Teaching Assistants Demand Prediction

Andrew Lau

BSc Computer Science

Submission Date: 26th April 2024

Supervisor: Sarah Sanders

*This report is submitted as part requirement for the BSc Degree in Computer Science at UCL. It is substantially the result of my own work except where explicitly indicated in the text.*

*The report may be freely copied and distributed provided the source is explicitly acknowledged.*

Abstract

The recruitment of Postgraduate Teaching Assistants (PGTAs) around University College London (UCL) Computer Science Department has been an ambiguous process with no clear metrics that can predict demand accurately. Despite the obvious answers of number of students, workload of the module and the difficulty of the syllabus, there is no linear relation between these metrics and the demand of PGTAs after plotting graphs to look at the results. Experimenting with AI models and different statistical tool has also proved to be inaccurate due to the unpredictability nature of modules. Hence, this project research deeper into what creates demand for PGTAs and how it can be better predicted for future references.

**Contents List**

1. Introduction
   1. Aims & Goals
   2. Deliverables
2. Background Information
3. Requirements and Analysis
4. Design and Implementation
5. Results Evaluation
6. Conclusion

**1 Introduction**

Problem Statement

At University College London (UCL) Computer Science Department, the recruitment of Postgraduate Teaching Assistants (PGTAs) is done manually through a recruitment process filled with ambiguity around the demand for PGTAs. For modules that are new, there isn’t a tried-and-tested method or official guideline on how to measure demand for PGTAs and it is usually done based on a review of the module content and intuition. This often leads to inaccurate estimation of workforce needed and inefficient allocation of resources within the department. The research is initially done around features that directly affect the demand of PGTAs such as number of students enrolled, coursework to exam ratio, module difficulty, etc. However, due to inconsistent datasets and unpredictability, further research is done in the field of AI prediction models and Natural Language Processing to analyse the module content and structure in more detail.

**1.1 Aims & Goals**

The project aims to better understand the demand PGTAs through analytical tools that drives data-driven decisions in the recruitment process. The analytical tool hopes to bring new insights into PGTA recruitment to realise its potential for ensuring that resources are better allocated to improve the learning experiences of students at UCL. Here are the aims and goals that should be achieved by the end of the project:

* Learn about the problems behind PGTA recruitment.
* Determine the factors that affects the demand of PGTA by researching into the recruitment process.
* Process data from module information and past records to create a combined dataset for training machine learning models.
* Accurately predict the necessary number of PGTAs to hire for a new and unfamiliar module.

**1.2 Deliverables**

The project submission would include the source files a documented and functional Dash app including:

1. Dashboards and analytical tools built using Plotly, including different types of graphs that provide a visual representation of the dataset:
   1. PGTAs Requested vs PGTAs Recruited Graph
      * Provides insights into which modules have inaccurate estimation of PGTA demand and its scale
   2. Students Enrolled vs PGTAs Recruited Graph
      * To determine whether the number of students enrolled affect the number of PGTAs recruited for the academic year 2022-2023 (only data for one year can be found)
   3. Exam-Coursework Ratio vs PGTAs Recruited Graph
      * To determine whether the weight of coursework and exams affect the number of PGTAs recruited
   4. Module Delivery Code vs PGTAs Recruited Graph
      * To determine whether module delivery code (modules taken by students in different years of study. In other words, difficulty of the module) affect the number of PGTAs recruited
2. Machine learning models drawn from Scikit-learn libraries.
   1. Several different machine learning models has been trained and tested with different variants. The one that was chosen as the final prediction model is a ridge regression model with cross-validation. A more detailed analysis can be found at section \_\_\_

**2 Background Information**

2.1 Background Research

The project kicked-off with a brief of the problem statement given by my project supervisor, explaining the problem with ambiguity around the hiring of PGTAs.

Data Sources

* Write about how data are given, and where it has been collected, the ethics behind retaining data and accessing them
* The data are sourced by my supervisor in the format of excel sheets. The data are collected and recorded manually when students signed up to be a teaching assistant.
* The data provided did not contain any sensitive/personal information (eg. Students’ email, phone number, date of birth, etc.) and hence, the ethical report is not needed to address the ethical conducts within this project

Technology Stack

Selecting the technology stack for a project can be dependent on specific requirements of the project, such as the complexity of data, real-time processing needs, interactivity level, and the intended user’s technical proficiency. My supervisor will be the intender user after completion of this project and hence, the technology stack chosen must suit her technical proficiency and preferences. Hence, the main requirements of the technology stack include ease of use, Python based and freely available. A few different platforms that meet the main requirements was considered for the implementation of this project. Their strengths and weaknesses in different areas of the project are considered carefully to enable a successful implementation of this project. Here are the platforms considered and their evaluations:

Grafana

* Best for real-time monitoring and observability
* Specialises in complex time-series data visualisation
* Offers extensive customisation options through a wide range of plugins

Streamlit

* Ideal for quickly creating data apps, reports, and dashboards
* More suitable with static data with smaller datasets
* Easy to set up, configure and deploy with minimal infrastructure requirements
* Limited in terms of real-time capabilities and complex interactive features

Plotly (Dash)

* Focuses more on data visualisation
* Has better graphing tools
* Better suited for advanced interactivity

Jupyter Notebooks

* Interactive computing environment that supports live code, equations, visualizations, and narrative text.
* Highly popular in the data science community for exploratory data analysis and prototyping.
* Integrates well with Python libraries like Matplotlib, Seaborn, Plotly for visualization.

In conclusion, Streamlit is chosen for its quick and easy setup along with Dash for its extensive data visualisation tools to enable easy plotting of graphs. Jupyter Notebooks are used for its support in data analysis and suitability in training machine learning models. It also works and integrates well with Plotly.

**3 Requirements and Analysis**

**4 Design and Implementation**

Data Processing

* Functions written to process/clean the data to enable them to be processed (handle missing values, etc.)

Graphs

Demo Graph

* Get familiar with the syntax and testing out the different callback to get an idea of the capabilities of Plotly library.

PGTAs Requested vs Recruited Graph and Module History Graph

* Get an idea of the modules that have an underestimated or overestimated PGTA demand and its scale.

Variables vs PGTAs Recruited Graphs

* Get an idea of how different variables relate to the recruit of PGTAs
* Found out that none of the variables relate linearly with the PGTAs’ demand

Generalised Additive Model

ML Models

Linear regression

* Advantages and disadvantages

Ridge regression

* Advantages and disadvantages

Prediction Prompt

User input -> machine learning model -> output

**5 Results Evaluation**

Cross-validation (CV)

* A model usually creates a prediction function from the same set of data, which might lead to a bias towards that dataset. This might produce inaccuracies when prompted with un-seen data (a situation called overfitting).
* Cross-validation solves this problem by using the K-fold strategy. The dataset is split into k smaller datasets and the model is trained and tested k times on k-1 folds and 1-fold respectively. E.g. a CV with 5 folds running on a dataset of 100 data will produce 20 data per fold.
* Cross-validation is computationally done using the cross\_val\_score helper function on the dataset by splitting the data, fitting a model, and computing the score for k number of consecutive times (with different splits each time). The Root Mean Squared Error (RMSE) is used as a performance metric.
* The results of each of the selected number of folds are documented below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| K - folds | Linear Regression | | Ridge Regression | |
| Mean RSME | Standard Deviation | Mean RSME | Standard Deviation |
| 5 | 101.66 | 12.34 | 96.09 | 11.00 |
| 10 | 101.00 | 24.29 | 93.96 | 22.67 |
| 15 | 99.06 | 25.88 | 92.84 | 28.69 |
| 20 | 99.10 | 30.38 | 92.24 | 30.91 |

* Low Mean RMSE: This indicates that the model, on average, has a low prediction error, directly relates to the model's predictive power.
* Low Standard Deviation: This suggests that the model's performance is consistent across different subsets of the dataset, crucial for ensuring that the model is reliable and not just performing well on specific types of data.
* Considering the trade-offs between accuracy and consistency, a model with a balanced mean RMSE and standard deviation is chosen for the context of predicting PGTAs to hire. Hence, the optimal model is the ridge regression model with 5-fold cross-validation.

**6 Conclusion**