COMP 432 Final Project

AmirHossein Hadinezhad, Axel Solano, Annika Timermanis

Dec 2021

Table of Contents

Introduction

This is Project of Machine Learning course, during which we had to develop a variety of models for different detests

Our Goals

- Organization
- Weekly live meetings
- Structure
- Cohesion
- Gain practical experience

Challenges

- We realize we would have less people on our team as one of our teammates dropped the course
- lack of time
- Project and courses workload

Tools

Version Control:

- GitHub

IDEs:

- Jupyter

Libraries:

- Variety of libraries including matplotlib, numpy, pandas ..

Training Overview

Our data investigation example

Data preparation and Analysis

```
print(np.count_nonzero(dataset=="?"))
print(np.count_nonzero(dataset==np.nan))
dataset.isnull().sum()
```

```
16
0
```

```
0     0
1     0
2     0
3     0
4     0
5     0
6     0
7     0
8     0
9     0
10    0
dtype: int64
```

```
dataset.isin(["?"]).sum()
```

```
0      0
1      0
2      0
3      0
4      0
5      0
6     16
7      0
8      0
9      0
10     0
dtype: int64
```

```
dataset[dataset[6]=='?']
```

	0	1	2	3	4	5	6	7	8	9	10
23	1057013	8	4	5	1	2	?	7	3	1	4
40	1096800	6	6	6	9	6	?	7	8	1	2
139	1183246	1	1	1	1	1	?	2	1	1	2
145	1184840	1	1	3	1	2	?	2	1	1	2
158	1193683	1	1	2	1	3	?	1	1	1	2

Important Feature of Breast Cancer Wisconsin

Loading Data

```
Dataset = pd.read_csv('Regression_Datasets/5_Bike_Sharing/Bike_Sharing.csv', delimiter=',')
Dataset.head()
```

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0	3	13	16
1	2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0	8	32	40
2	3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0	5	27	32
3	4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0	3	10	13
4	5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0	0	1	1

Data Prepration And Analysis

```
data= Dataset.iloc[:,2:].drop(['casual','registered'],axis =1)
data.head()
```

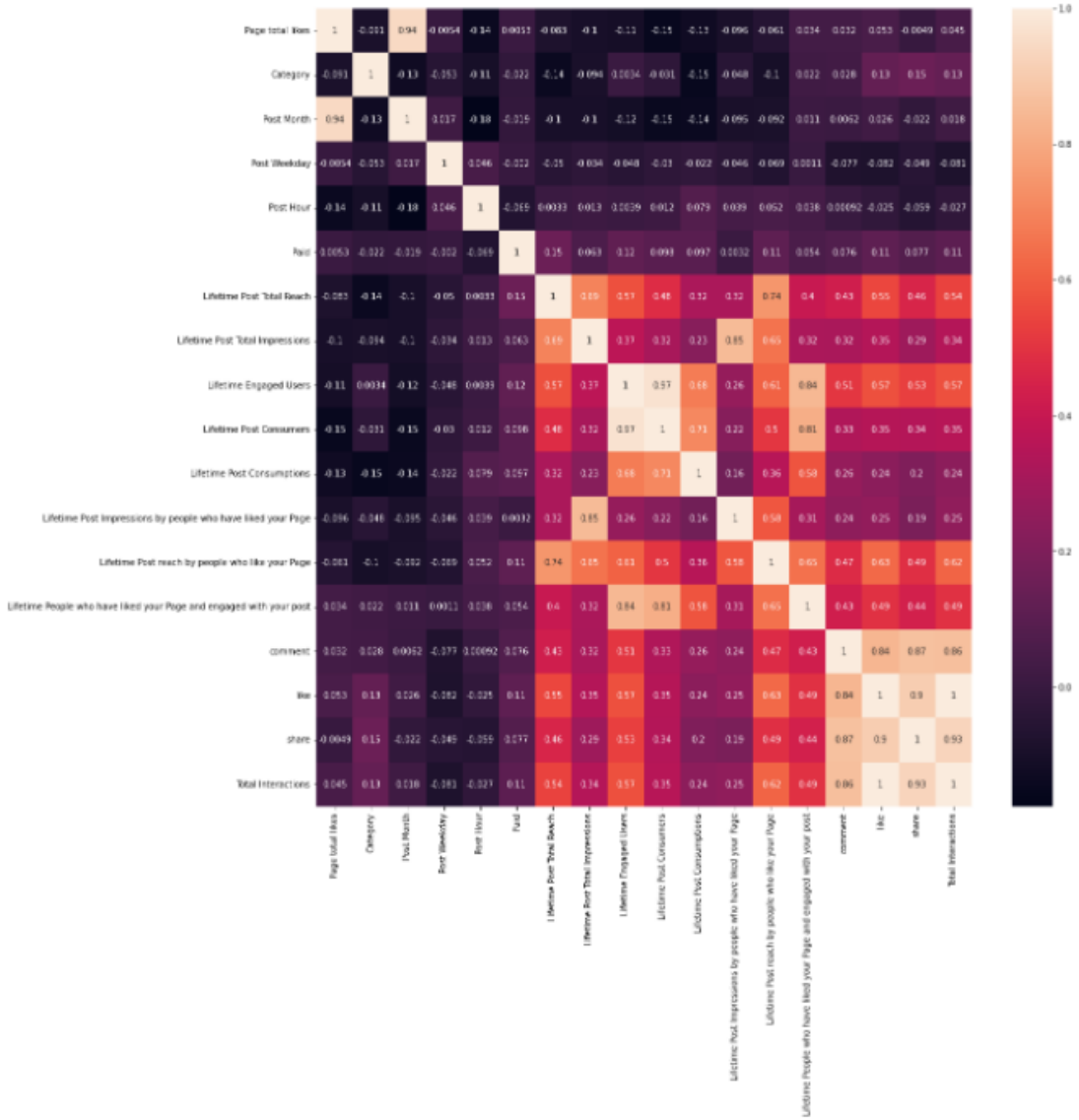
	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	cnt
0	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0	16
1	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0	40
2	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0	32
3	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0	13
4	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0	1

```
data.shape
```

```
(17379, 13)
```

```
data.isnull().sum()
```

Our data investigation example

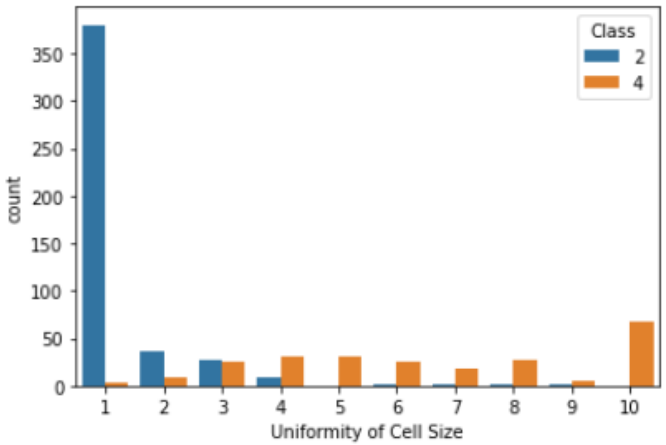


Important Feature of Breast Cancer Wisconsin

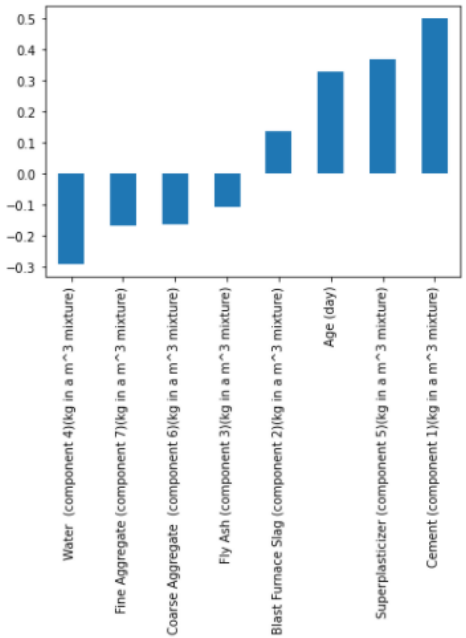
	precision	recall	f1-score	support
2	0.91	0.94	0.93	66
4	0.89	0.85	0.87	39
accuracy			0.90	105
macro avg	0.90	0.89	0.90	105
weighted avg	0.90	0.90	0.90	105

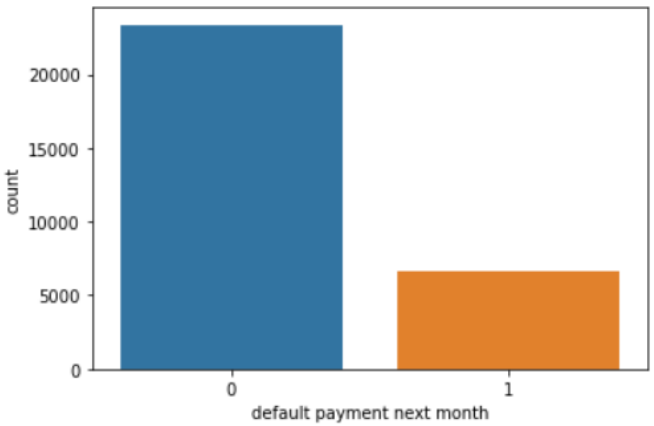
Uniformity of Cell Size

50]: <AxesSubplot:xlabel='Uniformity of Cell Size', ylabel='count'>

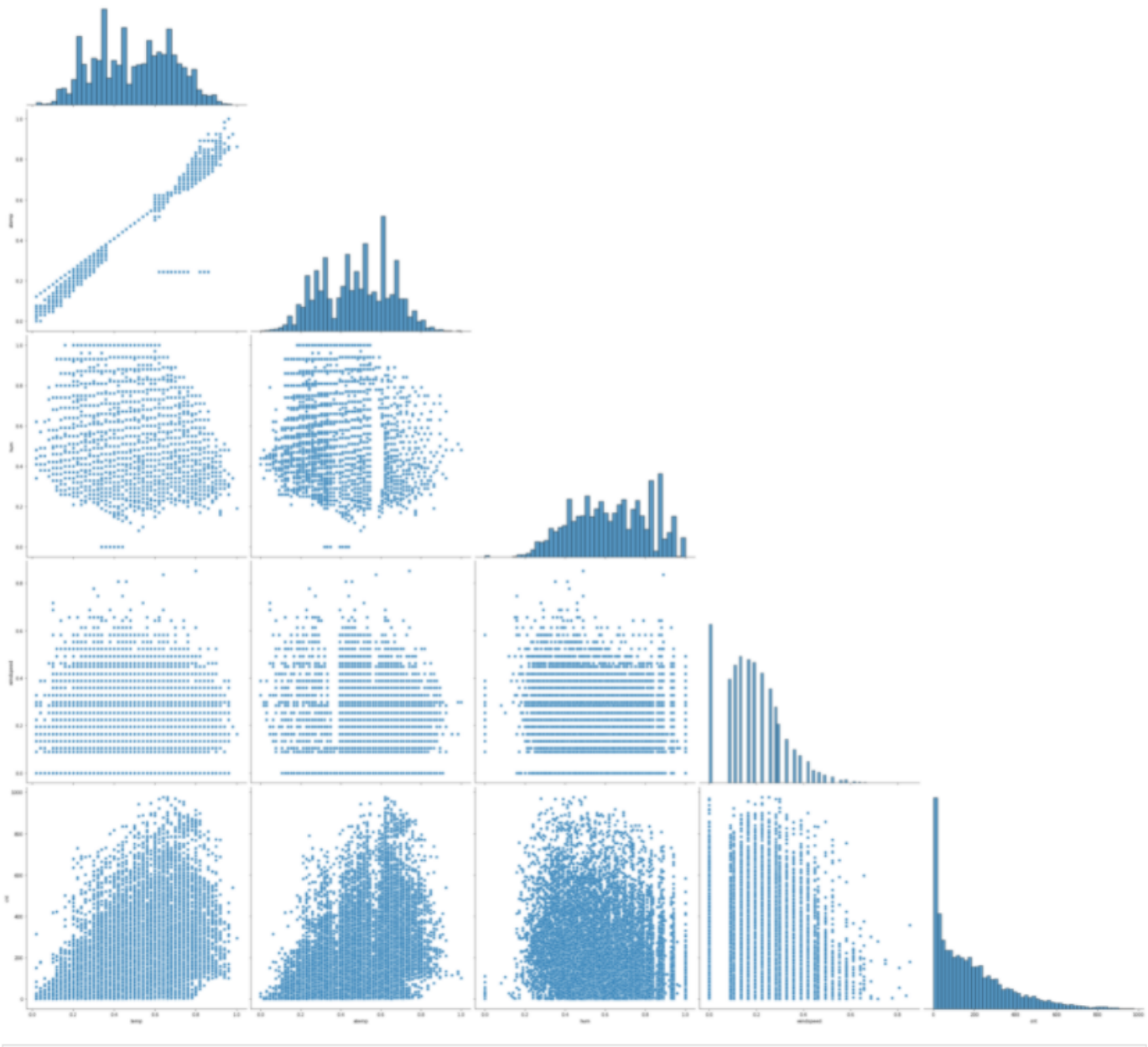


Target distribution and its relation to features

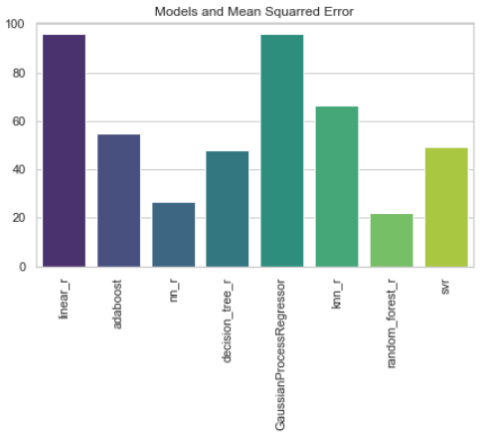


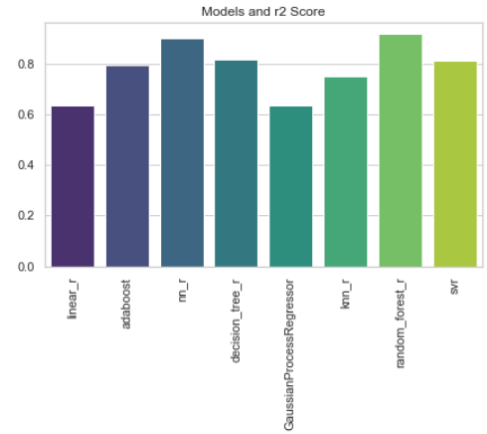


Similar Features



Result of Regression and Classification

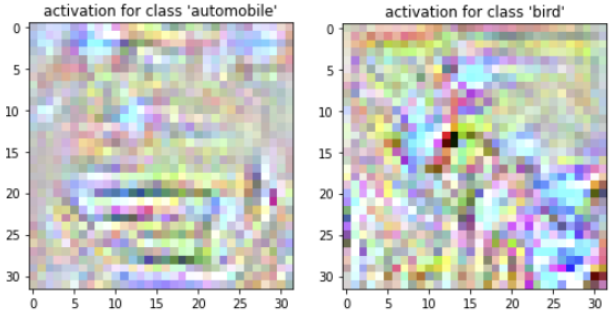


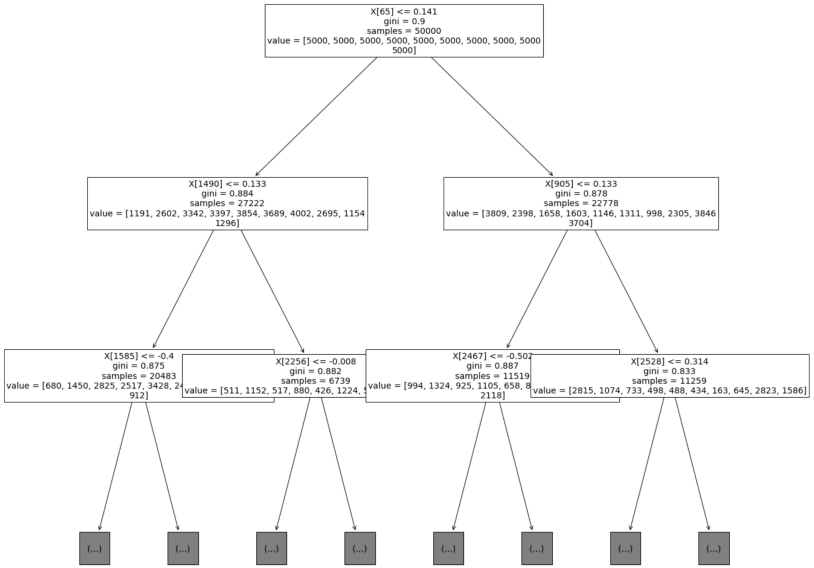


model	acc	w.f1	w.rec	clas.rec	tr.acc	tr.w.f1	tr.w.rec	tr.clas.rec
tree	0.842857	0.841698	0.842857	0.763636	0.953488	0.953293	0.953488	0.913978
knn	0.921429	0.921766	0.921429	0.927273	1	1	1	1
forest	0.842857	0.841698	0.842857	0.763636	0.953488	0.953293	0.953488	0.913978
adaboost	0.942857	0.94266	0.942857	0.909091	0.991055	0.991037	0.991055	0.978495
gnb	0.871429	0.872917	0.871429	0.963636	0.887299	0.88919	0.887299	0.913978
nn	0.921429	0.921766	0.921429	0.927273	0.951699	0.951857	0.951699	0.94086
logistic_r	0.892857	0.894021	0.892857	0.963636	0.874776	0.876718	0.874776	0.887097
svc	0.9	0.900757	0.9	0.927273	0.996422	0.996427	0.996422	1

Overall Conclusions: 3. CIFAR: Decision Tree VS CNN

- Most interpretable
  - The CNN was more interpretable than the Decision Tree
- Reasons::
  - CNN activation maximization figures
  - Decision Tree plot





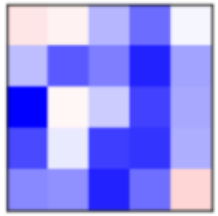
Novelty Component

The novelty component chosen: Tried inspecting the CNN's filters



CNN First Layer Filters





CNN Final Layer filters

Outcome

What We Succeeded VS What We Could Have Improved

- Succeeded
  - Trained and tuned every model
  - Did an analysis, and cross comparison of models
- Improved:
  - Timing
  - Organization towards the end of the semester
  - Class weights
  - Dummy estimator