Final Data Science Project

Project Overview

For better environmental and social welfare, the city of Ourra has recently implemented a bike sharing system called Drpia. The secretary of the department of transportation is looking to understand drivers of rental bike demands so as to properly balance the available supply at any given time.

Congratulations! You have been recently hired as a Data Science Consultant for the department of transportation. Your task, as outlined above, is to **predict and model rental bike counts (demand)** for the bike sharing system. In order to tackle this problem, you have been given a set of historical data of weather as well as seasonal information.

Two datasets are included: *train.csv* and *test.csv*. The training dataset consists of 6552 observations that include the rental count along with a number of weather and seasonal features. The test dataset contains 2208 observations and does not include the rental bike count. You will use this training data to build, implement and test models. You will then generate predictions based on whatever you determine is the "best" model. The data dictionary on the last page of this document gives a description of each variable.

This project should be treated as a take-home exam and is to be completed completely independently. You may not consult with anyone about any aspect of the project (including even simple coding questions) other than the professor and TA.

The final data science project is due by **11:59 PM EST** on **Tuesday April 19.**

—— Absolutely no late projects will be accepted ——

Project Deliverables

1) **Predictions:** A single csv file with 2208 test observations named "testing_predictions_XXX.csv" where XXX is your student ID number that contains three labeled columns in the following order:
    - id: bike Id provided in the test dataset
    - price: predicted rental bike counts.
    - student_id: your student id number
2) **Technical Report:** A pdf report that outlines your process from start to finish in technical detail. Please name it "technical_report_XXX.pdf" where XXX is your student ID number. The report should NOT include any code. You may use at most 4 figures or tables. Limit of 5 pages (double spaced). This should (at a minimum) touch on the following:
    - Introduction and description of exploratory data analysis.
    - Identification of Data oddities e.g. missing data, extreme values, etc. and how you handled them.
    - Summary of models considered. How many models seemed to perform "best" in terms of predictive accuracy? How'd you measure?
    - What were the most important variables? How did you measure variable importance? Were the variables deemed most important consistent across the top-performing models?
    - What were the most challenging aspects of this particular dataset? Were you able to mitigate these issues? Do you really trust your "best" model? If your job depended on this model, how worried would you be? Is there other information you may want in order to improve the final model/predictions/recommendations further?
3) **Final (non-technical) Report:** Discuss your findings to a non-technical decision maker. Introduce your project and summarize some key findings that could be useful to understand rental bike demand. Limit 1.5 pages (double-spaced). "final_report_XXX.pdf" where XXX is your ID number. (Note: Non-technical decision maker means that they will not know phrases concepts such as (but not limited to): lasso, ridge regression, random forests, gradient boosted trees, tuning parameter, cross validation, mean squared error, bias, variance, overfitting, etc.)
4) **Code:** A single (or multiple) .R or .Rmd named "code_XXX.R" where XXX is your student Id. Your code should be thoroughly commented and able to be run from another machine provided necessary packages and data are loaded.

Rubric

1) **Accuracy (10%):** Model predictions on the test dataset will be graded based on Mean Squared Error. This is largely an "all or nothing" category -- to earn full points, you simply need to have a model with lower test MSE than than a pre-established base rate. You do not need the best possible model available and you should not spend all your time trying *ad hoc* things in search of the lowest possible MSE. This project isn't a "predictive competition" like you might find on kaggle.com. This goal of this project is to find strong model(s) obtained by correct reasoning and to understand what those variables imply as well as the uncertainty surrounding them.
2) **Technical Report (50%):** The technical report will make up a significant chunk of your grade and should contain the guts of your process. The four main components of the technical report you will be graded on are:
    - *Introduction / EDA* – This should give an overview of the problem, general information of the data, identify data oddities, summary statistics, etc.
    - *Methods Overview/Details* - This should contain a summary of the methods explored and the various approaches that were considered.
    - *Summary of Results* - This should provide an overview of all of the results obtained. Comment on overall trends, contradictions between models, etc. You can include a table here if it helps summarize the findings. Include test error estimates from the best overall model(s) as well as from the model(s) you ultimately chose to rely on.
    - *Conclusions / Takeaways* - Based on the results described in the previous section ('Summary of Results'), describe what you feel can safely be concluded. If there are further tests/models that you think would be relevant to pursue given the overall results, note this.
3) **Final (non-technical) Report (30%):** How would you explain the results to someone interested in your findings that doesn't have a statistics background? Discuss your project and findings in a non-technical manner. Identify and summarize at least 3 specific key takeaways from your work. These can include any useful and potentially actionable findings and/or specific aspects of the work that decision makers should keep in mind.
4) **Quality of Code (10%):** Does code run from another machine provided necessary packages and data loaded? Is code "readable" and well commented.

Data Dictionary

Variable	Type	Description
Count	Numeric	Rented bike counts
Date	Character	Date
Hour	Numeric	Hour of the day
Temperature	Numeric	Temperature in Celsius
Humidity	Numeric	Humidity (%)
Wind	Numeric	Wind speed (m/s)
Visibility	Numeric	Visibility (10m)
Dew	Numeric	Dew point temperature in Celsius
Solar	Numeric	Solar Radiation (MJ/m2)
Rainfall	Numeric	Rainfall(mm)
Snowfall	Numeric	Snowfall (cm)
Seasons	Character	Winter, Spring, Summer, Autumn
Holiday	Character	Holiday/No holiday
Functioning	Character	NoFunc(Non Functional Hours), Fun(Functional hours)
ID	Character	Unique identifier