

4-th year of Engineering School

PYTHON FOR DATA ANALYSIS PROJECT

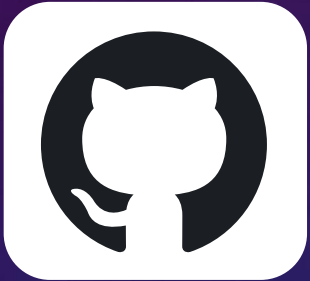
BY ESTELLE FIKET

01/05/2022

INTRODUCTION

For this project, I worked alone.

The objectives of this project are to implement a python notebook to analyse a dataset. What we have to do:



Create a Github
Repository
(to version some
document on it)



Implemented a
notebook with my
work on the dataset
assigned



Explaining the
context, the problem
and what I have done
on the notebook



Implemented a Flask
API

REQUIREMENTS

For each dataset, I used the language Python (version: 3.9 with respect to the PIP8 rules), and the following libraries:

- Pandas (Data Manipulation and Data Analysis)
- Seaborn and Matplotlib (Data Visualization and Data Analysis)
- Sklearn (Machine Learning)
- Flask (API)

I used DataSpell IDE to work on this project.

All the instructions to use the code implemented in the repository are written on the Read.me file.

OBJECTIVES FIXED FOR EACH DATASET

- To check if the information given is correct.
- To get other information in applying analysis on the different features such as:
 - Type of problem (regression or classification),
 - Which feature is the target,
 - Type of the different features + some statistics,
 - Checking the quality of the dataset.
- To fit several model
- To choose the best models and to search for the hyperparameters
- Discussion/Conclusion about the dataset, the results obtained, and the project in general.

WHAT WILL BE EXPLAIN IN THIS SLIDE SHOW?

Table of Contents

In this slide show, I will only explain the context, the problem for each dataset, and what I have done in the notebooks. Each step is already detailed in each notebook.

And then I will provide some explications about the main results and a discussion to conclude the analysis of each dataset.

Dataset

The dataset that has been assigned to me was the Blocks Classification one. During all the process, some things pushed me to choose another dataset in the end to work on, and to implement other method that we have seen in this course. My goal was to implement almost each subject that we have seen from each practical work.

I will first present my work on this first dataset. Then, I will detail what I have done on the second dataset (Seoul Bike), chosen randomly among the dataset from the Google Sheet shared to the class. To conclude this slideshow, a description of the almost finished implemented API will be given with a discussion around this project.

PAGE BLOCKS CLASSIFICATION DATASET

Link: <https://archive.ics.uci.edu/ml/datasets/Page+Blocks+Classification>

BLOCKS CLASSIFICATION: PRESENTATION

There are 5473 examples coming from 54 distinct documents and each observation concerns one block.

The problem consists in classifying all the blocks of the page layout of a document that has been detected by a segmentation process. This is an essential step in document analysis in order to separate text from graphic areas. There are five classes:

- Text (1),
- Horizontal line (2),
- Picture (3),
- Vertical line (4)
- Graphic (5).

BLOCKS CLASSIFICATION: PROBLEM

Supervised → Classification Problem




10 Features

Classifying all the blocks of the page layout of a document:



BLOCKS CLASSIFICATION: STEPS FOLLOWED FOR THIS WORK



- Data Collection
- First Exploratory Data Analysis
- Data Preprocessing
 - Missing Values
 - Null & Zero Values
 - Outliers
 - Categorical Features
- Imbalanced Class & Outliers Management
- Second Exploratory Data Analysis
 - Univariate Analysis
 - Bivariate Analysis
 - Correlation
- Data Processing & Feature Engineering
- Models Fitting
 - Model Selection: Training and Cross Validation
 - Hyperparameters for the selected models

BLOCKS CLASSIFICATION: DATA PREPROCESSING (1/2)



No NaN
&
No Null or Zero
values



Almost all features
have outliers
→ Treatment to do

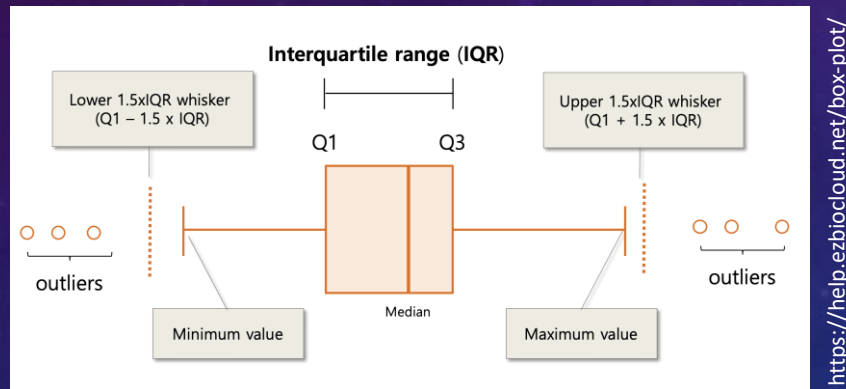


No categorical
features

BLOCKS CLASSIFICATION: DATA PREPROCESSING (2/2)

Outliers

→ Interquartile method:



If the data is an outlier, I checked if it came from the class 1.
If yes, then we delete, else we keep then since there is not so much outliers from the other class

Imbalanced Classes

→ Random sample by fraction:

I decided to keep 1/8 of the class 1, this sample taken randomly.

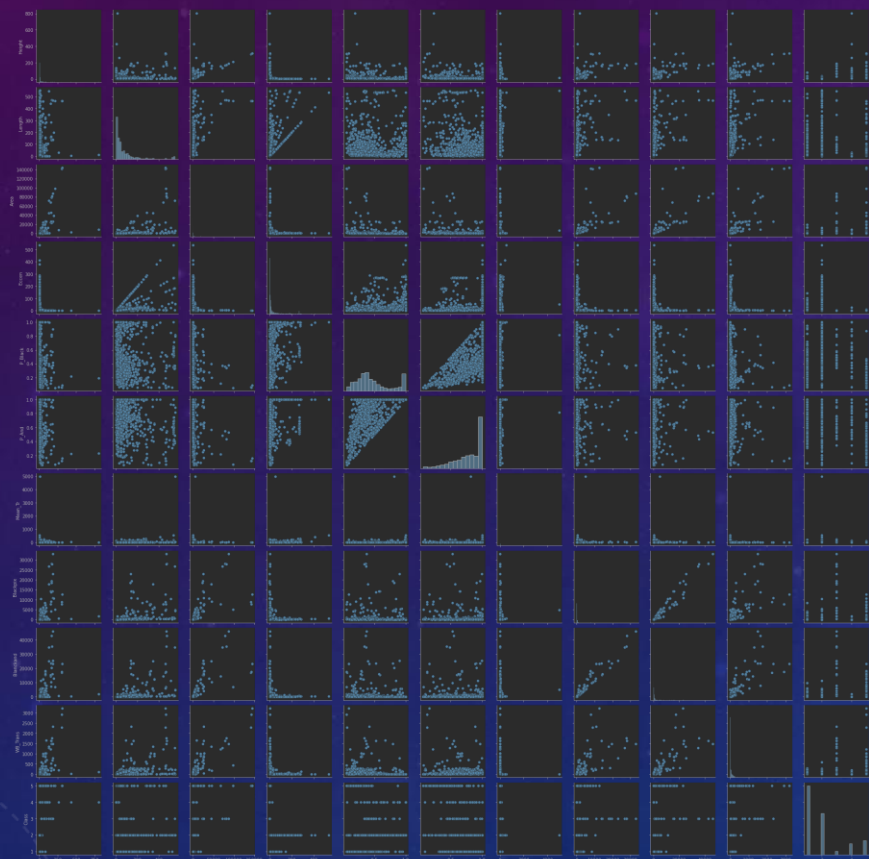
| | |
|---|------|
| 1 | 4394 |
| 2 | 329 |
| 5 | 115 |
| 4 | 88 |
| 3 | 28 |

→

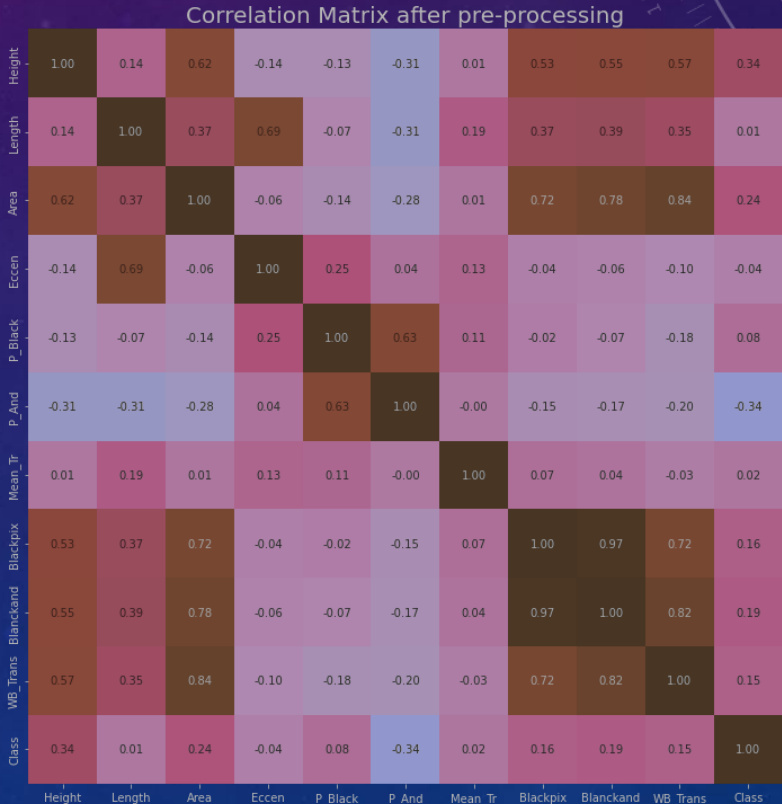
After outliers removing from the class 1, I treated the imbalanced classes problem

| | |
|---|-----|
| 1 | 549 |
| 2 | 329 |
| 5 | 115 |
| 4 | 88 |
| 3 | 28 |

BLOCKS CLASSIFICATION: EDA – SUMMARY



Bivariate Analysis



Correlation Matrix

BLOCKS CLASSIFICATION: MAIN RESULTS - TRAINING

Feature Selection:

→ Height, Area, Blanckand, Blackpix, WB_Trans

Models Tested:

→ Kfold Cross-Validation

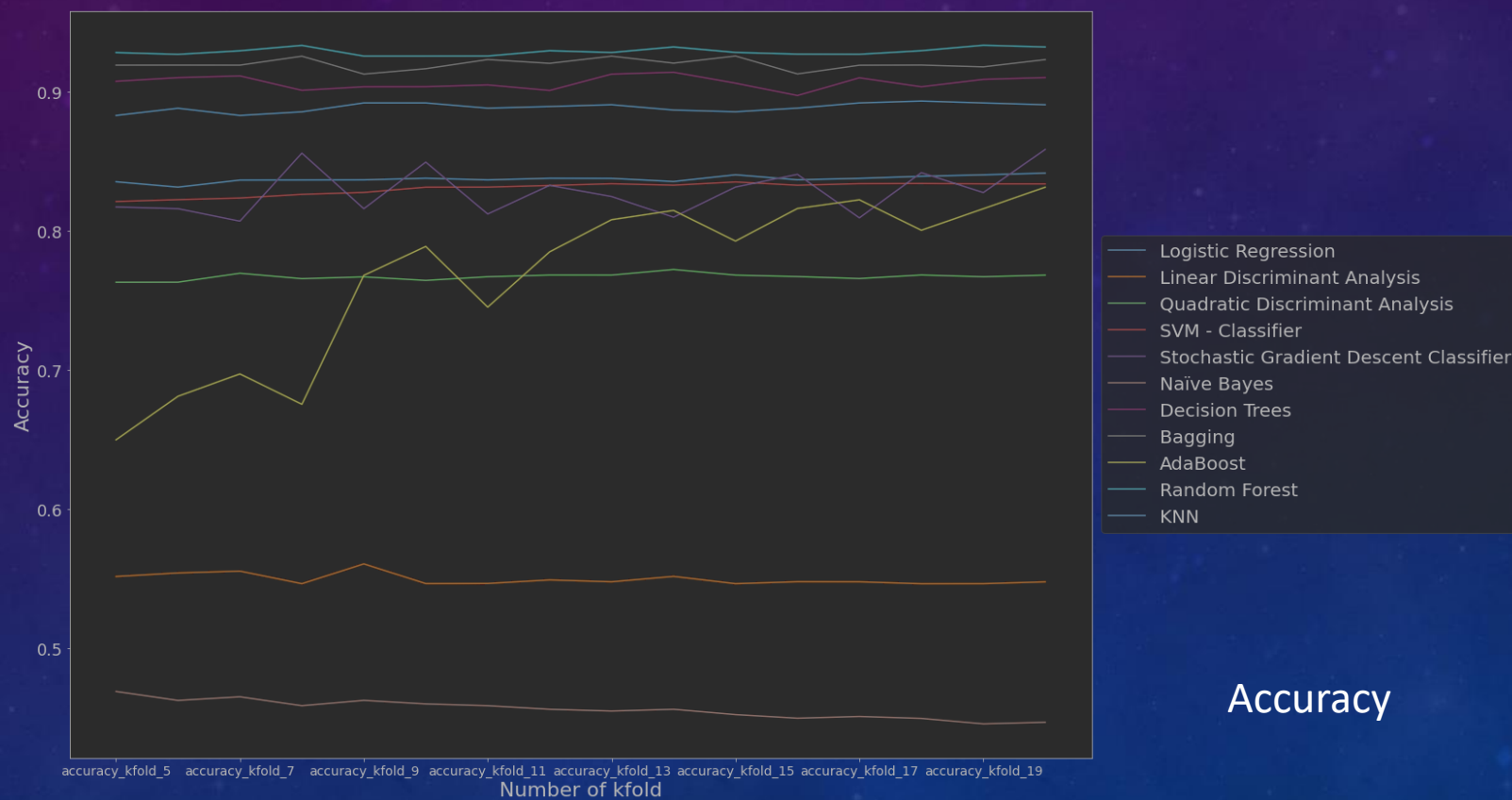
→ Logistic Regression, LDA, QDA, SVC, Stochastic Gradient Descent Classifier, Naïve Bayes, Decision Trees, Bagging, AdaBoost, Random Forest, KNN

Models for hyperparameters tuning:

I took the best model from the step training, the middle one and an other which had bad performances

→ Random Forest, KNN, SVC

SEOUL BIKES: MAIN RESULTS - METRICS



Accuracy

BLOCKS CLASSIFICATION: MAIN RESULTS - HYPERPARAMETERS

Random Forest

- Initial Score: 0.5646
- Hyperparameters:
 - { 'criterion': 'entropy',
 - 'max_depth': 60,
 - 'min_samples_leaf': 1,
 - 'min_samples_split': 2,
 - 'n_estimators': 500 }
- Final Score: 0.5976

KNN

- Initial Score: 0.7958
- Hyperparameters:
 - { 'learning_rate': 0.5,
 - 'n_estimators': 100 }
- Final Score: 0.8048

SVC

- Initial Score: 0.6728
- Hyperparameters { 'C': 500,
'degree': 3, 'gamma': 1,
'kernel': 'linear' }
- Final Score: 0.7898

Problem here when I was searching for the hyperparameters. With only a max_depth of 60, I had the best score which was the same score. So the other parameters do not impact the model (but another one impacts the scores, check the conclusion (2/2))

Design Document Template

Version X.X - 01 December 2016

Company Name - Address - Telephone - Email - www.website.com

© [year] [Company Name] or a [Company Name] affiliate company. All rights reserved.

[Company Name]

[Phone/Fax/Website]

Trademark

[Product Name] are registered trademarks of [Company]. All other trademarks or registered trademarks are the property of their respective owners.

Disclaimer

The information provided in this document is provided "as is" without warranty of any kind. [Company Name] disclaims all warranties, either expressed or implied, including the warranty of merchantability and fitness for a particular purpose. [Company Name] is liable for any damages, including direct, indirect, incidental, consequential, loss of business or other special damages, even if [Company Name] or its suppliers have been advised of the possibility of such damages.

Document History

[Company Name] may occasionally update online documentation to reflect changes in the related software. Consequently, if the document state not documented recently, it may not contain the most up-to-date information. Please refer to [www.\[website\].com](#) for the most current information.

Please refer to this table, providing also document title if not in document if the document is updated, as indicated by a change in the table.

Other legal help

[Company Name] support, product, and training information is obtained as follows.

Product information— Documentation, release notes, software updates, and other information [Company Name] products, training, and services are at [Company Name] website at [http://www.\[website\].com](#)

Technical support— [http://support.\[website\].com](#) and [http://support.\[website\].com](#) support pages, you can get technical support, including trouble-shooting, software support, hardware support, and other support. You must have a valid support agreement.

Your comments

Your suggestions will help us continue to improve the accuracy, organization, and overall quality of the user interface. Please send your comments to the document to:

[DocumentName@Company Name.com](#)

If you have issues, comments, or questions about specific information or content, please include the file name, location, the part number, the version, the page number, and any other details that will help us locate the subject that you are addressing.

[Company Name]

[Product Name]

[Company Name]

[Product Name]

Preface

Style Conventions

The following style conventions are used in this document:

Table

Names of commands, options, programs, processes, services, and utilities

Names of interface elements (such as windows, dialog boxes, menus, fields, and menus)

Interface elements (such as windows, dialog boxes, menus, and menus)

Table

Publication date released in use

Example (for example a new name)

Table

System output, such as an error message or output

URLs, sample paths, filenames, programs, and options

Table

Variables in command line

User input variables

Example: `arg1` indicates an argument or variable value supplied by the user

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

Example: `arg1` indicates an argument or variable value

BLOCKS CLASSIFICATION: DISCUSSION & CONCLUSION (1/2)

The first dataset allowed me to work on a classification problem.

Some aspects were challenging like the outliers management or the imbalanced classes to deal with. Those aspect are big problematic in professional real life project, so it was pretty interesting to discovered some methods even if I didn't have the time to implement them.

The treatments that I apply on my outliers and imbalanced classes have impacted the score of my models. But that allowed to avoid biased models because of the first class which was the most important in this dataset.

Table of Contents

| | | |
|-------|------------------------------------|---|
| 1 | Introduction | 1 |
| 1.1 | Purpose | 1 |
| 1.2 | Background | 1 |
| 1.3 | Scope | 1 |
| 1.4 | Methodology | 1 |
| 1.5 | Evaluation Criteria | 1 |
| 2 | Executive Summary | 2 |
| 2.1 | Purpose of this document | 2 |
| 2.2 | Scope of this document | 2 |
| 2.3 | Methodology, Tools, and Techniques | 2 |
| 2.4 | Evaluation Criteria | 2 |
| 2.5 | Results and Discussion | 2 |
| 2.6 | Conclusion | 2 |
| 3 | Detailed Design | 3 |
| 3.1 | Introduction | 3 |
| 3.2 | System Architecture | 3 |
| 3.3 | Data Design | 3 |
| 3.4 | User Interface Design | 3 |
| 3.5 | Implementation Details | 3 |
| 3.6 | Testing and Validation | 3 |
| 3.7 | Deployment and Maintenance | 3 |
| 3.8 | Appendix | 3 |
| 3.9 | Glossary | 3 |
| 3.10 | Bibliography | 3 |
| 3.11 | Index | 3 |
| 3.12 | Revision History | 3 |
| 3.13 | Change Log | 3 |
| 3.14 | Document History | 3 |
| 3.15 | Document Control | 3 |
| 3.16 | Document Approval | 3 |
| 3.17 | Document Release | 3 |
| 3.18 | Document Archiving | 3 |
| 3.19 | Document Retention | 3 |
| 3.20 | Document Disposal | 3 |
| 3.21 | Document Security | 3 |
| 3.22 | Document Access | 3 |
| 3.23 | Document Modification | 3 |
| 3.24 | Document Distribution | 3 |
| 3.25 | Document Communication | 3 |
| 3.26 | Document Collaboration | 3 |
| 3.27 | Document Integration | 3 |
| 3.28 | Document Interoperability | 3 |
| 3.29 | Document Compatibility | 3 |
| 3.30 | Document Portability | 3 |
| 3.31 | Document Scalability | 3 |
| 3.32 | Document Flexibility | 3 |
| 3.33 | Document Adaptability | 3 |
| 3.34 | Document Resilience | 3 |
| 3.35 | Document Robustness | 3 |
| 3.36 | Document Reliability | 3 |
| 3.37 | Document Availability | 3 |
| 3.38 | Document Usability | 3 |
| 3.39 | Document Learnability | 3 |
| 3.40 | Document Interactability | 3 |
| 3.41 | Document Attractability | 3 |
| 3.42 | Document Desirability | 3 |
| 3.43 | Document Feasibility | 3 |
| 3.44 | Document Viability | 3 |
| 3.45 | Document Desirability | 3 |
| 3.46 | Document Feasibility | 3 |
| 3.47 | Document Viability | 3 |
| 3.48 | Document Desirability | 3 |
| 3.49 | Document Feasibility | 3 |
| 3.50 | Document Viability | 3 |
| 3.51 | Document Desirability | 3 |
| 3.52 | Document Feasibility | 3 |
| 3.53 | Document Viability | 3 |
| 3.54 | Document Desirability | 3 |
| 3.55 | Document Feasibility | 3 |
| 3.56 | Document Viability | 3 |
| 3.57 | Document Desirability | 3 |
| 3.58 | Document Feasibility | 3 |
| 3.59 | Document Viability | 3 |
| 3.60 | Document Desirability | 3 |
| 3.61 | Document Feasibility | 3 |
| 3.62 | Document Viability | 3 |
| 3.63 | Document Desirability | 3 |
| 3.64 | Document Feasibility | 3 |
| 3.65 | Document Viability | 3 |
| 3.66 | Document Desirability | 3 |
| 3.67 | Document Feasibility | 3 |
| 3.68 | Document Viability | 3 |
| 3.69 | Document Desirability | 3 |
| 3.70 | Document Feasibility | 3 |
| 3.71 | Document Viability | 3 |
| 3.72 | Document Desirability | 3 |
| 3.73 | Document Feasibility | 3 |
| 3.74 | Document Viability | 3 |
| 3.75 | Document Desirability | 3 |
| 3.76 | Document Feasibility | 3 |
| 3.77 | Document Viability | 3 |
| 3.78 | Document Desirability | 3 |
| 3.79 | Document Feasibility | 3 |
| 3.80 | Document Viability | 3 |
| 3.81 | Document Desirability | 3 |
| 3.82 | Document Feasibility | 3 |
| 3.83 | Document Viability | 3 |
| 3.84 | Document Desirability | 3 |
| 3.85 | Document Feasibility | 3 |
| 3.86 | Document Viability | 3 |
| 3.87 | Document Desirability | 3 |
| 3.88 | Document Feasibility | 3 |
| 3.89 | Document Viability | 3 |
| 3.90 | Document Desirability | 3 |
| 3.91 | Document Feasibility | 3 |
| 3.92 | Document Viability | 3 |
| 3.93 | Document Desirability | 3 |
| 3.94 | Document Feasibility | 3 |
| 3.95 | Document Viability | 3 |
| 3.96 | Document Desirability | 3 |
| 3.97 | Document Feasibility | 3 |
| 3.98 | Document Viability | 3 |
| 3.99 | Document Desirability | 3 |
| 3.100 | Document Feasibility | 3 |
| 3.101 | Document Viability | 3 |
| 3.102 | Document Desirability | 3 |
| 3.103 | Document Feasibility | 3 |
| 3.104 | Document Viability | 3 |
| 3.105 | Document Desirability | 3 |
| 3.106 | Document Feasibility | 3 |
| 3.107 | Document Viability | 3 |
| 3.108 | Document Desirability | 3 |
| 3.109 | Document Feasibility | 3 |
| 3.110 | Document Viability | 3 |
| 3.111 | Document Desirability | 3 |
| 3.112 | Document Feasibility | 3 |
| 3.113 | Document Viability | 3 |
| 3.114 | Document Desirability | 3 |
| 3.115 | Document Feasibility | 3 |
| 3.116 | Document Viability | 3 |
| 3.117 | Document Desirability | 3 |
| 3.118 | Document Feasibility | 3 |
| 3.119 | Document Viability | 3 |
| 3.120 | Document Desirability | 3 |
| 3.121 | Document Feasibility | 3 |
| 3.122 | Document Viability | 3 |
| 3.123 | Document Desirability | 3 |
| 3.124 | Document Feasibility | 3 |
| 3.125 | Document Viability | 3 |
| 3.126 | Document Desirability | 3 |
| 3.127 | Document Feasibility | 3 |
| 3.128 | Document Viability | 3 |
| 3.129 | Document Desirability | 3 |
| 3.130 | Document Feasibility | 3 |
| 3.131 | Document Viability | 3 |
| 3.132 | Document Desirability | 3 |
| 3.133 | Document Feasibility | 3 |
| 3.134 | Document Viability | 3 |
| 3.135 | Document Desirability | 3 |
| 3.136 | Document Feasibility | 3 |
| 3.137 | Document Viability | 3 |
| 3.138 | Document Desirability | 3 |
| 3.139 | Document Feasibility | 3 |
| 3.140 | Document Viability | 3 |
| 3.141 | Document Desirability | 3 |
| 3.142 | Document Feasibility | 3 |
| 3.143 | Document Viability | 3 |
| 3.144 | Document Desirability | 3 |
| 3.145 | Document Feasibility | 3 |
| 3.146 | Document Viability | 3 |
| 3.147 | Document Desirability | 3 |
| 3.148 | Document Feasibility | 3 |
| 3.149 | Document Viability | 3 |
| 3.150 | Document Desirability | 3 |
| 3.151 | Document Feasibility | 3 |
| 3.152 | Document Viability | 3 |
| 3.153 | Document Desirability | 3 |
| 3.154 | Document Feasibility | 3 |
| 3.155 | Document Viability | 3 |
| 3.156 | Document Desirability | 3 |
| 3.157 | Document Feasibility | 3 |
| 3.158 | Document Viability | 3 |
| 3.159 | Document Desirability | 3 |
| 3.160 | Document Feasibility | 3 |
| 3.161 | Document Viability | 3 |
| 3.162 | Document Desirability | 3 |
| 3.163 | Document Feasibility | 3 |
| 3.164 | Document Viability | 3 |
| 3.165 | Document Desirability | 3 |
| 3.166 | Document Feasibility | 3 |
| 3.167 | Document Viability | 3 |
| 3.168 | Document Desirability | 3 |
| 3.169 | Document Feasibility | 3 |
| 3.170 | Document Viability | 3 |
| 3.171 | Document Desirability | 3 |
| 3.172 | Document Feasibility | 3 |
| 3.173 | Document Viability | 3 |
| 3.174 | Document Desirability | 3 |
| 3.175 | Document Feasibility | 3 |
| 3.176 | Document Viability | 3 |
| 3.177 | Document Desirability | 3 |
| 3.178 | Document Feasibility | 3 |
| 3.179 | Document Viability | 3 |
| 3.180 | Document Desirability | 3 |
| 3.181 | Document Feasibility | 3 |
| 3.182 | Document Viability | 3 |
| 3.183 | Document Desirability | 3 |
| 3.184 | Document Feasibility | 3 |
| 3.185 | Document Viability | 3 |
| 3.186 | Document Desirability | 3 |
| 3.187 | Document Feasibility | 3 |
| 3.188 | Document Viability | 3 |
| 3.189 | Document Desirability | 3 |
| 3.190 | Document Feasibility | 3 |
| 3.191 | Document Viability | 3 |
| 3.192 | Document Desirability | 3 |
| 3.193 | Document Feasibility | 3 |
| 3.194 | Document Viability | 3 |
| 3.195 | Document Desirability | 3 |
| 3.196 | Document Feasibility | 3 |
| 3.197 | Document Viability | 3 |
| 3.198 | Document Desirability | 3 |
| 3.199 | Document Feasibility | 3 |
| 3.200 | Document Viability | 3 |
| 3.201 | Document Desirability | 3 |
| 3.202 | Document Feasibility | 3 |
| 3.203 | Document Viability | 3 |
| 3.204 | Document Desirability | 3 |
| 3.205 | Document Feasibility | 3 |
| 3.206 | Document Viability | 3 |
| 3.207 | Document Desirability | 3 |
| 3.208 | Document Feasibility | 3 |
| 3.209 | Document Viability | 3 |
| 3.210 | Document Desirability | 3 |
| 3.211 | Document Feasibility | 3 |
| 3.212 | Document Viability | 3 |
| 3.213 | Document Desirability | 3 |
| 3.214 | Document Feasibility | 3 |
| 3.215 | Document Viability | 3 |
| 3.216 | Document Desirability | 3 |
| 3.217 | Document Feasibility | 3 |
| 3.218 | Document Viability | 3 |
| 3.219 | Document Desirability | 3 |
| 3.220 | Document Feasibility | 3 |
| 3.221 | Document Viability | 3 |
| 3.222 | Document Desirability | 3 |
| 3.223 | Document Feasibility | 3 |
| 3.224 | Document Viability | 3 |
| 3.225 | Document Desirability | 3 |
| 3.226 | Document Feasibility | 3 |
| 3.227 | Document Viability | 3 |
| 3.228 | Document Desirability | 3 |
| 3.229 | Document Feasibility | 3 |
| 3.230 | Document Viability | 3 |
| 3.231 | Document Desirability | 3 |
| 3.232 | Document Feasibility | 3 |
| 3.233 | Document Viability | 3 |
| 3.234 | Document Desirability | 3 |
| 3.235 | Document Feasibility | 3 |
| 3.236 | Document Viability | 3 |
| 3.237 | Document Desirability | 3 |
| 3.238 | Document Feasibility | 3 |
| 3.239 | Document Viability | 3 |
| 3.240 | Document Desirability | 3 |
| 3.241 | Document Feasibility | 3 |
| 3.242 | Document Viability | 3 |
| 3.243 | Document Desirability | 3 |
| 3.244 | Document Feasibility | 3 |
| 3.245 | Document Viability | 3 |
| 3.246 | Document Desirability | 3 |
| 3.247 | Document Feasibility | 3 |
| 3.248 | Document Viability | 3 |
| 3.249 | Document Desirability | 3 |
| 3.250 | Document Feasibility | 3 |
| 3.251 | Document Viability | 3 |
| 3.252 | Document Desirability | 3 |
| 3.253 | Document Feasibility | 3 |
| 3.254 | Document Viability | 3 |
| 3.255 | Document Desirability | 3 |
| 3.256 | Document Feasibility | 3 |
| 3.257 | Document Viability | 3 |
| 3.258 | Document Desirability | 3 |
| 3.259 | Document Feasibility | 3 |
| 3.260 | Document Viability | 3 |
| 3.261 | Document Desirability | 3 |
| 3.262 | Document Feasibility | 3 |
| 3.263 | Document Viability | 3 |
| 3.264 | Document Desirability | 3 |
| 3.265 | Document Feasibility | 3 |
| 3.266 | Document Viability | 3 |
| 3.267 | Document Desirability | 3 |
| 3.268 | Document Feasibility | 3 |
| 3.269 | Document Viability | 3 |
| 3.270 | Document Desirability | 3 |
| 3.271 | Document Feasibility | 3 |
| 3.272 | Document Viability | 3 |
| 3.273 | Document Desirability | 3 |
| 3.274 | Document Feasibility | 3 |
| 3.275 | Document Viability | 3 |
| 3.276 | Document Desirability | 3 |
| 3.277 | Document Feasibility | 3 |
| 3.278 | Document Viability | 3 |
| 3.279 | Document Desirability | 3 |
| 3.280 | Document Feasibility | 3 |
| 3.281 | Document Viability | 3 |
| 3.282 | Document Desirability | 3 |
| 3.283 | Document Feasibility | 3 |
| 3.284 | Document Viability | 3 |
| 3.285 | Document Desirability | 3 |
| 3.286 | Document Feasibility | 3 |
| 3.287 | Document Viability | 3 |
| 3.288 | Document Desirability | 3 |
| 3.289 | Document Feasibility | 3 |
| 3.290 | Document Viability | 3 |
| 3.291 | Document Desirability | 3 |
| 3.292 | Document Feasibility | 3 |
| 3.293 | Document Viability | 3 |
| 3.294 | Document Desirability | 3 |
| 3.295 | Document Feasibility | 3 |
| 3.296 | Document Viability | 3 |
| 3.297 | Document Desirability | 3 |
| 3.298 | Document Feasibility | 3 |
| 3.299 | Document Viability | 3 |
| 3.300 | Document Desirability | 3 |
| 3.301 | Document Feasibility | 3 |
| 3.302 | Document Viability | 3 |
| 3.303 | Document Desirability | 3 |
| 3.304 | Document Feasibility | 3 |
| 3.305 | Document Viability | 3 |
| 3.306 | Document Desirability | 3 |
| 3.307 | Document Feasibility | 3 |
| 3.308 | Document Viability | 3 |
| 3.309 | Document Desirability | 3 |
| 3.310 | Document Feasibility | 3 |
| 3.311 | Document Viability | 3 |
| 3.312 | Document Desirability | 3 |
| 3.313 | Document Feasibility | 3 |
| 3.314 | Document Viability | 3 |
| 3.315 | Document Desirability | 3 |
| 3.316 | Document Feasibility | 3 |
| 3.317 | Document Viability | 3 |
| 3.318 | Document Desirability | 3 |
| 3.319 | Document Feasibility | 3 |
| 3.320 | Document Viability | 3 |
| 3.321 | Document Desirability | 3 |
| 3.322 | Document Feasibility | 3 |
| 3.323 | Document Viability | 3 |
| 3.324 | Document Desirability | 3 |
| 3.325 | Document Feasibility | 3 |
| 3.326 | Document Viability | 3 |
| 3.327 | Document Desirability | 3 |
| 3.328 | Document Feasibility | 3 |
| 3.329 | Document Viability | 3 |
| 3.330 | Document Desirability | 3 |
| 3.331 | Document Feasibility | 3 |
| 3.332 | Document Viability | 3 |
| 3.333 | Document Desirability | 3 |
| 3.334 | Document Feasibility | 3 |
| 3.335 | Document Viability | 3 |
| 3.336 | Document Desirability | 3 |
| 3.337 | Document Feasibility | 3 |
| 3.338 | Document Viability | 3 |
| 3.339 | Document Desirability | 3 |
| 3.340 | Document Feasibility | 3 |
| 3.341 | Document Viability | 3 |
| 3.342 | Document Desirability | 3 |
| 3.343 | Document Feasibility | 3 |
| 3.344 | Document Viability | 3 |
| 3.345 | Document Desirability | 3 |
| 3.346 | Document Feasibility | 3 |
| 3.347 | Document Viability | 3 |
| 3.348 | Document Desirability | 3 |
| 3.349 | Document Feasibility | 3 |
| 3.350 | Document Viability | 3 |
| 3.351 | Document Desirability | 3 |
| 3.352 | Document Feasibility | 3 |
| 3.353 | Document Viability | 3 |
| 3.354 | Document Desirability | 3 |
| 3.355 | Document Feasibility | 3 |
| 3.356 | Document Viability | 3 |
| 3.357 | Document Desirability | 3 |
| 3.358 | Document Feasibility | 3 |
| 3.359 | Document Viability | 3 |
| 3.360 | Document Desirability | 3 |
| 3.361 | Document Feasibility | 3 |
| 3.362 | Document Viability | 3 |
| 3.363 | Document Desirability | 3 |
| 3.364 | Document Feasibility | 3 |
| 3.365 | Document Viability | 3 |
| 3.366 | Document Desirability | 3 |
| 3.367 | Document Feasibility | 3 |
| 3.368 | Document Viability | 3 |
| 3.369 | Document Desirability | 3 |
| 3.370 | Document Feasibility | 3 |
| 3.371 | Document Viability | 3 |
| 3.372 | Document Desirability | 3 |
| 3.373 | Document Feasibility | 3 |
| 3.374 | Document Viability | 3 |
| 3.375 | Document Desirability | 3 |
| 3.376 | Document Feasibility | 3 |
| 3.377 | Document Viability | 3 |
| 3.378 | Document Desirability | 3 |
| 3.379 | Document Feasibility | 3 |
| 3.380 | Document Viability | 3 |
| 3.381 | Document Desirability | 3 |
| 3.382 | Document Feasibility | 3 |
| 3.383 | Document Viability | 3 |
| 3.384 | Document Desirability | 3 |
| 3.385 | Document Feasibility | 3 |
| 3.386 | Document Viability | 3 |
| 3.387 | Document Desirability | 3 |
| 3.388 | Document Feasibility | 3 |
| 3.389 | Document Viability | 3 |
| 3.390 | Document Desirability | 3 |
| 3.391 | Document Feasibility | 3 |
| 3.392 | Document Viability | 3 |
| 3.393 | Document Desirability | 3 |
| 3.394 | Document Feasibility | 3 |
| 3.395 | Document Viability | 3 |
| 3.396 | Document Desirability | 3 |
| 3.397 | Document Feasibility | 3 |
| 3.398 | Document Viability | 3 |
| 3.399 | Document Desirability | 3 |
| 3.400 | Document Feasibility | 3 |
| 3.401 | Document Viability | 3 |
| 3.402 | Document Desirability | 3 |
| 3.403 | Document Feasibility | 3 |
| 3.404 | Document Viability | 3 |
| 3.405 | Document Desirability | 3 |
| 3.406 | Document Feasibility | 3 |
| 3.407 | Document Viability | 3 |
| 3.408 | Document Desirability | 3 |
| 3.409 | Document Feasibility | 3 |
| 3.410 | Document Viability | 3 |
| 3.411 | Document Desirability | 3 |
| 3.412 | Document Feasibility | 3 |
| 3.413 | Document Viability | 3 |
| 3.414 | Document Desirability | 3 |
| 3.415 | Document Feasibility | 3 |
| 3.416 | Document Viability | 3 |
| 3.417 | Document Desirability | 3 |
| 3.418 | Document Feasibility | 3 |
| 3.419 | Document Viability | 3 |
| 3.420 | Document Desirability | 3 |
| 3.421 | Document Feasibility | 3 |
| 3.422 | Document Viability | 3 |
| 3.423 | Document Desirability | 3 |
| 3.424 | Document Feasibility | 3 |
| 3.425 | Document Viability | 3 |
| 3.4 | | |

Design Document Template

Version X.X - 01 December 2016

Company Name • Address • Telephone • Email • www.website.com

© [year] [Company Name] or a [Company Name] affiliate company. All rights reserved.

[Company]([address])

[Phone]([phone])

Table of Contents

[Product Name] are registered trademarks of [Company]. All other trademarks or registered trademarks are the property of their respective owners.

Disclaimer

The information provided in this document is provided "as is" without warranty of any kind. [Company Name] disclaims all warranties, either expressed or implied, including the accuracy of completeness and timeliness of the information. [Company Name] is liable for any damages arising from the use of the information, including direct, indirect, incidental, consequential, loss of business or profits or special damages, even if [Company Name] or its suppliers have been advised of the possibility of such damages.

Document History

[Company Name] may occasionally update online documentation to reflect changes to the product software. Consequently, if the document state is discontinued monthly, it may not contain the most up-to-date information. Please refer to [www.\[company\].com](#) for the most current information.

For more information, please refer to the document state is discontinued if the document is updated, as indicated by a change in the document state.

Other legal help

[Company Name] support, product, and licensing information is contained in the document.

Product Information: Documentation, release notes, software updates, and other information about [Company Name] products, services, and updates are at [Company Name] website at [www.\[company\].com](#).

Technical Support: [Company Name] provides technical support for our products. For more information, please refer to the document state is discontinued if the document is updated, as indicated by a change in the document state.

Other legal help

For more information, please refer to the document state is discontinued if the document is updated, as indicated by a change in the document state.

For more information, please refer to the document state is discontinued if the document is updated, as indicated by a change in the document state.

For more information, please refer to the document state is discontinued if the document is updated, as indicated by a change in the document state.

[Company Name]

[Product Name]

[Company Name]

[Product Name]

Preface

Style Conventions

The following style conventions are used in this document:

Table

Names of commands, options, programs, processes, services, and utilities

Names of interface elements (such as windows, dialog boxes, menus, fields, and menus)

Interface elements (such as windows, dialog boxes, menus, fields, and menus)

Table

Publication date information (in use)

Examples (for example, a new name)

Table

System output, such as an error message or output

URLs, sample paths, filenames, prompts, and options

Table

Variables in command lines

User input variables

For example, the following command uses variables supplied by the user:

1) Example command line

2) Example command line

3) Example command line

4) Example command line

5) Example command line

6) Example command line

7) Example command line

8) Example command line

9) Example command line

10) Example command line

11) Example command line

12) Example command line

13) Example command line

14) Example command line

15) Example command line

16) Example command line

17) Example command line

18) Example command line

19) Example command line

20) Example command line

21) Example command line

22) Example command line

23) Example command line

24) Example command line

25) Example command line

26) Example command line

27) Example command line

28) Example command line

29) Example command line

30) Example command line

31) Example command line

32) Example command line

33) Example command line

34) Example command line

35) Example command line

36) Example command line

37) Example command line

38) Example command line

39) Example command line

40) Example command line

41) Example command line

42) Example command line

43) Example command line

44) Example command line

45) Example command line

BLOCKS CLASSIFICATION: DISCUSSION & CONCLUSION (2/2)

Especially if we take a look at the imbalanced classes management, I took randomly samples (1/8 of the first class) to redo the first class and to have a better proportion between each class. Thus, when I run many time the notebook, I obtained different scores more or less better, depending on the sample that the algorithm took (since a part of the outliers was treated).

So, because of the fact that I didn't succeed to do something that I can named as a "good job" with this dataset, I decided to take an other dataset to play with.

[Company Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

[Product Name]

1.3 Evaluation Criteria

Define the criteria used to evaluate software systems, such as organizational objectives, increased efficiency, and reduced operating costs.

2 Executive Summary

Provide a brief introduction to the system for which this design is being developed.

2.1 Purpose of this document

[Company Name]

[Product Name]

[Product Name]

1) Configuration Management Plan [version]

2) Software Quality Assurance Plan [version]

2.5 Methodology, Tools, and Techniques

Describe the software tools (or techniques) required for performing design document tasks.

SEOUL BIKE SHARING DEMAND DATASET

Link: <https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand>

SEOUL BIKES: PRESENTATION

According to the responsible of this dataset, the dataset contains count of public bikes rented at each hour in Seoul Bike haring System with the corresponding Weather data and Holidays information

"Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes. The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information."

SEOUL BIKES: PROBLEM

Supervised → Regression Problem

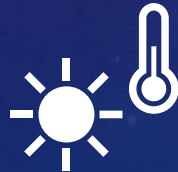


15 Features

Time



Weather




Prediction of bike count required at each hour
for the stable supply of rental bikes:

Target

Rented Bikes count



SEOUL BIKES: STEPS FOLLOWED FOR THIS WORK



- Data Collection
- First Exploratory Data Analysis
- Data Preprocessing
 - Missing Values
 - Null & Zero Values
 - Outliers
 - Categorical Features
- Second Exploratory Data Analysis
 - Univariate Analysis
 - Time
 - Weather
 - Bivariate Analysis
 - Correlation
- Data Processing & Feature Engineering
- Models Fitting
 - Model Selection: Training and Cross Validation
 - Hyperparameters for the selected models

SEOUL BIKES: DATA PREPROCESSING - SUMMARY



No NaN
&
almost no Null or
Zero values



Some features have
outliers but caused by
the random variation
of the weather



Three categorical
features:
- Seasons
- Holiday
- Functioning Day

→ No manipulation compared to the previous dataset

SEOUL BIKES: EDA – SUMMARY



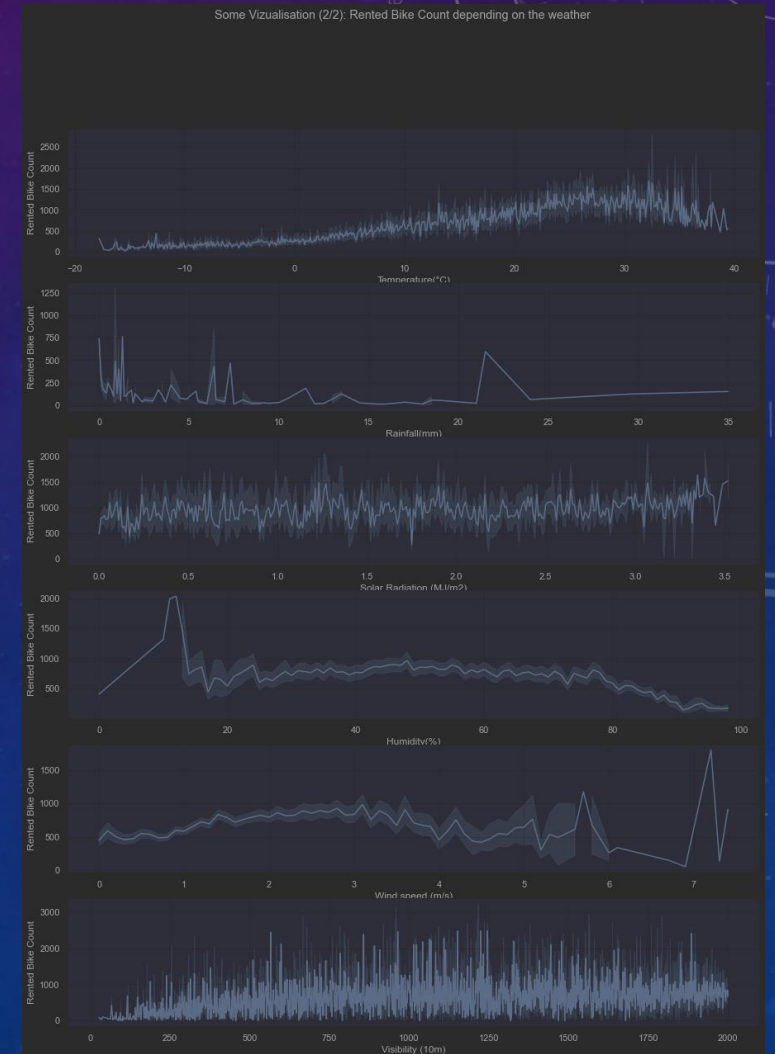
Univariate Analysis

Time
Month/Season and Hour
have an impact on the target

Weather
Temperature, visibility
and solar radiation have an
impact on the target for
the weather features

Bivariate Analysis

→ Dew point temperature has a
big correlation with the
temperature and the humidity



SEOUL BIKES: MAIN RESULTS - TRAINING

Feature Selection:

→ Temperature, Hour, Seasons, Solar Radiation, Visibility, Month, Wind speed, Business Day

Models Tested:

→ Kfold Cross-Validation

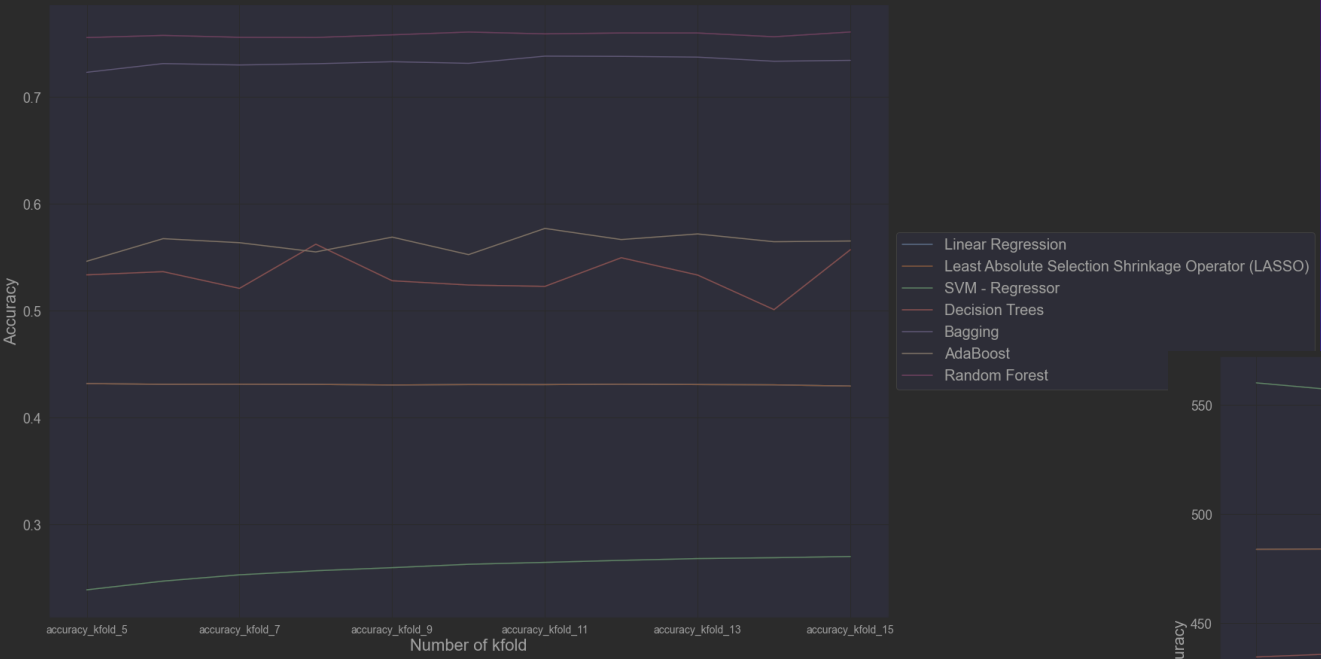
→ Linear Regression, Lasso, SVR, Decision Tree, Bagging, AdaBoost, Random Forest

Models for hyperparameters tuning:

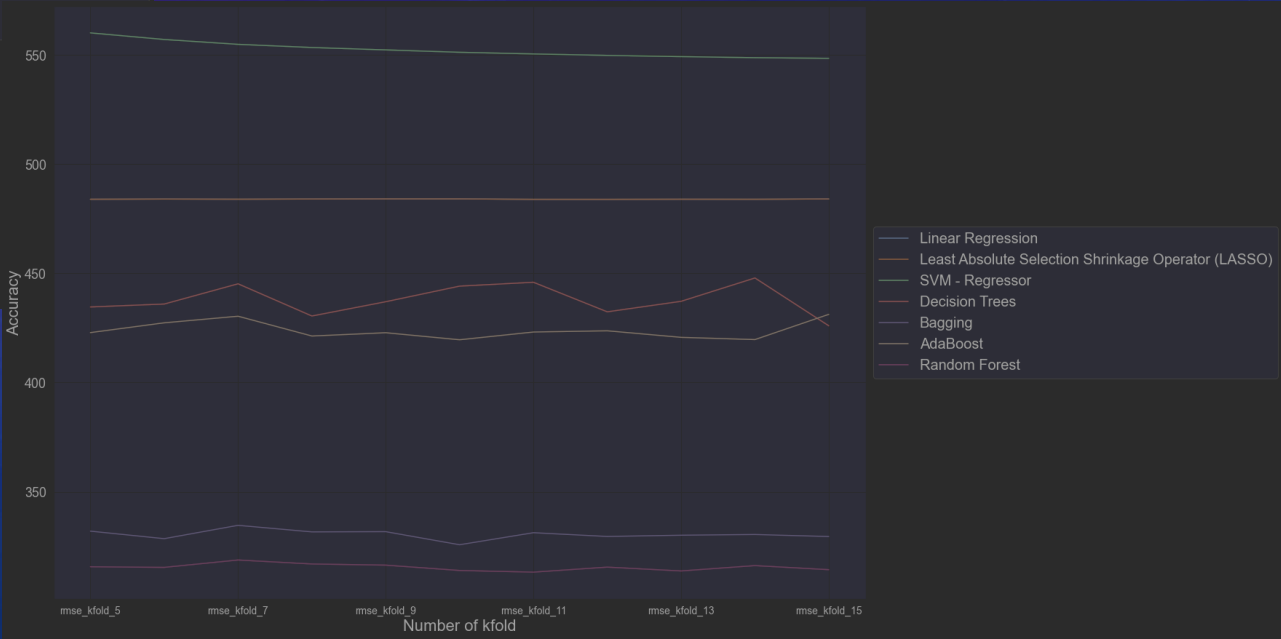
I took the best model from the step training, the middle one and an other which had bad performances

→ Random Forest, AdaBoost, Lasso

SEOUL BIKES: MAIN RESULTS - METRICS



Accuracy



RMSE

SEOUL BIKES: MAIN RESULTS - HYPERPARAMETERS

Random Forest

- Initial Score: 0.7960
- Hyperparameters: `{'criterion': 'squared_error', 'max_depth': 100, 'n_estimators': 1000}`
- Final Score: 0.7961

AdaBoost

- Initial Score: 0.5976
- Hyperparameters: `{'learning_rate': 0.5, 'n_estimators': 100}`
- Final Score: 0.6210

Lasso

- Initial Score: 0.4506
- Hyperparameters: `{'alpha': 0.1}`
- Final Score: 0.4508



SEOUL BIKES: DISCUSSION & CONCLUSION

The second dataset was a regression problem.

I had the opportunity to work on another models we haven't seen in class. One of the advantages of this dataset was that it was pretty easy to do some Exploratory Data Analysis (EDA) compared to the other dataset, as much as to do some univariate analysis.

FLASK API IMPLEMENTED

An API has been implemented.

In fact, just the beginning of the API has been pushed. Currently, it is possible to select the dataset either the blocks classification one or the Seoul bike one. And then, the user can add a value for each significant feature find from the work which has been done on the notebooks. When all the forms are filled, we can push the button “Predict!”.

Normally, when this button is pushed, a prediction of either the class of the block or the number of rented bike, is made and appeared. This major fonctionnality has not been implemented yet (lack of time to do it properly).

Moreover, the design is too simple, it can be more worked on.

Which Dataset?

Select a dataset

blocks_classification

blocks_classification

seoul_bike

Features Input:

Which Dimension for the Grid

Temperature

in °C

Hour

from 0 to 23

Seasons

W, SP, SU or A

Solar Radiation

...

Visibility

...

Month

From 1 to 12

Wind Speed

...

Business Day?

0: False or 1: True

Predict!

CONCLUSION: DISCUSSION ON THE PROJECT

For this project I wanted to use all we have seen during the practical works of this course.

I was able to work both on classification and regression problems, thus this project was complete with different aspects that we can run into from a supervised learning problem.

This project also pushes me to interpret each analysis I have made. It is a good thing to now how to solve problem in a practical manner but the interpretations are a big part of the job too. Even if we hadn't seen interpretations that much during the course, I used my knowledge from another course I had this semester.

CONCLUSION: PERSPECTIVES

- For the data preprocessing for each dataset, it is possible to do more than it was made for this project. First, the work on the outliers management could have been better. There are plenty of methods to treat them, not only the interquartile rule, and we can also treat them and not delete all of them (like using the median and so on). For the imbalanced classes, there are also some method that are better than taking randomly a define fraction of a class which is the more represented.
- More Data Visualizations could have been made. For this part, R is maybe more adequate but Python did the job with seaborn. I still have a lot to learn on this package.
- For the feature selection, I used only the correlation matrix. It is also possible to use the Principle Component Analysis (PCA). It is a good way to get the significant component and to support the interpretation of the previous method.
- Finally, the hyperparameters for each model can be improve considerably. I tried to start some hyperparameters running with large choice for different parameters of each model but I couldn't get anything in a good amount of time.

All the points describe above will be implemented after the evaluation of this project and the grade given.