Dimensionality Reduction for Studying Diffuse Circumgalactic Medium

GSOC 2021 Project Proposal

Proposal by:
Hassan Raza Bukhari

for Organization:
Machine Learning for Science (ML4Sci)

About me

Name: Hassan Raza Bukhari

Email: hbukhari.bese18seecs@seecs.edu.pk

Github: github.com/eruditehassan

Linkedin: linkedin.com/in/syedhassanbukhari/

University: National University of Sciences and Technology (NUST), Islamabad

Major Interests: Data Analysis, Machine Learning, Software Development & Problem solving with algorithms

Relevant Skills (in terms of technologies):

Experience level **Intermediate or Higher**

- **Python** (pandas, numpy, matplotlib, seaborn, scikit learn, pytorch, OOP)
- **R** (dplyr, ggplot2, tidyr, knitr)
- **Data Structures and Algorithms**
- **Machine Learning on Cloud and deployment of models** (hands on experience with AWS and Azure)
- Strong grip on **Data Cleaning and Preprocessing**
- **Databases**

Other Skills

Expeerience level **Basics to Intermediate**

- **Web Development** (HTML, CSS, Javascript, PHP, basics of React)

Community Activites:

- Microsoft Learn Student Ambassador and Marketing Lead at MLSA Islamabad
- Data Science lead at Google Developer Student Club NUST
- Vice Chair Publicity IEEE SEECS

Project Details

The Project details including background, motivation, task related and timeline information is given below:

Background of Cicrcumgalactic Medium (CGM) Study

The history of gas that cycles between the visible body of the galaxy and its Circumgalactic Diffuse Medium (CGM) encapsulate the history of Galaxy. Therefore, the history of CGM is being researched by using various techniques including Machine Learning.

Primary Task and Expected Outcome of Project

Primary Task: The Primary task of the project is the implementation of Machine Learning techniques that will reduce the dimensions of the datasets. These techniques should be applicable to quasar absorption spectra datasets.

Expected Outcome: The model implemented should reduce the dimensions and also maintain a high level of accuracy compared to the results from the original datasets.

Why did I chose this project?

The reason why I chose this specific project is because apart from having a hope of an excellent learning experience what excites me the most is that my contributions would help in solving a real world problem, thinking about that motivates me and excites me.

Proposed Flow

1. **Feature Understanding, Data Cleaning and Preprocessing**

   The first step would be to spend some time understanding the data and examining it. Then, looking for ways to make the data cleaner and more optimized for Machine Learning models.

2. **Training different models using Original Dataset**

   Finding various models that would be suitable for the project. Training the models using the actual dataset to get an understanding of how well each one of them perform on the dataset, then compring results to find the best model for the task at hand.

3. Tuning the model

Manual and Automated tuning techniques can be utilized to tune the model to perform optimally. The best proposed approach is running automated hyperparameter tuning job on a powerful cloud server to get the best results. This step would ideally be done incrementally when training each model so that we would know how well each model performs when tuned properly.

4. Studying and Using various Dimensionality Reduction techniques

One dimensionality reduction technique might work very well for one dataset, but not for another one. Therefore, various alternatives must be considered. It is proposed that significant effort and time would be put into examining various techniques, and using them on the dataset and then comparing the results generated by all of them. Some possible techniuqes could be PCA, KPCA, LDA, SVD, t-SNE, etc.

5. Reducing Dimensions, Training and Comparing Results

This stage will take most of the time for the whole project because evey time the dimensions are reduced using a technique the model would have to be trained again and the accuracy would be compared to the results obtained earlier. At the end of this stage, one dimensionality reduction technique will be shortlisted which provides the optimal results.

6. Tuning the model on Reduced Dimensions

This can also be done in parallel with the step 5 where the process is carried out incrementally by first reducing dimensions using a technique, then training a model on it and also tuning the model's hyperparameters and then estimating the accuracy. This would be done for all the different dimensionality reduction techniques. Alternatively, the accuracy can only be compred without tuning, and then the best performing model will be tuned. This would depend largely on the hardware resources available.

7. Documenting the results

Ideally it would be carried out at all stages of the process where everything would be documented so that anything can be traced back and corrected if and when required.

Project Timeline

No.	Time Period	Task
1	Community Bonding Period (May 17 - June 7)	The community bonding period would provide me with a chance to get familiar with the mentors and the community. In addition to that it would give me time to parallelly do research on what needs to be done and to improve myself in areas where I lack so that I perform best during the actual coding period. I have following tasks in mind: • Studying previous work done on the project • Examining what additional skills or knowledge would be required for proper flow of the project. • Seeking advice from mentors regarding this • Making a plan about self improvements for better work on the project and also collecting resources that might be helpful. • Working on the plan and being in touch with the mentors and getting their feedback on my learning. • Making notes to be used during Coding Period

No.	Time Period	Task
2	Week 1 (June 7 - June 13)	This week would be dedicated to better understanding of the feature which would be helpful in later stages of work. This includes: • Understanding the features of the data by: ◦ Examining the ranges of values of each feature ◦ Understanding the importance and significance of each feature ◦ Dependency of features to see if any feature depend on the other (would be useful for reducing dimensions) ◦ Doing exploratory data visualization to get better understanding of the features. • Documenting the results obtained at each stage and making notes for self.
3	Week 2 (June 14 - June 20)	Having good understanding of features and data, this week would be focused around preprocessing and cleaning the data for optimal results in the later stages: • Preprocessing / Cleaning the dataset ◦ Handling missing values by either removal or imputation (also checking which imputation technique will give the best results) ◦ Encoding the categorical data ◦ Doing Feature Scaling (Standardization / Normalization) ◦ Splitting dataset for future use • Documenting the tasks during each processing step along with rationale for the decision.

No.	Time Period	Task
4	Week 3 (June 21 - June 27)	This week would be dedicated to finding the most suitable models for the project. This would include following tasks: • Doing research to shortlist the best Machine Learning / Deep Learning models for the task at hand • Studying results obtained by other Data Scientists using same models and understanding the preprocessing techniques and other factors influencing their results, also getting an idea of the type of data being used. • Testing various models on a subset of our dataset to shortlist the models to be used at later stages. • Documenting the rationale for choice.
5	Week 4 (June 28 - July 4) & Week 5 (July 5 - July 11)	Training and tuning the models would be the prime objective of these two weeks. Ideally the training and tuning of different models would be carried out in parallel to get the most out of the time. • Running training and tuning jobs on the shortlisted models • Comparing results to choose the best model which not only provides an optimal accuracy, but also which effectively minimzes the overfitting. • Documenting the process at each stage.

No.	Time Period	Task
6	Week 6 (July 12 - July 18)	The prime objective of this week would be to research about and shortlist possible Dimensionality reduction techniques. It would involve following tasks: • Studying in detail about various dimensionality reduction techniques that can be applied to the task at hand to understand what kind of techniques would give the best results. • Testing on subset of dataset and examining the way each technique is working and what criteria is being used to reduce dimensions. • Doing an analysis of which technique works best. • Experimenting by combining various techniques to get results. • Training a model on subset of data using some of the techniques to see how well they maintain accuracy when dimensions are reduced. • Documenting the results obtained
7	Week 7 (July 19 - July 25)	This week will continue the work done in the last week and build on top of that. It will involve following tasks: • Benchmarking the performance of the remaining dimensionality reductio techniques. • Shortlisting 2 or 3 best techniques obtained using the above process. • Running those techniques on the entire data set and examining the features after reduction. • Documenting the results obtained

No.	Time Period	Task
8	Week 8 (July 26 - August 1) & Week 9 (August 2 - August 8)	The work in this period would focus on training, testing and tuning of the model using the reduced dimension data. • Training the model using the new reduced dimensions data. • Testing the model to compare its accuracy with the model that was trained on original dataset. • Running hyperparameter tuning jobs on it to tune it properly and then repeating the same process above to check its accuracy when the model is tuned. • Repeating the training, testing and tuning for other shortlisted Dimensionality reduction techniques as well • Documenting during all the steps
9	Week 10 (August 9 - August 15 & August 16)	Buffer Period to be used in case of any unforseen delay

Experience & More Information about Me

A brief section about the relevant experience that I have:

Pull Requests

I have made the following pull requests:

- ML4SCI.github.io Pull Request #2 (merged) : Added Website Documentation section
- ML4SCI.github.io Pull Request #3 (merged) : Added additional GSOC related documenation

Projects done as a Professional

These are the projects that I have completed as a professional freelancer (listing in order of latest first):

1. Improvements in **Voting Data Analysis System**
2. Fixing a buggy **Recommendation System**
3. **Data Analysis** Code Documentation
4. Data Analysis using R (involved **implementation of custom algorithm** for data analysis and **statistical time series forecasting**)
5. **Voting Polarization** calculation with **Pandas** and **Jupyter Notebooks**
6. **App Score Prediction** using R
7. **Non-parametric Model** Implementation in Python
8. Fixing a **Lead Scoring Algorithm**
9. **Forecasting model** Implementation in Python
10. Bio statistics **Data Analysis**

Reference: My Upwork Profile

Relevant Community Work

Working with various communities I have done following work that is relevant:

1. **Main speaker** for Data Science training session (**Data Science Scifi)** organized by combined efforts of 12 Google Developer Student Clubs in Pakistan that had **2 Google Developer Experts** as speakers as well (Reference Post)

2. **Speaker** for **Basics of Data Analysis with R** session organized by **Microsoft Learn Student Ambassadors Islamabad** (Reference Post)

Research Experience

Worked as a **Data Science research Assistant** for **IoT Lab SEECS, NUST.** It involved the following work:

- Analysis of Water Quality using Satelite imagery
- Study and review of relevant research papers
- Analysis of data obtained using IoT devices

Other Notable Experiences

- Wrote a reseach report titled **Innovation of Digital Marketing with Machine Learning and Data Science** (availalbe on featured content section of my Linkedin profile)
- **Udemy Instructor** (having over 10k students from over 130 countries)

Availability

Due to exams I might be away during the end of Community bonding period or start of coding time for one week, but I will make up for that. The exam dates are not confirmed due to uncertainity of condition of Covid in Pakistan. Apart from that, I would be readily available to work. I intend to spend 5 - 6 hours each day and do not mind working on weekends as well.

Plan after GSOC

Being able to work on something so exciting would be a great learning experience for me, it would not only groom me in terms of my technical knowledge, but also improve my collaboration and teamwork skills, and at the same time giving me an opportunity to work on something that is intended to solve a real world problem. Therefore, **I plan on continuing to contribute after GSOC as well.**