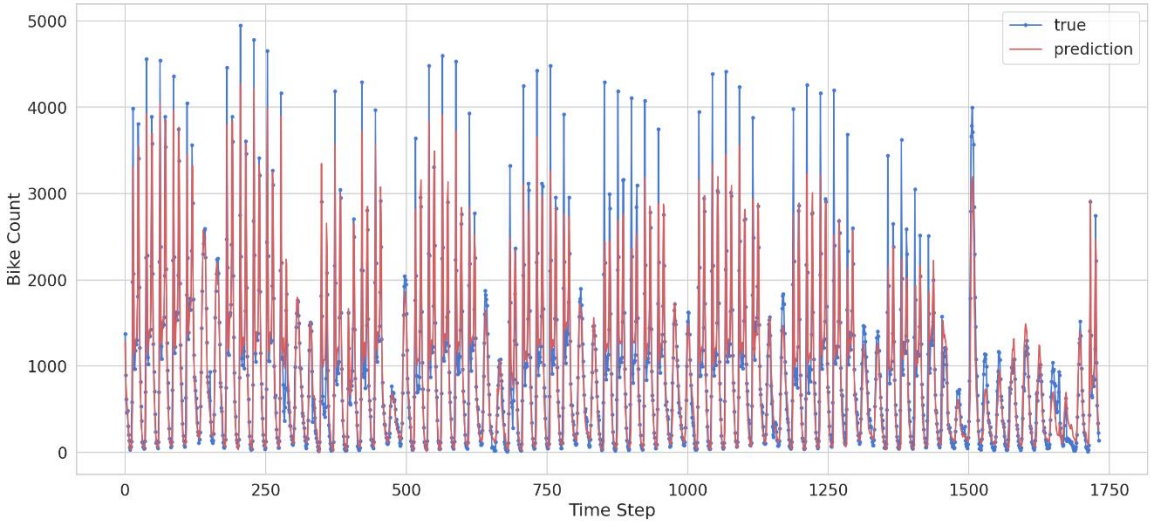


Result:

Evaluation loss **0.05637**

RMSE score: **344.758**

r2 score: **0.862**



Conclusion

- From the above analysis, it's certain that vanilla bi-LSTM model performs better than stacked bi-LSTM models. The loss on the test set is ~0.027
- To improve the model performance different approaches such as - cross-validation, hyperparameter tuning (different number of epochs, number of neurons and layers, optimizer and learning rate etc.,) should be performed.

Full Project Can be found here –GitHub-Demand-Prediction