Bike Sharing Demand Prediction

Python for Data Analysis Final Project

Théo DEMESSANCE – ESILV A5
January 2021

Getting started

- Bike sharing systems are a means of renting bicycles.
- People are able rent a bike from a one location and return it to a different place on an as-needed basis.
- Currently, there are over 500 bike-sharing programs around the world.
- The data generated by these systems makes them attractive :
  - The duration of travel, departure location, arrival location, and time elapsed
  - Meteorological data
  - Can be used to study mobility in a given city

Seoul City

- 1500 bike renting stations all over the city

- Built to be utilized by all kinds of people including women, elderly persons and infirm.

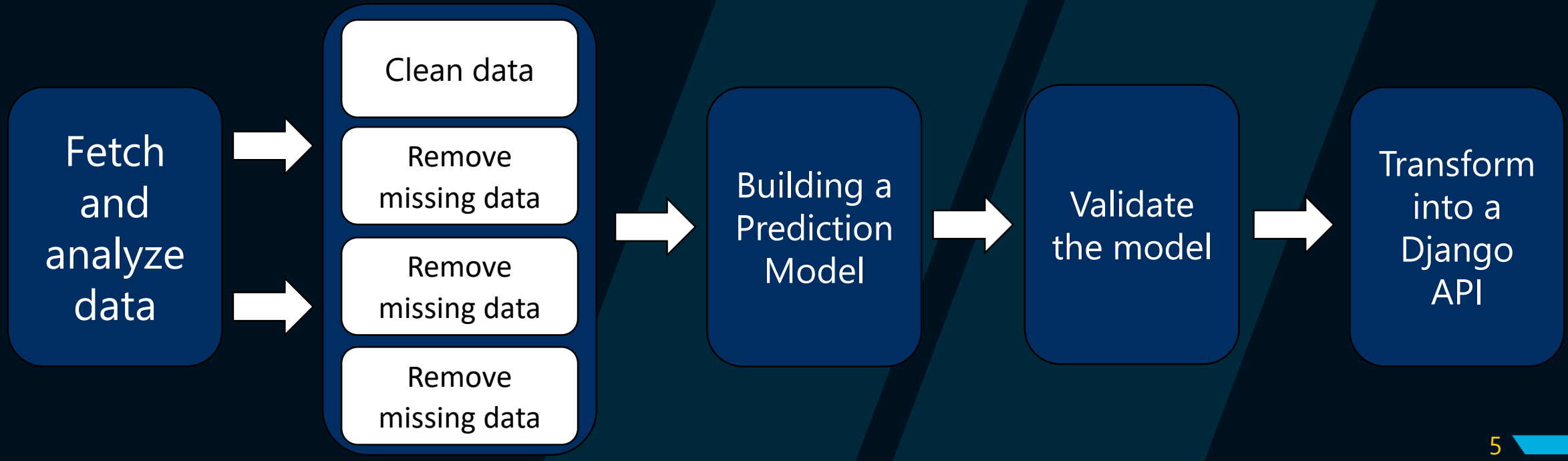
- Manufactured using durable and light-weight materials.

Problematic

- Different objectives :
  - To perform a data analysis of the dataset
  - To modelize different Machine Learning algorithms which predict the number of bike rented in a given date
  - Transformation of a model into a Django API

All different results and code can be found at : https://github.com/TheoDemessance/Seoul-Bike-Sharing-Demand

Proposed method



About the data

- The dataset represents records starting from the 1st December 2017 to the 30th november 2018.
- Each line is a record of the number of bike rented in the city with meteorological data.
- We have 24 records (one per hour) over 365 days, i.e. 8760 line with 14 columns.

Dataset Description

Field	Data Type	Description
Date	Datetime64 & hour	The date and hour of the record
Rented Bike Count	Int64	The number of bikes rented for the hour of the date
Temperature	Float64	The temperature in °C recorded
Humidity	Int64	The relative humidity in %
Wind speed	Float64	The speed that the air is moving in m/s
Visibility	Int64	Distance (10m scale) of clearness
Dew point temperature	float64	temperature to which the air would have to cool to reach saturation.
Solar radiation	Float64	the electromagnetic radiation emitted by the sun in MJ/m^2
Rainfall	Float64	Height of the precipitation in mm
Snowfall	Float64	Height of the snowfalls in cm
Seasons	Category	Winter, Spring, Summer and Autumn
Holiday	Category	Holiday and No Holiday
Functioning Day	Category	Day Worked or Not

Feature Engineering

- We create new columns:
  - By splitting the date, we isolate the year, month, day and day of the week
  - We drop the date column after the split
- We encode the categorical data using One-Hot encoder
- We log the target data and we drop the outliers in order to have a better distribution of our data

Metric used

- RMSE : Root Mean Squared Error

- The use of RMSE is very common, and it is considered an excellent general-purpose error metric for numerical predictions.

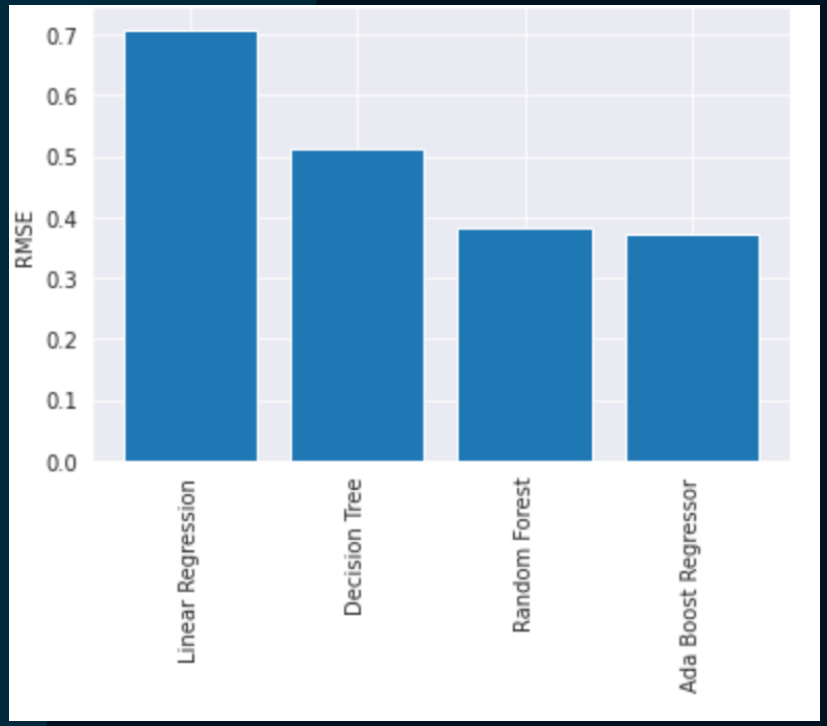
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2}$$

- where O_i are the observations, S_i predicted values of a variable, and n the number of observations available for analysis.

Different models used

We model our data using:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- Ada Boosting Regressor



References

1. Bike Sharing Demand : https://www.kaggle.com/c/bike-sharing-demand

2. Dataset : https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

3. A rule-based model for Seoul Bike sharing demand prediction using weather data, Sathishkumar VE & Yongyun Cho.