PROJECT REPORT:

PYTHON FOR DATA SCIENCE

Ousmane Bawa Gaoh Moustapha

Shishir Dubey

Shashank Sharma

International Master of Science

Big Data

National School of Statistics and Information Analysis (ENSAI)

This report is submitted for the subject of

Python for Data Science

February 12, 2016

CONTENTS

INTRODUCTION

The project is aimed at the implementation of a case study in linear regression ('DAT21 Linear Regression Assignment'), published on the kaggle website. Machine learning techniques learnt during the lecture on 'Python for Data Science'at ENSAI were utilized to implement the case study. Training and testing datasets from the kaggle website were used to implement the project. We used IPython Notebook as the primary tool to implement and test our code. Mean Absolute Error (MAE) results obtained using different machine learning models were tested on the kaggle website and final scores were reported.

We have discussed our problem solving methodology and MAE scores obtained in the subsequent sections.

1.1 Methodology

Step 1: Theoretical concepts and practical exercises in Python, learnt during the classroom training at Ensai were revised.

Step 2: The problem statement was thoroughly understood in order to formulate the machine learning algorithms to be implemented.

Step 3: The training and testing datasets were studied, in order to spot the data processing methods to be implemented.

Step 4: The datasets were cleaned and processed, and all major discrepancies were removed.

Step 5: A descriptive analysis of the data was carried out in order to understand the relevancy of variables and observations.

Step 6: Categories were created for each of the variables of interest in the dataset.

Step 7: Various methods for machine learning and prediction were carried out. Initially, Linear Regression was carried out using one variable, and then by adding one additional variable at a time. Subsequently, Decision Tree, Support Vector Machine, and K Nearest Neighbours regression were carried out by using the variables that gave us the best performance during Linear Regression.

Step 8: Different values of MAE were tested on the kaggle website and reported.

Step 10: Advanced feature engineering tools, such as TF-IDF, were implemented.

2. RESULTS

The following MAE scores were obtained by testing on the kaggle website, and were submitted on the online shared MS-Excel workbook:

1. 9558.13645
2. 10155.67498
3. 10375.40612
4. 10967.524
5. 11064.54668
6. 12568.42989

3. REFERENCES

- 'Python for Data Science' by M. Koby Karp, M. Christophe Blefari, and M. Herve Mignot.

- 'matplotlib' library for Python (http://matplotlib.org/users/index.html)

- 'pandas' library for Python (http://pandas.pydata.org/)

- 'numpy' library for Python (http://www.numpy.org/)

- 'sklearn' library for Python (http://scikit-learn.org/stable/)

- 'bs4' library for Python (https://pypi.python.org/pypi/beautifulsoup4/)

- 'nltk' library for Python (https://pypi.python.org/pypi/nltk/)