3 Must-have tools if you're serious about machine learning

Pandas profiling

Pandas is one of the most widely used libraries for loading and processing data in Python. It has a gre statistical operations on your data.

One of these methods is the describe method, which gives you a compact summary of your data on t notebook.

This method is very basic, maybe a little too basic for anyone that's serious about machine learning.

There is an alternative, called **Pandas profiling**.

This library generates a complete report for your dataset, which includes:

- Basic data type information (which columns contain what)
- Descriptive statistics (mean, average, etc.)
- Quantile statistics (tells you about how your data is distributed)
- Histograms for your data (again, for visualizing distributions)
- Correlations (Let's you see what features are related)

Using this library is simple:

```
pip install -U pandas-profiling
```

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

```
Collecting pandas-profiling
  Downloading https://files.pythonhosted.org/packages/e8/b0/bd5e3aaf37302fbe581b6947dc5e
     |████████████████████████████████| 194kB 4.8MB/s
Requirement already satisfied, skipping upgrade: numpy>=1.16.0 in /usr/local/lib/python3
Requirement already satisfied, skipping upgrade: scipy>=1.4.1 in /usr/local/lib/python3.
Requirement already satisfied, skipping upgrade: pandas==0.25.3 in /usr/local/lib/python
Requirement already satisfied, skipping upgrade: matplotlib>=3.0.3 in /usr/local/lib/pyt
Collecting confuse==1.0.0
  Downloading https://files.pythonhosted.org/packages/4c/6f/90e860cba937c174d8b3775729cc
Requirement already satisfied, skipping upgrade: jinja2==2.11.1 in /usr/local/lib/python
Collecting visions==0.2.2
  Downloading https://files.pythonhosted.org/packages/07/4a/ab37f8bafda516b66c4f475b221a
Collecting htmlmin==0.1.12
  Downloading https://files.pythonhosted.org/packages/b3/e7/fcd59e12169de19f0131ff281207
Requirement already satisfied, skipping upgrade: missingno==0.4.2 in /usr/local/lib/pyth
Collecting phik==0.9.9
  Downloading https://files.pythonhosted.org/packages/03/cf/b8cef2778104dc5d319f36dd836e
     |████████████████████████████████| 614kB 14.4MB/s
Requirement already satisfied, skipping upgrade: astropy>=3.2.3 in /usr/local/lib/python
Collecting tangled-up-in-unicode==0.0.3
  Downloading https://files.pythonhosted.org/packages/43/fc/e3c970c5007b405827a4623e70fc
     |████████████████████████████████| 1.5MB 28.7MB/s
Collecting tqdm==4.42.0
  Downloading https://files.pythonhosted.org/packages/cc/2e/4307206db63f05ed37e21d4c0d84
     |████████████████████████████████| 61kB 10.6MB/s
Requirement already satisfied, skipping upgrade: kaggle==1.5.6 in /usr/local/lib/python3
Requirement already satisfied, skipping upgrade: ipywidgets==7.5.1 in /usr/local/lib/pyt
Collecting requests==2.22.0
  Downloading https://files.pythonhosted.org/packages/51/bd/23c926cd341ea6b7dd0b2a00aba9
     |████████████████████████████████| 61kB 8.6MB/s
Requirement already satisfied, skipping upgrade: python-dateutil>=2.6.1 in /usr/local/li
Requirement already satisfied, skipping upgrade: pytz>=2017.2 in /usr/local/lib/python3.
Requirement already satisfied, skipping upgrade: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.
Requirement already satisfied, skipping upgrade: kiwisolver>=1.0.1 in /usr/local/lib/pyt
Requirement already satisfied, skipping upgrade: cycler>=0.10 in /usr/local/lib/python3.
```

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

```
                                                                                       st
                                                                                       th
Requirement already satisfied, skipping upgrade: networkx in /usr/local/lib/python3.6/di
Collecting attr
  Downloading https://files.pythonhosted.org/packages/de/be/ddc7f84d4e087144472a38a373d3
Requirement already satisfied, skipping upgrade: seaborn in /usr/local/lib/python3.6/dis
Collecting pytest-pylint>=0.13.0
  Downloading https://files.pythonhosted.org/packages/0f/2e/12d3e83e90341fc7aff0e42955ab
Requirement already satisfied, skipping upgrade: nbconvert>=5.3.1 in /usr/local/lib/pyth
Requirement already satisfied, skipping upgrade: jupyter-client>=5.2.3 in /usr/local/lib
Requirement already satisfied, skipping upgrade: joblib>=0.14.1 in /usr/local/lib/python
Collecting pytest>=4.0.2
  Downloading https://files.pythonhosted.org/packages/a5/c0/34033b2df7718b91c667bd259d5c
     |████████████████████████████████| 235kB 43.0MB/s
Requirement already satisfied, skipping upgrade: numba>=0.38.1 in /usr/local/lib/python3
Requirement already satisfied, skipping upgrade: certifi in /usr/local/lib/python3.6/dis
Requirement already satisfied, skipping upgrade: six>=1.10 in /usr/local/lib/python3.6/d
Requirement already satisfied, skipping upgrade: urllib3<1.25,>=1.21.1 in /usr/local/lib
Requirement already satisfied, skipping upgrade: python-slugify in /usr/local/lib/python
Requirement already satisfied, skipping upgrade: widgetsnbextension~=3.5.0 in /usr/local
Requirement already satisfied, skipping upgrade: nbformat>=4.2.0 in /usr/local/lib/pytho
Requirement already satisfied, skipping upgrade: ipython>=4.0.0; python_version >= "3.3"
Requirement already satisfied, skipping upgrade: traitlets>=4.3.1 in /usr/local/lib/pyth
Requirement already satisfied, skipping upgrade: ipykernel>=4.5.1 in /usr/local/lib/pyth
Requirement already satisfied, skipping upgrade: idna<2.9,>=2.5 in /usr/local/lib/python
Requirement already satisfied, skipping upgrade: chardet<3.1.0,>=3.0.2 in /usr/local/lib
Requirement already satisfied, skipping upgrade: setuptools in /usr/local/lib/python3.6/
Requirement already satisfied, skipping upgrade: decorator>=4.3.0 in /usr/local/lib/pyth
Collecting pylint>=2.0.0
  Downloading https://files.pythonhosted.org/packages/e9/59/43fc36c5ee316bb9aeb7cf5329cd
    |████████████████████████████████| 307kB 37.2MB/s
Requirement already satisfied, skipping upgrade: jupyter-core in /usr/local/lib/python3.
Requirement already satisfied, skipping upgrade: mistune<2,>=0.8.1 in /usr/local/lib/pyt
Requirement already satisfied, skipping upgrade: pandocfilters>=1.4.1 in /usr/local/lib/
Requirement already satisfied, skipping upgrade: bleach in /usr/local/lib/python3.6/dist
Requirement already satisfied, skipping upgrade: testpath in /usr/local/lib/python3.6/di
Requirement already satisfied, skipping upgrade: defusedxml in /usr/local/lib/python3.6/
Requirement already satisfied, skipping upgrade: pygments in /usr/local/lib/python3.6/di
Requirement already satisfied, skipping upgrade: entrypoints>=0.2.2 in /usr/local/lib/py
Requirement already satisfied, skipping upgrade: pyzmq>=13 in /usr/local/lib/python3.6/d
Requirement already satisfied, skipping upgrade: tornado>=4.1 in /usr/local/lib/python3.
Requirement already satisfied, skipping upgrade: packaging in /usr/local/lib/python3.6/d
Requirement already satisfied, skipping upgrade: importlib-metadata>=0.12; python_versio
Requirement already satisfied, skipping upgrade: py>=1.5.0 in /usr/local/lib/python3.6/d
Requirement already satisfied, skipping upgrade: more-itertools>=4.0.0 in /usr/local/lib
Requirement already satisfied, skipping upgrade: wcwidth in /usr/local/lib/python3.6/dis
Collecting pluggy<1.0,>=0.12
  Downloading https://files.pythonhosted.org/packages/a0/28/85c7aa31b80d150b772fbe4a2294
Requirement already satisfied, skipping upgrade: attrs>=17.4.0 in /usr/local/lib/python3
Requirement already satisfied, skipping upgrade: llvmlite>=0.31.0dev0 in /usr/local/lib/
Requirement already satisfied, skipping upgrade: text-unidecode>=1.3 in /usr/local/lib/p
Requirement already satisfied, skipping upgrade: notebook>=4.4.1 in /usr/local/lib/pytho
Requirement already satisfied, skipping upgrade: ipython-genutils in /usr/local/lib/pyth
Requirement already satisfied, skipping upgrade: jsonschema!=2.5.0,>=2.4 in /usr/local/l
Requirement already satisfied, skipping upgrade: prompt-toolkit<2.0.0,>=1.0.4 in /usr/lo
Requirement already satisfied, skipping upgrade: simplegeneric>0.8 in /usr/local/lib/pyt
Requirement already satisfied, skipping upgrade: pexpect; sys_platform != "win32" in /us
Requirement already satisfied, skipping upgrade: pickleshare in /usr/local/lib/python3.6
```

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

```
4e
Collecting astroid<2.4,>=2.3.0
  Downloading https://files.pythonhosted.org/packages/ad/ae/86734823047962e7b8c8529186a1
    |████████████████████████████████| 215kB 40.0MB/s
Collecting isort<5,>=4.2.5
  Downloading https://files.pythonhosted.org/packages/e5/b0/c121fd1fa3419ea9bfd55c7f9c4f
    |████████████████████████████████| 51kB 9.4MB/s
Requirement already satisfied, skipping upgrade: webencodings in /usr/local/lib/python3.
Requirement already satisfied, skipping upgrade: zipp>=0.5 in /usr/local/lib/python3.6/d
Requirement already satisfied, skipping upgrade: terminado>=0.3.3; sys_platform != "win3
Requirement already satisfied, skipping upgrade: ptyprocess>=0.5 in /usr/local/lib/pytho
Collecting typed-ast<1.5,>=1.4.0; implementation_name == "cpython" and python_version <
  Downloading https://files.pythonhosted.org/packages/90/ed/5459080d95eb87a02fe860d44719
    |████████████████████████████████| 747kB 46.0MB/s
Collecting lazy-object-proxy==1.4.*
  Downloading https://files.pythonhosted.org/packages/0b/dd/b1e3407e9e6913cf178e506cd0de
    |████████████████████████████████| 61kB 9.7MB/s
Requirement already satisfied, skipping upgrade: wrapt==1.11.* in /usr/local/lib/python3
Building wheels for collected packages: pandas-profiling, confuse, visions, htmlmin, tan
  Building wheel for pandas-profiling (setup.py) ... done
  Created wheel for pandas-profiling: filename=pandas_profiling-2.5.0-py2.py3-none-any.w
```

```
    Stored in directory: /root/.cache/pip/wheels/9b/c9/f1/4a2f30c760e017f3e2f46be999c4597a
    Building wheel for confuse (setup.py) ... done
    Created wheel for confuse: filename=confuse-1.0.0-cp36-none-any.whl size=17487 sha256=
    Stored in directory: /root/.cache/pip/wheels/b0/b2/96/2074eee7dbf7b7df69d004c9b6ac4e32
    Building wheel for visions (setup.py) ... done
    Created wheel for visions: filename=visions-0.2.2-cp36-none-any.whl size=53058 sha256=
    Stored in directory: /root/.cache/pip/wheels/53/87/68/294a9e88d82e395b38571df18f7cb71e
    Building wheel for htmlmin (setup.py) ... done
    Created wheel for htmlmin: filename=htmlmin-0.1.12-cp36-none-any.whl size=27084 sha256
    Stored in directory: /root/.cache/pip/wheels/43/07/ac/7c5a9d708d65247ac1f94066cf1db075
    Building wheel for tangled-up-in-unicode (setup.py) ... done
    Created wheel for tangled-up-in-unicode: filename=tangled_up_in_unicode-0.0.3-cp36-non
    Stored in directory: /root/.cache/pip/wheels/c4/57/cc/5f58206efb00418d4dcae8d08a3cb406
    Building wheel for attr (setup.py) ... done
    Created wheel for attr: filename=attr-0.3.1-cp36-none-any.whl size=2459 sha256=4569758
    Stored in directory: /root/.cache/pip/wheels/f0/96/9b/1f8892a707d17095b5a6eab0275da9d3
  Successfully built pandas-profiling confuse visions htmlmin tangled-up-in-unicode attr
  ERROR: google-colab 1.0.0 has requirement requests~=2.21.0, but you'll have requests 2.2
  ERROR: datascience 0.10.6 has requirement folium==0.2.1, but you'll have folium 0.8.3 wh
  Installing collected packages: confuse, tangled-up-in-unicode, attr, visions, htmlmin, p
    Found existing installation: pluggy 0.7.1
      Uninstalling pluggy-0.7.1:
        Successfully uninstalled pluggy-0.7.1
    Found existing installation: pytest 3.6.4
      Uninstalling pytest-3.6.4:
        Successfully uninstalled pytest-3.6.4
    Found existing installation: tqdm 4.28.1
      Uninstalling tqdm-4.28.1:
        Successfully uninstalled tqdm-4.28.1
    Found existing installation: requests 2.21.0
      Uninstalling requests-2.21.0:
        Successfully uninstalled requests-2.21.0
    Found existing installation: pandas-profiling 1.4.1
      Uninstalling pandas-profiling-1.4.1:
        Successfully uninstalled pandas-profiling-1.4.1
                                                                                        .2
```

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

```
  You must restart the runtime in order to use newly installed versions.
```

RESTART RUNTIME

```
import pandas as pd
import pandas_profiling
```

```
pandas_profiling.__version__
```

```
'2.5.0'
```

```
!wget  'https://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/hepatitis.data'
```

```
--2020-02-26 10:36:48--  https://archive.ics.uci.edu/ml/machine-learning-databases/hepat
Resolving archive.ics.uci.edu (archive.ics.uci.edu)... 128.195.10.252
Connecting to archive.ics.uci.edu (archive.ics.uci.edu)|128.195.10.252|:443... connected
HTTP request sent, awaiting response... 200 OK
Length: 7545 (7.4K) [application/x-httpd-php]
Saving to: ‘hepatitis.data’

hepatitis.data      100%[===================>]   7.37K  --.-KB/s    in 0s
```

-1. Title: Hepatitis Domain

-2. Sources: -(a) unknown -(b) Donor: G.Gong (Carnegie-Mellon University) via Bojan Cestnik Jozef Ste
Yugoslavia (tel.: (38)(+61) 214-399 ext.287) } -(c) Date: November, 1988

-3. Past Usage: -1. Diaconis,P. & Efron,B. (1983). Computer-Intensive Methods in Statistics. Scientific
reported a 80% classfication accuracy -2. Cestnik,G., Konenenko,I, & Bratko,I. (1987). Assistant-86: A k
Sophisticated Users. In I.Bratko & N.Lavrac (Eds.) Progress in Machine Learning, 31-45, Sigma Press.

-4. Relevant Information: Please ask Gail Gong for further information on this database.

-5. Number of Instances: 155

-6. Number of Attributes: 20 (including the class attribute) 'Class','AGE','SEX',STEROID','ANTIVIRALS','FA
BIG','LIVER FIRM','SPLEEN PALPABLE','SPIDERS','ASCITES','VARICES','BILIRUBIN','ALK PHOSPHATE','SG(
Attribute information: -1. Class: DIE, LIVE

```
  -2. AGE: 10, 20, 30, 40, 50, 60, 70, 80

  -3. SEX: male, female
```

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

```
  -5. ANTIVIRALS: no, yes

  -6. FATIGUE: no, yes

  -7. MALAISE: no, yes

  -8. ANOREXIA: no, yes

  -9. LIVER BIG: no, yes

 -10. LIVER FIRM: no, yes

 -11. SPLEEN PALPABLE: no, yes

 -12. SPIDERS: no, yes
```

```
-13. ASCITES: no, yes


-14. VARICES: no, yes


-15. BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00


    -- see the note below
-16. ALK PHOSPHATE: 33, 80, 120, 160, 200, 250


-17. SGOT: 13, 100, 200, 300, 400, 500,


-18. ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0


-19. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90


-20. HISTOLOGY: no, yes
```

The BILIRUBIN attribute appears to be continuously-valued.

-8. Missing Attribute Values: (indicated by "?")

```
 Attribute Number:    Number of Missing Values:
               1:    0
               2:    0
               3:    0
               4:    1
```

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

```
               7:    1
               8:    1
               9:    10
                    10:    11
                    11:    5
                    12:    5
                    13:    5
                    14:    5
                    15:    6
                    16:    29
                    17:    4
                    18:    16
                    19:    67
                    20:    0
```

-9. Class Distribution:

```
- DIE: 32
- LIVE: 123
```

```
df = pd.read_csv('hepatitis.data',sep=',',names=['Class','AGE','SEX','STEROID','ANTIVIRALS','
```

```
df.head()
```

	Class	AGE	SEX	STEROID	ANTIVIRALS	FATIGUE	MALAISE	ANOREXIA	LIVER BIG	LIVER FIRM	SPLI PALPAI
0	2	30	2	1	2	2	2	2	1	2	
1	2	50	1	1	2	1	2	2	1	2	
2	2	78	1	2	2	1	2	2	2	2	
3	2	31	1	?	1	2	2	2	2	2	
4	2	34	1	2	2	2	2	2	2	2	

```
df.describe()
```

	Class	AGE	SEX	ANTIVIRALS	HISTOLOGY
count	155.000000	155.000000	155.000000	155.000000	155.000000
std	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
min	1.000000	7.000000	1.000000	1.000000	1.000000
25%	2.000000	32.000000	1.000000	2.000000	1.000000
50%	2.000000	39.000000	1.000000	2.000000	1.000000
75%	2.000000	50.000000	1.000000	2.000000	2.000000
max	2.000000	78.000000	2.000000	2.000000	2.000000

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

```
pandas_profiling.ProfileReport(df)
```

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

```
missing [matrix]: 100%                    2/2 [00:02<00:00, 1.06s/it]

warnings [correlations]: 100%             3/3 [00:00<00:00, 32.16it/s]

package: 100%                             1/1 [00:00<00:00, 3.40it/s]

build report structure: 100%              1/1 [02:39<00:00, 159.73s/it]
```

3.8	9
4.4	9
Other values (25)	83

Toggle details

PROTIME

Categorical

Distinct count	45
Unique (%)	29.0%
Missing	0
Missing (%)	0.0%
Memory size	1.3 KiB

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

74	4
46	4
85	4
Other values (40)	65

Toggle details

HISTOLOGY

Categorical

Distinct count	2

pandas_profiling.ProfileReport(df)

This outputs a bunch of HTML, containing all the information mentioned above.

For me, this tool saves a lot of time. Normally I spend quite a bit of time typing in all the commands tc one to achieve the same results.

```
df.profile_report(title='Hepatitis Data Visualization')
```

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

```
missing [matrix]: 100%                  2/2 [00:06<00:00, 3.05s/it]

warnings [correlations]: 100%           3/3 [00:00<00:00, 26.56it/s]

package: 100%                           1/1 [00:00<00:00, 3.22it/s]

build report structure: 100%            1/1 [02:48<00:00, 168.72s/it]
```

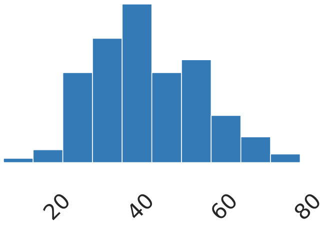
AGE

Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	49
Unique (%)	31.6%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	41.2
Minimum	7
Maximum	78
Zeros (%)	0.0%
Memory size	1.3 KiB

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu



Toggle details

SEX

Categorical

Categorical

```
profile = df.profile_report()
profile.to_file('Pandas Profiling Report — Hepatitis.html')
```

```
variables: 100%                                     20/20 [00:17<00:00, 1.15it/s]

correlations [recoded]: 100%                        6/6 [00:10<00:00, 1.82s/it]

interactions [continuous]: 100%                     1/1 [00:06<00:00, 6.05s/it]

table: 100%                                         1/1 [00:05<00:00, 5.84s/it]

missing [matrix]: 100%                              2/2 [00:05<00:00, 2.89s/it]

warnings [correlations]: 100%                       3/3 [00:00<00:00, 24.00it/s]
```

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

```
build report structure: 100%                        1/1 [31:55<00:00, 1915.05s/it]
```

FeatureTools

Another thing that takes a lot of time when building **a model is feature engineering**. On an average pr
time extracting and transforming data into a proper machine learning dataset.

We can however reduce this time by quite a bit if we use the right tools for the job. One such tool is th

This tool automatically gathers up features from a bunch of tables and transforms these f
learning dataset.

It works like this:

```
import featuretools as ft
```

```
2,3,4,5,6,7,8,9,10],'name':['A','B','C','D','E','F','G','H','I','J']})
,4,5,6,7,8,9,10,3,5,6],
','#55672','#3677','#456','#45577','#423534','#3125','#256253','#1237','#4444','#123457','#123
 products','cake','cereals','bread','cooking oil','books','crockery','spices','condiments','ve
```

```
entities = { 'customers': (customer_df, 'customer_id'), 'orders': (orders_df, 'order_id') }
```

```
relationships = [ ('customers','customer_id','orders','customer_id') ]
```

```
feature_matrix, feature_defs = ft.dfs( entities=entities, relationships=relationships, target_
```

```
feature_defs
```

```
[<Feature: name>,
 <Feature: COUNT(orders)>,
 <Feature: NUM_UNIQUE(orders.order_item)>,
 <Feature: MODE(orders.order_item)>]
```

```
feature_matrix
```

	name	COUNT(orders)	NUM_UNIQUE(orders.order_item)	MODE(orders.order_item)
customer_id				
1	A	1	1	dairy products
2	B	1	1	cake
				als
				ead
5	E	2	2	cooking oi
6	F	2	2	books
7	G	1	1	crockery
8	H	1	1	spices
9	I	1	1	condiments
10	J	1	1	vegetable

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

- First we define the entities in our database, which in our case are customers and orders. Next w customers. The customer is the parent entity (one) and orders is the child entity (many).
- We then ask FeatureTools to build a dataset for us, where we choose customers as the target er dataset with customer as the parent and engineer features from orders as observations related

- As you can see, FeatureTools is especially useful when you have a large database with many tal learning dataset from that database. It also works great on temporal data.
- We can specify exactly which feature engineering primitives the tool should use, such as sum, c

Please note that FeatureTools doesn't handle normalization, scaling and other operations the common issues with data.

LIME

If you're an active machine learning engineer you may have noticed that more customers than ever as decision. They no longer blindly trust the model.

This can be quite a challenge as most models are hard to explain to a customer. But there is a solutio

LIME (Local Interpretable Model Explainer) is a tool that allows you to explain a decision made by yo a decision tree, random forest or even a neural network.

LIME explains a prediction so that even the non experts could compare and improve on an untrustwo An ideal model explainer should contain the following desirable properties:

- **Interpretable** It should provide qualitative understanding between the input variables and the re
- **Local Fidelity** It might not be possible for an explanation to be completely faithful unless it is th Having said that it should be at least locally faithful i.e it must replicate model's behaviour in the
- **Model Agnostic** The explainer should be able to explain any model and should not make any as explanations.
- **Global perspective** The explainer should explain a representative set to the user, so that the use

io

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

Sigma Cab's Surge Pricing Type Classification

This was a multiclass classification problem. Here we will see how to make predictions for this datas **interpretable**.

The intent here is not to build the best possible model but rather the focus is on the aspect of interp

https://www.analyticsvidhya.com/blog/2017/06/building-trust-in-machine-learning-models/

Import sample data

```
from google.colab import files
files.upload()
```

Choose Files test_XaoFywY.csv

- **test_XaoFywY.csv**(application/vnd.ms-excel) - 4921161 bytes, last modified: 2/26/2020 - 100% done

```
Saving test_XaoFywY.csv to test_XaoFywY.csv
{'test_XaoFywY.csv': b'Trip_ID,Trip_Distance,Type_of_Cab,Customer_Since_Months,Life_Styl
```

Load Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.base import TransformerMixin
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_validate
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
```

Import Data

```
train_df = pd.read_csv('/content/train_63qYitG.csv')
test_df = pd.read_csv('/content/test_XaoFywY.csv')
```

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

	Trip_ID	Trip_Distance	Type_of_Cab	Customer_Since_Months	Life_Style_Index	Cor
0	T0005689460	6.77	B	1.0	2.42769	
1	T0005689461	29.47	B	10.0	2.78245	
2	T0005689464	41.58	NaN	10.0	NaN	
3	T0005689465	61.56	C	10.0	NaN	
4	T0005689467	54.95	C	10.0	3.03453	

```
test_df.head()
```

	Trip_ID	Trip_Distance	Type_of_Cab	Customer_Since_Months	Life_Style_Index	Cor
0	T0005689459	9.44	A	10.0	2.57438	
1	T0005689462	32.15	B	10.0	2.85143	
2	T0005689463	10.38	C	4.0	2.70530	
3	T0005689466	14.94	NaN	6.0	2.48159	
4	T0005689468	32.03	B	7.0	2.81598	

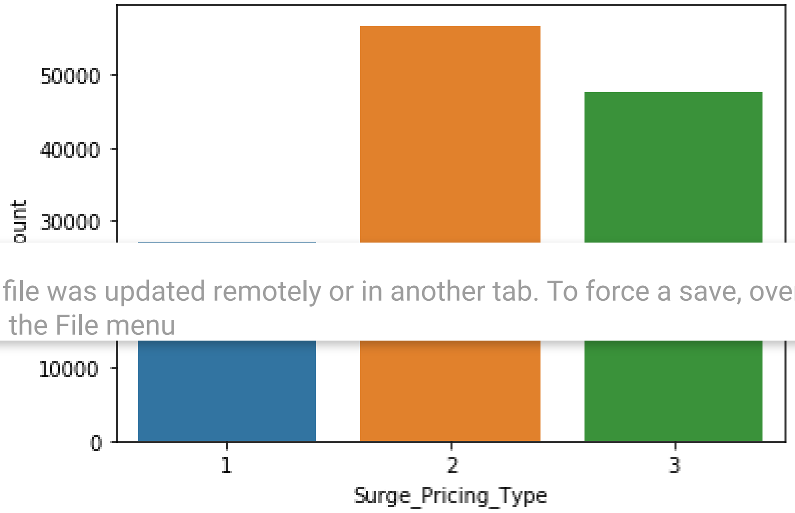
```
train_df.shape,test_df.shape
```

```
((131662, 14), (87395, 13))
```

Visualize Labels

```
import seaborn as sns
import pandas_profiling
import matplotlib.pyplot as plt
%matplotlib inline
sns.countplot(x='Surge_Pricing_Type',data=train_df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9ec4c11a20>
```



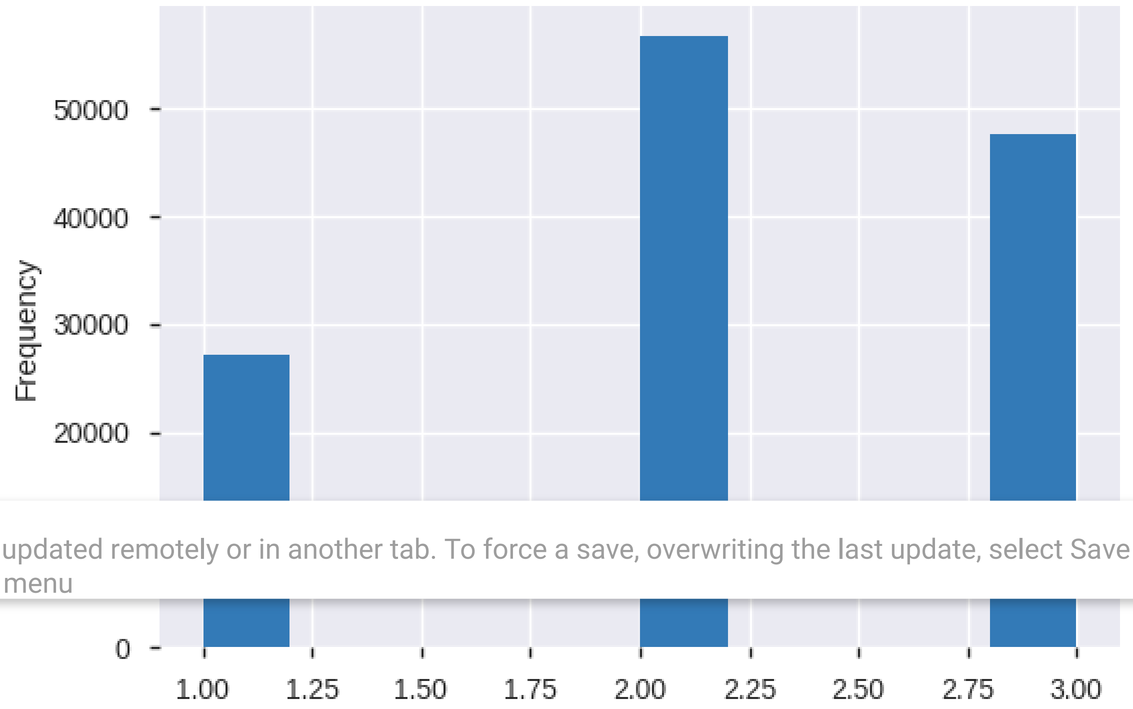
This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

```
pandas_profiling.ProfileReport(train_df)
```

95-th percentile	3
Maximum	3
Range	2
Interquartile range	1

Descriptive statistics

Standard deviation	0.73816
Coef of variation	0.34242
Kurtosis	-1.1368
Mean	2.1557
MAD	0.61199
Skewness	-0.25515
Sum	283830
Variance	0.54489
Memory size	1.0 MiB



This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

Value	Count	Frequency (%)
2	56728	43.1%
3	47720	36.2%
1	27214	20.7%

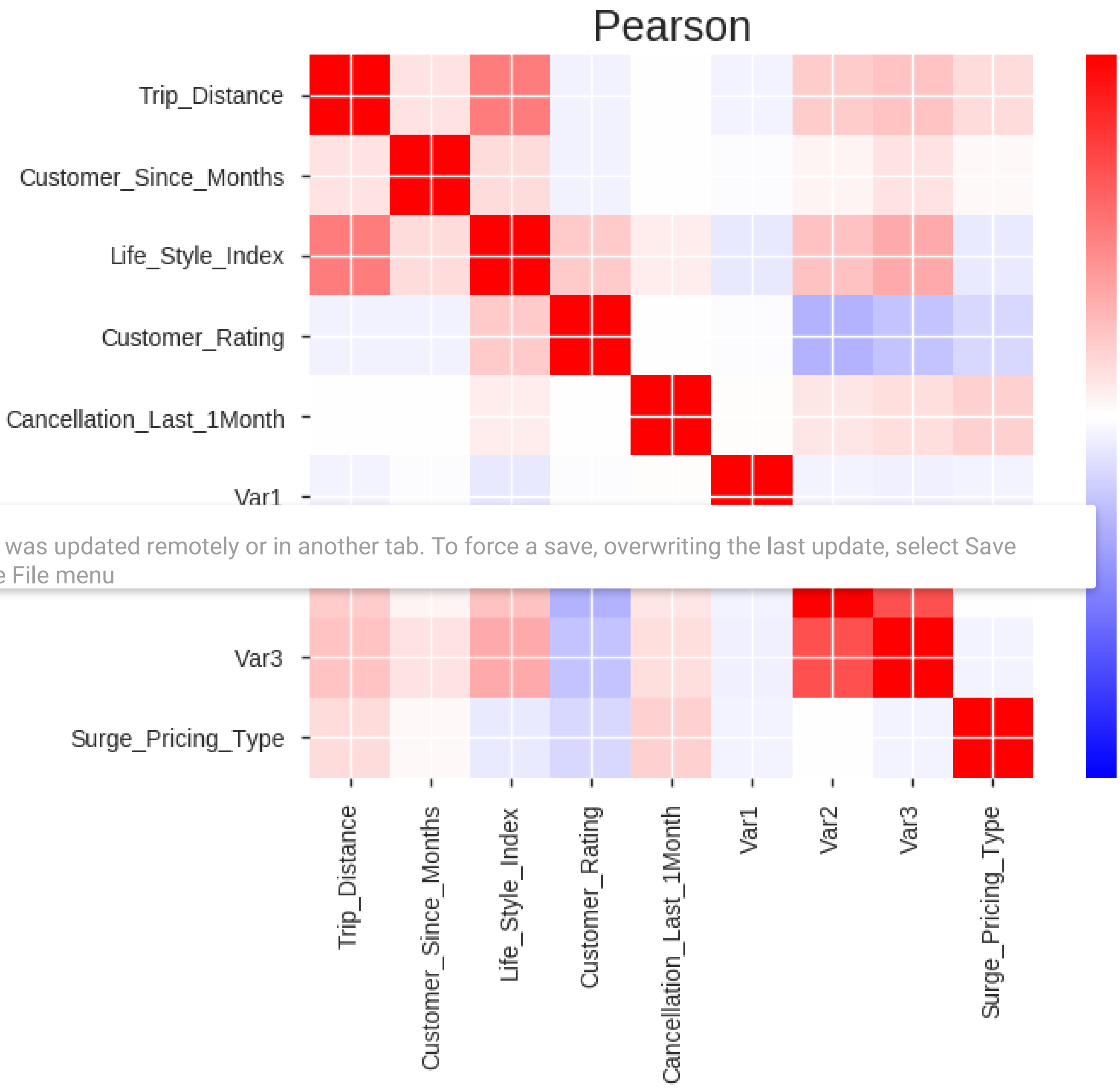
Minimum 5 values

Value	Count	Frequency (%)
1	27214	20.7%
2	56728	43.1%
3	47720	36.2%

Maximum 5 values

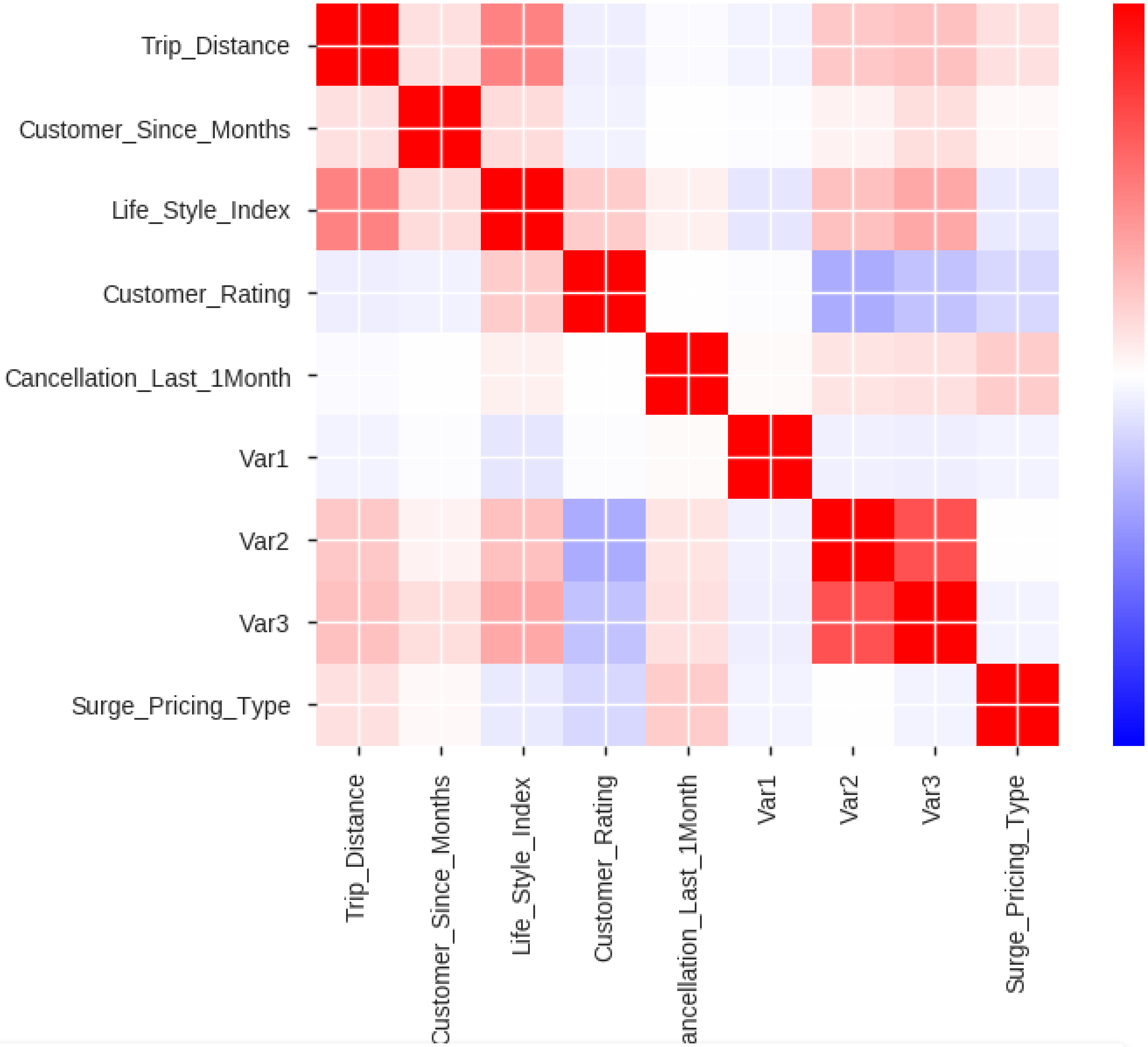
Value	Count	Frequency (%)
1	27214	20.7%
2	56728	43.1%
3	47720	36.2%

Correlations



This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

Spearman



This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

Sample

	Trip_ID	Trip_Distance	Type_of_Cab	Customer_Since_Months	Life_
0	T0005689460	6.77	B	1.0	
1	T0005689461	29.47	B	10.0	
2	T0005689464	41.58	NaN	10.0	
3	T0005689465	61.56	C	10.0	
4	T0005689467	54.95	C	10.0	

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu

This file was updated remotely or in another tab. To force a save, overwriting the last update, select Save from the File menu