PERSONAL DETAILS

Name: Rajat Gupta

Course: PGDBA (Post Graduate Diploma in Business Analytics)

Course Website: https://www.isical.ac.in/~pgdba/Programme_Overview.html

Universities: IIM Calcutta, IIT Kharagpur, ISI Kolkata

Time Zone: Indian Standard Time (IST) – GMT + 5:30

GitHub: https://github.com/rajatguptakgp

LinkedIn: https://www.linkedin.com/in/rajatguptakgp

Contact Details:

1. **Primary email-id:** rajatgba2021@email.iimcal.ac.in
2. **Secondary email-id:** rajatgupta069@gmail.com

BIOGRAPHICAL INFORMATION

Hi, I am Rajat Gupta, student of the PGDBA program jointly offered by IIM Calcutta, IIT Kharagpur and ISI Kolkata. I am currently interning at **Amazon** under the role of **Data Scientist II**. My internship started on **4th January 2021** and will end on **4th June 2021**.

I have been a part of variety of projects in the fields of **Computer Vision**, **NLP**, and **Reinforcement Learning** among others that I have done in the duration of course, as can be seen in my resume. I have a strong knowledge of **Python, PyTorch/Tensorflow/Keras** and machine learning algorithms from Linear/Logistic Regression to ensemble methods like bagging and boosting to advanced neural network frameworks like Autoencoders and GANs. I am also an enthusiastic participant in Data Science competitions and quizzes. Last year, I stood **1st** among **2035 aspirants** of Data Science Quiz, conducted by Analytics Club, IIT Guwahati.

I have been part of three remote academic internships during the course. Last year, I worked under my two professors from **Computer Science & Engineering Department, IIT Kharagpur** for two projects for a duration of four months from April '20 till July '20. The two projects that I had worked on were:

1. **Generating explainable recommendations**, where I built a machine learning framework using graphs (Python library used: **Networkx**)
2. **Building knowledge-aware chatbots:** Transformer framework with problem modelled as Reinforcement Learning

I am also an active open-source contributor. I maintain my code works of all the projects and learnings in my GitHub repository. I like to code machine learning algorithms from scratch to have a deeper knowledge of algorithms and to be an efficient coder.

I want to participate in GSOC '21 as this is my last opportunity to be a part of something so unique and contributing to open-source. In order to showcase my skills, sincerity and dedication, and to increase chances of selection, I have worked on evaluation tasks of **six** projects offered by ML4SCI organization and as listed below and **all of my three proposals** are for these projects:

1. Deep Regression Techniques for Decoding Dark Matter with Strong Gravitational Lensing
2. Domain Adaptation for Decoding Dark Matter with Strong Gravitational Lensing
3. Equivariant Neural Networks for Dark Matter Morphology with Strong Gravitational Lensing
4. Machine Learning for Turbulent Fluid Dynamics
5. Decoding quantum states through nuclear magnetic resonance
6. Dimensionality Reduction for Studying Diffuse Circumgalactic Medium

SYNOPSIS

PROJECT PROPOSAL 1: Dimensionality Reduction for Studying Diffuse Circumgalactic Medium

In this project, we will explore efficacy of different dimension reduction techniques in Machine Learning on simulated quasar absorption spectra to determine gas composition, temperature and density. We start by building machine learning classifiers on original simulated dataset using models like

1. Logistic Regression
2. Ensemble methods
    a. Bagging:
        i. Random Forest Classifier
    b. Boosting:
        i. LightGBM Classifier
        ii. XGBoost Classifier
3. Neural Networks - MLP

We then use dimension reduction techniques like – **PCA, SVD and Isomap** to reduce dimensions of data and assess classifier performance across same set of above classifiers. We will further look at advanced neural network frameworks like **Autoencoders** and **Semi-Supervised GANs**, for classification while generating embeddings as latent representation of data in reduced dimensions.

Finally, we compare the model performances across original and dimensionally-reduced dataset and conclude with the best model.

PROJECT PROPOSAL 2: Decoding quantum states through nuclear magnetic resonance

Since it is hard to classify materials at low temperatures, in this project, we will study time evolution of nuclear spins via simulations of nuclear magnetic resonance (NMR). We will explore different machine learning frameworks in classification from Logistic Regression to Bagging and Boosting Classifiers to Neural Networks, to determine type of electronic interaction based on simulated data.

Additionally, **important features responsible for classification will be extracted** from the model using following methods:

1. Feature Importances criterion in Ensemble models
2. Permutation Feature Importance
3. SHAP

Finally, we will build the optimization algorithm for estimating best parameters borrowing approach of **Grid Search CV** and **Randomized Search CV**.

PROJECT PROPOSAL 1: Dimensionality Reduction for Studying Diffuse Circumgalactic Medium

PROJECT MILESTONES + DELIVERABLES

Week Number	**Dates**	**Milestones**	**Deliverable (apart from Weekly Report)**
Week 1	June 7 – June 14	Understanding absorption spectrum	
Week 2	June 14 – June 21		Literature Survey Report
Week 3	June 21 – June 28	Simulating Quasar absorption spectra	Dataset to be used for further analysis
Week 4	June 28 – July 5	Modelling classifier on original dataset	Model performance through metric **(AUROC)**
Week 5	July 5 – July 12	Dimensionally reduce data using PCA, LDA, SVD	Comparing classifier performance of dimensionally reduced dataset with original dataset + Code-base of all techniques explored
Week 6	July 12 – July 19	**Mid - Evaluation**	**Presentation + Documentation of work done in past 5 weeks**
Week 7	July 19 – July 26	Exploring neural networks for Dimension Reduction	
Week 8	July 26 – August 2		Building a neural network with embeddings as latent representation
Week 9	August 2 – August 9	Exploring advanced neural network frameworks	Modelling **Autoencoder NN + Semi-supervised GAN (SGAN)**
Week 10	August 9 – August 16	Comparing results for all techniques explored + wrap-up	**Code Base + Documentation of work done in last 4 weeks**
Week 11	August 16 – August 23	**Code Submission + Final Evaluation**	Final presentation

I plan to deliver the following during the course of project:

1. I will be submitting a weekly report to give updates about the work that I have done for the week. This report can go up to 3-4 pages depending on the level of detail. This report will help me align with the objectives of the project and will also give a clear idea about my update to my mentor.

2. Apart from the weekly report, I intend to deliver the following:

    a. A literature survey report highlighting my understanding of absorption spectrum and relevant theory

    b. Dataset to be used for analysis

    c. Code-base for classifiers tested on original dataset

    d. Code-based for classifiers tested on dimensionally-reduced dataset using techniques discussed above

e. **Table of model performances across original dataset and dimensionally-reduced dataset**.

The table should look like as follows:

Classifier	**TEST AUROC**	
	Original Dataset	**Dimensionally reduced dataset**
Logistic Regression	---	---
LightGBM	---	---
MLP	---	---
Autoencoder NN	---	---
SGAN	---	---

f. Documentation of the complete approach during mid and final evaluation

PROJECT OUTLINE

Insights from the evaluation test: In task 3, I had trained the original dataset using LGBM (Light Gradient Boosting Machine) Classifier and Neural Network (Multi-Layer Perceptron). The results that I had got are as follows:

Model	Test AUROC
Logistic Regression	0.634
LGBM Classifier	0.733
Multi-Layer Perceptron (MLP)	0.706

I had also ensured that my models are **not overfitting**.

1. For **MLP**, I had checked for training and validation loss which were close to each other
2. For **LGBM**, I had plotted **Train AUROC and Test AUROC** and ensured that they were close to each other.

Next, I had used two dimension reduction approaches – **Principal Component Analysis (PCA)** and **Autoencoder Neural Network** for reducing dimensions ensuring that there isn't significant degradation in classifier performance.

In case of autoencoder neural network, I had learnt the distribution of data-points of class 0 from training set and reconstructed the inputs for both class labels 1 and 0 from test set. **Logically, reconstruction loss for data-points of class 1 is going to be higher than for data-points of class 0, since the model has learnt distribution of data from class 0, and so we can label data-points as of class 1 if the reconstruction loss is above a certain threshold and 0 otherwise.** Based on that, we can get ROC metrics (FPR, TPR) for a range of thresholds.

For both approaches, I had considered reducing dimensions of data from **28 features** to **20 features** and the results of classifiers on reduced dataset are as follows:

Model	Test AUROC
LGBM Classifier	0.653
Multi-Layer Perceptron (MLP)	0.524

However, **reducing dimensions from 28 to 20, does not seem to be a huge reduction** and I believe there lies a scope to reduce dimensions further while achieving similar model performances. Keeping this in mind, my approach for the project is going to be as follows:

1. After understanding the theory behind absorption spectra and relevant materials, I will perform simulations on quasar absorption spectra to create the dataset.
2. Next, I will model a variety of classifiers from basic like Logistic Regression, Decision Tree to ensemble methods like Bagging and Boosting. I will try to achieve best model performances with tuning hyperparameters.
3. Next, I will try neural networks to model a classifier and compare performance with above models. I intend to complete this by mid-evaluation.

4. Now, in order to dimensionally reduce the data, I will try approaches like **PCA, SVD and Isomap** and follow steps in similar order as described above – training classifiers, hyperparameter tuning.
5. Finally, I would like to explore advanced neural network framework like Autoencoder NN and Semi-Supervised GAN (SGAN) for classification.
6. Comparing performances for all models and concluding with the best technique.
7. Final Presentation + Documentation + Submission of Code Base

PROJECT PROPOSAL 2: Decoding quantum states through nuclear magnetic resonance

PROJECT MILESTONES + DELIVERABLES

Week Number	Dates	Milestones	Deliverable (apart from Weekly Report)
Week 1	June 7 – June 14	Understanding NMR Simulation Code	
Week 2	June 14 – June 21		Dataset to be used for analysis
Week 3	June 21 – June 28	Building classification models – **Logistic Regression, LGBM, XGBoost**	
Week 4	June 28 – July 5		Code-base for classifiers tested
Week 5	July 5 – July 12	Hyperparameter tuning + Model Evaluation (overfitting/underfitting)	Summary of model performances + Important Features of dataset
Week 6	July 12 – July 19	**Mid - Evaluation**	**Presentation + Documentation of work done in past 5 weeks**
Week 7	July 19 – July 26	Building Neural Network for predictions (MLP)	
Week 8	July 26 – August 2		Model performance
Week 9	August 2 – August 9	Developing optimization algorithm for estimating parameters	
Week 10	August 9 – August 16	Analysis wrap-up	**Code Base + Documentation of work done in last 4 weeks**
Week 11	August 16 – August 23	**Code Submission + Final Evaluation**	Final presentation

I plan to deliver the following during the course of project:

1. I will be submitting a weekly report to give updates about the work that I have done for the week. This report can go up to 3-4 pages depending on the level of detail. This report will help me align with the objectives of the project and will also give a clear idea about my update to my mentor.
2. Apart from the weekly report, I intend to deliver the following:
   a. Dataset to be used for analysis
   b. Comparison of results of different classifiers tested on dataset
   c. Description of most important features of dataset calculated through **feature importances attribute in ensemble classifiers**, **Permutation Feature Importance** method and **SHAP (Shapley Additive Explanations)**
   d. Model performance for neural network trained (Multi-Layer Perceptron) on dataset
   e. Presentation + Documentation + Code-base

PROJECT OUTLINE

Insights from the evaluation test: In this task, I had taken three approaches to estimate parameters - α, ξ and Γ, which are as follows:

1. Multi-output regression:
    a. Linear Regression
    b. Random Forest Regressor
2. Chained Regression:
    a. LightGBM Regressor
3. Neural Networks:
    a. Multi-Layer Perceptron (MLP)

and the results I got are as follows:

Model	Train MSE	Test MSE
Linear Regression	0.559	**1.183**
Random Forest Regressor	0.623	3.136
LightGBM Regressor	**0.027**	2.946
Multi-Layer Perceptron (MLP)	1.678	1.974

We can make the following observations from the results above:

a. Linear Regression has the best generalizing power since it has the lowest Test MSE. Also, Test R2 for the model is **0.956**, which means that the model can capture **95% of variance in data**

b. MLP is underfitting as both **Train MSE** (Mean Squared Error) and **Test MSE** are high

c. It seems that both bagging (Random Forest Regressor) and boosting models (LightGBM Regressor) are over-fitted models as the difference between Train and Test MSE are relatively large than other models

Keeping these in mind, my approach for the project is going to be as follows:

1. After understanding NMR Simulation Code, I would start by preparing the dataset that will be used for analysis. In case I need to predict the parameters α, ξ and Γ from the data to be used for classification later, I will make a regressor that **does not overfit and underfit**.
2. Next, I will model a variety of classifiers from basic like Logistic Regression, to ensemble methods like Bagging and Boosting classifiers.
3. Next, I will try to achieve best model performances with tuning hyperparameters **while ensuring that the models do not overfit**.

4. Next, I will extract the most important features of the dataset using a number of techniques as follows. I intend to complete this by mid-evaluation.
    a. **Feature Importances using ensemble classifiers:** Those features which are responsible for a split in decision tree
    b. **Permutation Feature Importance:** Shuffling values within a column and comparing accuracy with original case. If there is a drastic change in accuracy, this would mean that the feature is important
    c. **SHAP (Shapley Additive Explanations):** Explaining predictions by computing contribution of each feature in making predictions.
5. Next, I will try neural networks for making predictions of variables and compare performance with above models.
6. Finally, I will develop the optimization algorithm. I think the approach to be taken can be similar to **Grid Search CV** for **exhaustive search** and finding global extrema or **Randomized Search CV** for **randomized search** which is **computationally faster** but is **not exhaustive**.
7. Final Presentation + Documentation + Submission of Code Base

RELATED WORK

SGAN: https://towardsdatascience.com/semi-supervised-learning-with-gans-9f3cb128c5e

Isomap: https://blog.paperspace.com/dimension-reduction-with-isomap/

Permutation Importance: https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html

SHAP: https://christophm.github.io/interpretable-ml-book/shap.html