ESILV - PYTHON FOR DATA ANALYSIS

Project 2022

Made by Charles Delemazure

TABLE OF CONTENTS

SEOUL BIKE SHARING DEMAND DATA SET

I. Reading and Understanding the data

Bike-sharing demand, the dataset, the target, the features, thoughts

II. Preprocessing and data visualization

Data scaling, data encoding, features selection, variable creation, visualization with Seaborn and Pyplot

III. Model predictions

Linear Regression, Random Forest, Decision Tree, XGBoost, Bayesian Ridge, AdaBoost, Support Vector Regression, LGBM, KNN

IV. API with Streamlit

Seoul bike sharing demand

Bike sharing is one of the ways to reduce urban traffic. It also reduces air pollution by reducing the number of cars on the road. The bike sharing system is a new generation of traditional bike rental systems, and the entire process has been automated. Users can borrow bicycles for free or for a fee and return them to another place.

The Data Set

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

8760 rows

13 Features

1 target

The target

RENTED BIKE COUNT

The hypothesis in the research is that the bike sharing is highly related with the time of the day, season, and weather conditions. The research will try to predict the bike shares in the future.

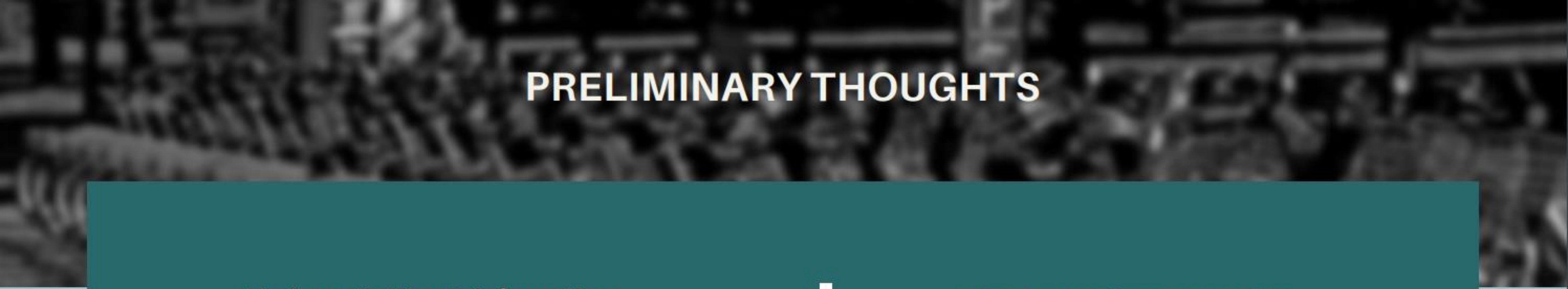
The features

TIME INFORMATION (5)

- Date - year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

WEATHER INFORMATIONS (8)

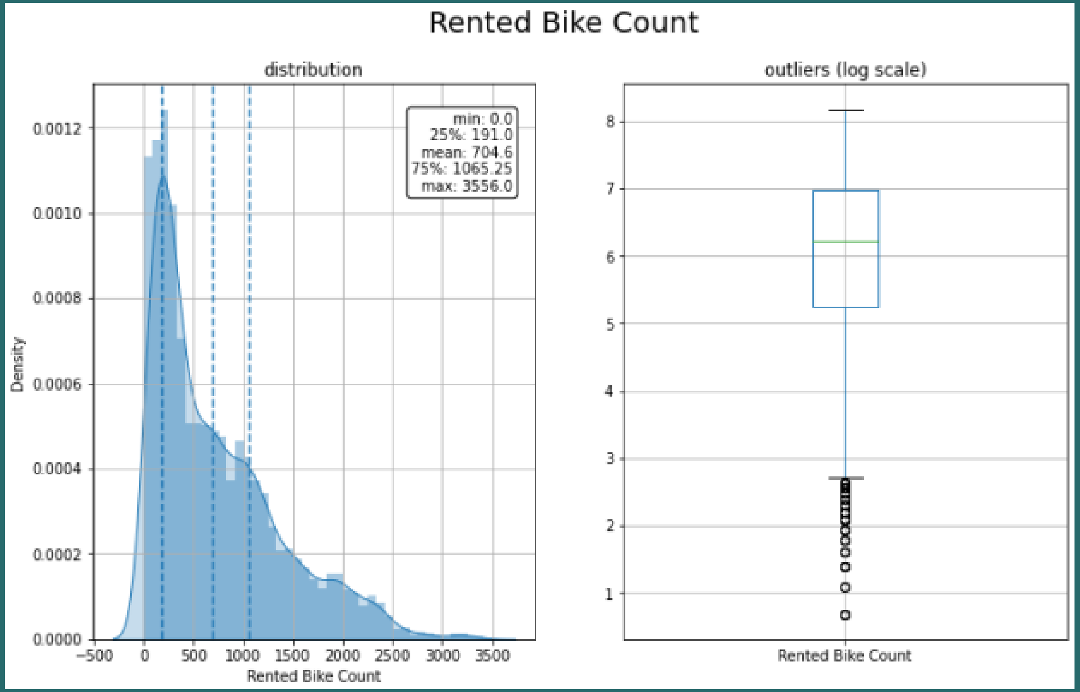
- Temperature -Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm



PRELIMINARY THOUGHTS

I believe that basic information such as **Temperature** and **Hour** have a good impact on the number of bike rented. Logically, people are more likely to use bikes during warm days.

The datas of some weather conditions may to be significant for the predictions. Seoul is not a city known for its snow so we may delete the Snowfall information.

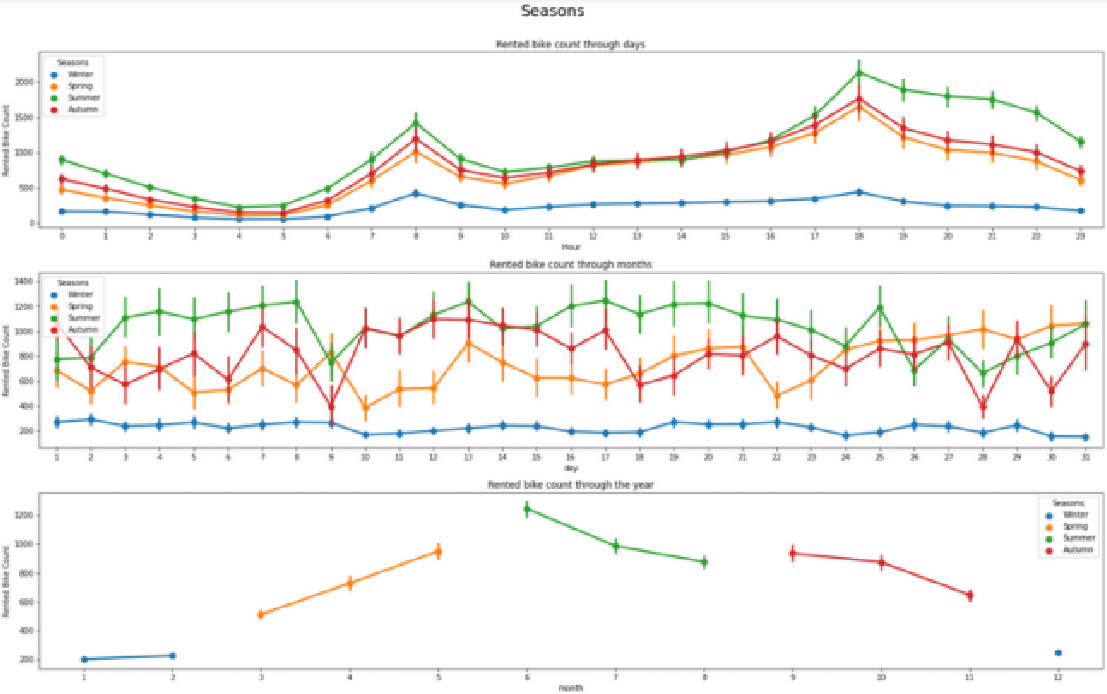


Visualization with seaborn

I have made many graphs using Seaborn. I liked it because it allows me to make very nice plots in a few lines and it is easy to use. We can see on the example just bellow that there is a higher bike sharing demand during summer, from june to august. It is also very useful to visualize the correlation of the variables between them.

Visualization with pyplot

I know that the Seaborn library is based on matplotlib and that mastering it is necessary to make more accurate and customized graphics. Still in an effort to understand the data and theircorrelations with our target, we learned to use matplotlib to plot relevant graphs.



ENCODING DATA

One hot encoding

Holiday and Functioning Dat are categoriacal datas where no relationship exists between their categories.
It involves representing each categorical variable with a binary vector.

Holiday : Holiday - 1 / No Holiday - 0

Functioning Day : Yes - 1 / No - 0

Label encoding

Seasons is a categorical data but doesn't have an ordered relationship between the categories.
I mapped the Seasons column like this

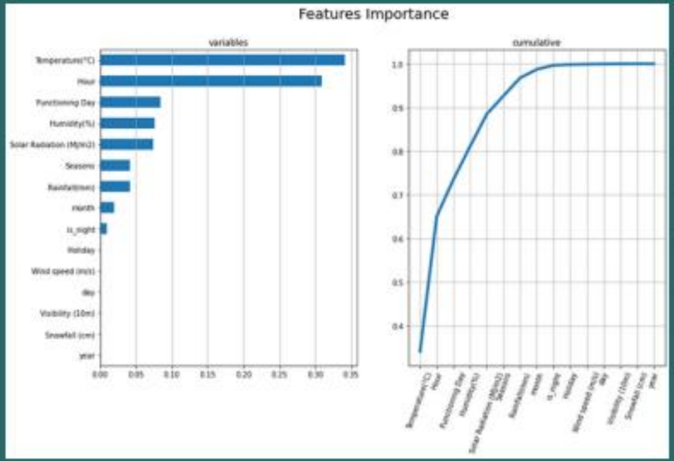
Winter - 1
Spring - 2
Summer - 3
Autmun - 4

New variable / Feature selection/ Data scaling

Creation of the variable is_night which depends of the column Hour and specify how dark it is outside

Feature selection

I use the Gradient Boosting Regressor model to calculate the import for each attribute in the dataset. This model construct boosted trees, and the most important features are the ones which help the most constructing the boosted decision trees, the most useful and valuable features.



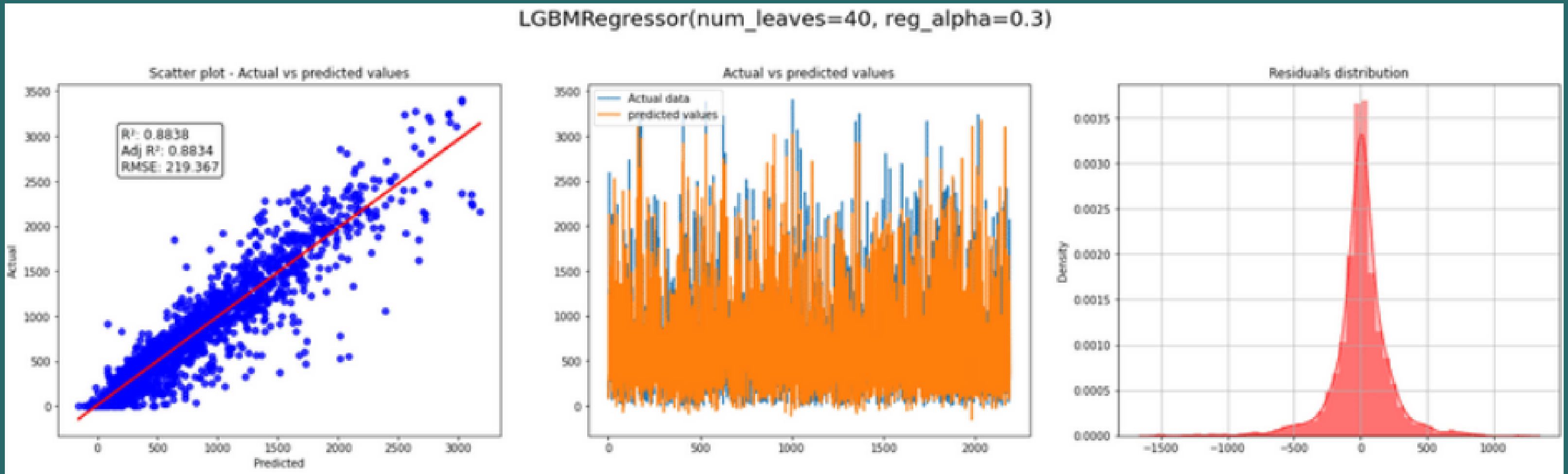
After i scale the data. I chose the standardization because it is more efficient to compare measurements that have differents units.

MODELS AND PREDICTIONS

RANDOM FOREST, DECISION TREE, LINEAR REGRESSION, BAYESIAN RIDGE, XGBOOST, ADABOOST, LGBM, SVR, KNN

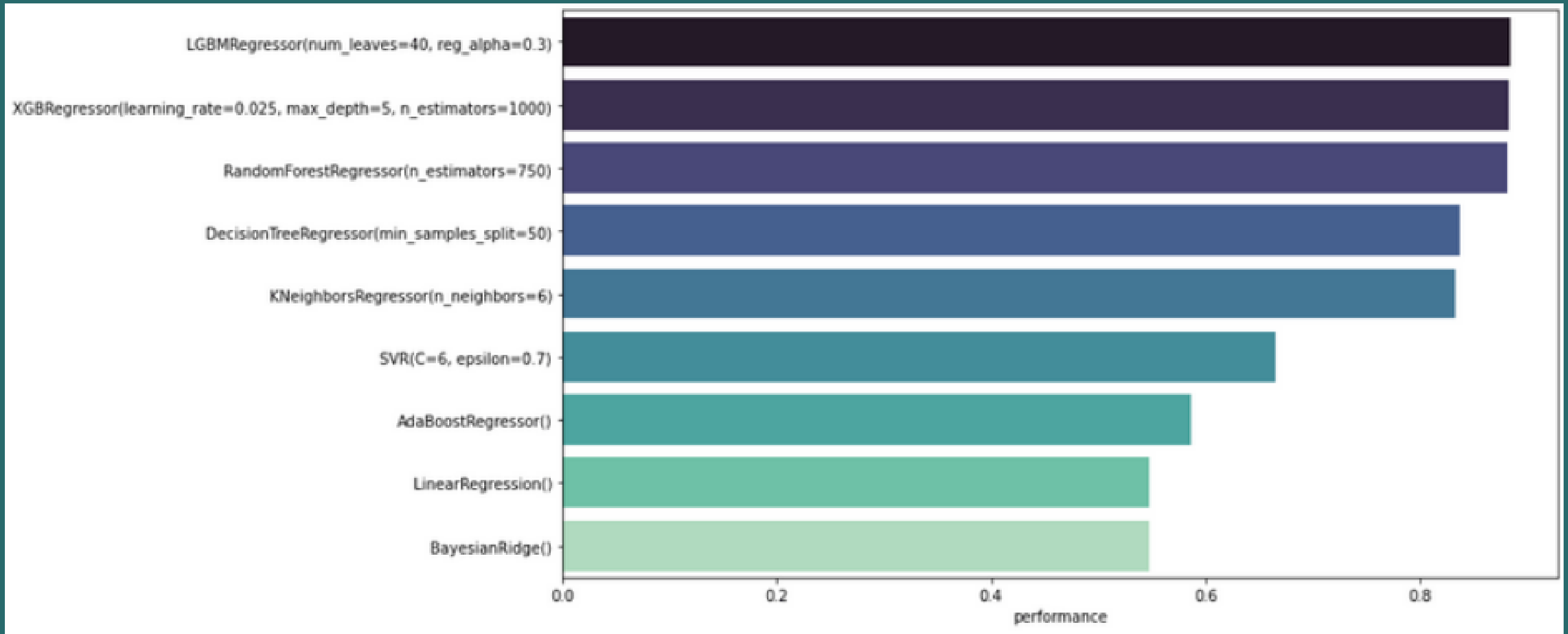
BEST MODEL

LIGH GBM



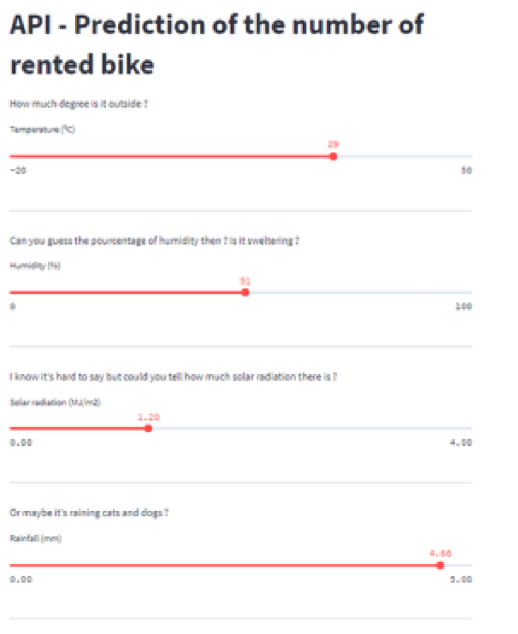
After all this, I try differents models to try to have the best score (the best r2 squared) . I used the GridSearchCV method seen in class to optimize my hyperparameters. The results obtained are pretty satisfying, the tree first ones are very closed with a score of :

- LightGBM : 0,8838
- XGBRegressor : 0,8822
- RandomForestRegressor : 0,8808



API with Streamlit

Finally, i have made an API with StreamLit which allows you to select (with sliders and element boxes) the values of the time and weather conditions and predict the number of rented bikes necessary.



API - Prediction of the number of rented bike

How much degree is it outside ?

Temperature (°C)

Can you guess the pourcentage of humidity then ? is it sweltering ?

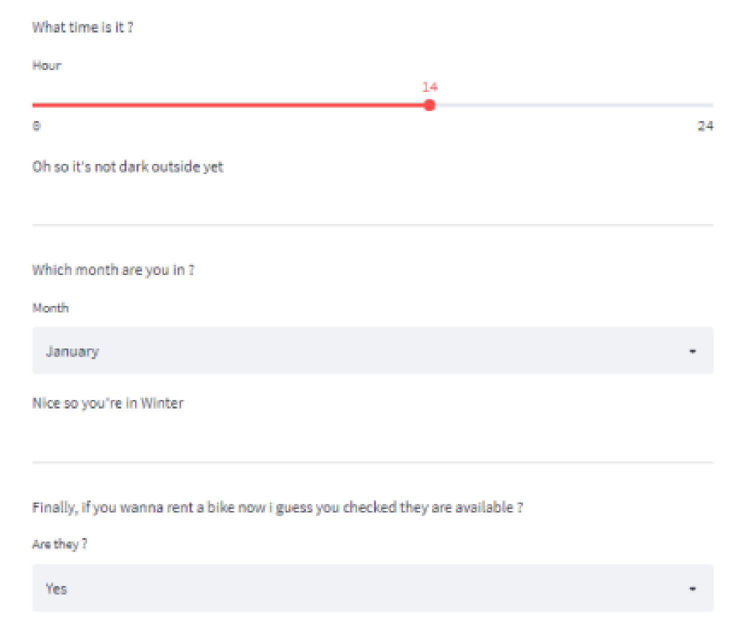
Humidity (%)

I know it's hard to say but could you tell how much solar radiation there is ?

Solar radiation (MJ/m2)

Or maybe it's raining cats and dogs ?

Rainfall (mm)



What time is it ?

Hour

Oh so it's not dark outside yet

Which month are you in ?

Month

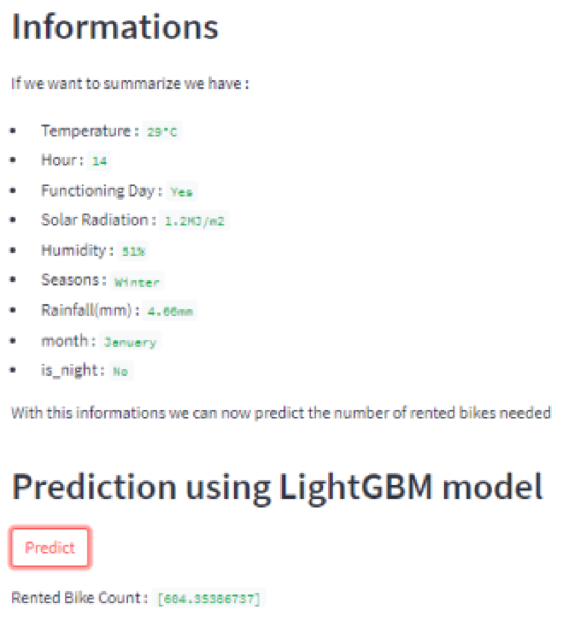
January

Nice so you're in Winter

Finally, if you wanna rent a bike now i guess you checked they are available ?

Are they ?

Yes



Informations

If we want to summarize we have :

- Temperature : `29°C`
- Hour : `14`
- Functioning Day : `Yes`
- Solar Radiation : `1.2MJ/m2`
- Humidity : `51%`
- Seasons : `Winter`
- Rainfall(mm) : `4.66mm`
- month : `January`
- is_night : `No`

With this informations we can now predict the number of rented bikes needed

Prediction using LightGBM model

Rented Bike Count : `[684.35386737]`