# **SLR FINAL EXAM PAPER FOR 70 MARKS**

**Data Set Information:** Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousand bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

**Attribute Information:** Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv

1. instant: record index
2. dteday : date
3. season : season (1:winter, 2:spring, 3:summer, 4:fall)
4. yr : year (0: 2011, 1:2012)
5. mnth : month ( 1 to 12)
6. hr : hour (0 to 23)
7. holiday : weather day is holiday or not (extracted from [[Web Link]](http://dchr.dc.gov/page/holiday-schedule))
8. weekday : day of the week
9. workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
10. weathersit :  
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy  
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist  
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds  
    - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
11. temp : Normalized temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale)
12. atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-16, t\_max=+50 (only in hourly scale)
13. hum: Normalized humidity. The values are divided to 100 (max)
14. windspeed: Normalized wind speed. The values are divided to 67 (max)
15. casual: count of casual users
16. registered: count of registered users
17. cnt: count of total rental bikes including both casual and registered18 - GT Turbine decay state coefficient. (Target variable)

Instructions: Based on the above dataset the students are expected to follow the below mentioned steps.

* **Instructions:**  Based on the above dataset the students are expected to follow the below mentioned steps.

|  |  |  |
| --- | --- | --- |
| STEPS | Instructions | MARKS |
| 1.Data Ingestion: | Read the dataset (tab, csv, xls, txt, inbuilt dataset). Do the descriptive statistics and do null value condition check, write an inference on it**.** | 5 |
| Split Data set: | (train – 70% and test – 30%), Using test and train data, calculate R square value and check if the model Over-fit / under-fit model.  Write a detailed inference on it. | 5 |
| 3. Build a base model: | (which will include all the parameter as input) and check which are the significant parameters based on the summary of the model and write a brief summary | 5 |
| 4. EDA - Plots: | 1. Check for Outlier using Box plot and apply the formula to check how many of these data has outliers 2. Apply correlation plot and display with color code and numerical indication 3. Plot best X with Y and find how the Y is varying with X. Like if X is continuous, go for scatter plot. if X is categorical, go for box plot. 4. Apply density plot and check if the data is normally distributed | 5 |
| 5. EDA - Inference: | Write detailed inference (not observations) about the question 4, few pointers like Outlier detection, outlier treatment, conditions for linearity, impact because of transformation on variables, distributions etc. | 5 |
| 6. Multi Collinearity Check: | Check if there is any multi-collinearity for the given dataset using different techniques and write inference. | 5 |
| 7. Apply Feature selection technique: | (at least 3 techniques – your choice) to identify which are the best features and build models accordingly. Write inference on the model created, detailed out how Feature selection model is better than the previous models | 10 |
| 8. Data Transformation: | Apply scaling to the numerical data and apply label encoding to the categorical data, build a model with the transformed data and check if this transformation is adding value or improving the accuracy. Write detailed inference on it. | 10 |
| 9. Apply feature engineering: | if needed to improve your accuracy and write inference. Summarize how EDA is helping to do the feature engineering and how to improve the model accuracy. Relate EDA with you model and incorporate your inference | 10 |
| 10. Output | Display the output in tabular format, which will compare all the model, model tuning methods, metrics.  Compare all the models create and write detailed summary on the **Final** **model** and why is this best model, enrich with the factor why it is best model based on what? (e.g. EDA inference, transformation, feature selection etc.) | 10 |