Applied Data Science

Define Problem /Understanding Problem

Group 4

Introduction:

There are several excellent notebooks to study data science competition entries. However many will skip some of the explanation on how the solution is developed as these notebooks are developed by experts for experts. The objective of this notebook is to follow a step-by-step workflow, explaining each step and rationale for every decision we take during solution development.

Work Flow Stages

- 1.Question or problem definition.
- 2.Acquire training and testing data.
- 3.Wrangle, prepare, cleanse the data.
- 4.Analyze, identify patterns, and explore the data.
- 5.Model, predict and solve the problem.
- 6.Visualize, report, and present the problem solving steps and final solution.
- 7.Supply or submit the results.

INPUT:

- # This Python 3 environment comes with many helpful analytics libraries installed
- # It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python
-
- #load packages
- import sys #access to system parameters https://docs.python.org/3/library/sys.html
- print(“Python version: {}". Format(sys.version))

```
import pandas as pd #collection of functions for data processing and analysis modeled after R
dataframes with SQL like features print("pandas version: {}". format(pd.__version__))
```

```
import pandas as pd #collection of functions for data processing and analysis
modeled after R dataframes with SQL like features print("pandas version: {}".
format(pd.__version__))
```

```
import matplotlib #collection of functions for scientific and
publication-ready visualization print("matplotlib version: {}".
format(matplotlib.__version__))
```

```
import numpy as np #foundational package for scientific
computing print("NumPy version: {}". format(np.__version__))
```

```
import scipy as sp #collection of functions for scientific
computing and advance mathematics print("SciPy version: {}".
format(sp.__version__))
```

```
import IPython from IPython import display #pretty printing of
dataframes in Jupyter notebook print("IPython version: {}".
format(IPython.__version__))
```

```
import sklearn #collection of machine learning algorithms
print("scikit-learn version: {}". format(sklearn.__version__))

#misc libraries import random import time

 #ignore warnings import warnings
 warnings.filterwarnings('ignore') print('-'*25)

   from subprocess import check_output
   print(check_output(["ls", "../input"]).decode("utf8"))
```

Output:

```
• Python version: 3.6.3 |Anaconda custom (64-bit)| (default, Nov 20 2017, 20:41:42)
• [GCC 7.2.0]
• pandas version: 0.20.3
• matplotlib version: 2.1.1
• NumPy version: 1.13.0
• SciPy version: 1.0.0
• Ipython version: 5.3.0
• scikit-learn version: 0.19.1
• -------------------------
• gender_submission.csv
• test.csv
• train.csv
```