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

At University College London (UCL) Computer Science Department, the recruitment of Postgraduate Teaching Assistants (PGTAs) is done manually through a recruitment process with limited data to grasp the demand, especially for new modules, leading to inaccurate estimation of workforce needed. 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.

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

**Contents**

1. Introduction

• What problem are you solving? • Why are you solving it? • How does this relate to other work in this area? • What work does it build on? • What is the scope of your work? • What is included in the scope and what is outside the scope?

Project Plan

Title:

* Understanding Demand of PGTAs

Aim:

* To gain insights into the demand of PGTAs for undergraduate courses and better manage PGTA allocations and applications

Objective:

* Review history of PGTAs allocations
* Develop dashboards to present high level overview of PGTAs
* Document findings from datasheets of PGTAs and evaluate their significance

Deliverables:

* A documented and functional dash app showing insights into PGTAs with dashboards and analytical tools

Work Plan:

* Project start to mid-October: Research on frameworks and platform for development.
* Mid-October to mid-December: System design, simple prototypes, system implementation.
* January to February: System implementation, evaluation and testing.
* February to early March: Research on new methods and further analysis.
* Mid-February to end of March: Wrap up and final report.

Ethics review:

* Ethics approval is not required as sensitive data is not collected/accessed in any part of the project

Interim Report:

The progress of the project is documented from the first week to this point in time:

* Week 1: Discussed the context of the project, aims and objectives. Researched on the problem area and solutions around it.
* Week 2: Technology review with considerations of ease of use, freely available, Python-oriented.
* Week 3: Had an overview of how the dataset looks and was briefed on its different components. Discussed ideas of data storing and analysis needed to be performed.
* Week 4: Getting familiarised with selected platforms and understanding the syntax. Plotted a sample graph to get comfortable with developing within the framework.
* Week 5: Plotted a simple graph showing PGTA requested and PGTA recruited, showing modules where demand exceeds expectation.

A link to the project is attached [here](https://github.com/Andrew0000000/Final-Year-Project).

Further Development plan:

* Research on useful statistics to capture/analyse and create plots for those.
* Design and develop UI of the app, allowing easy navigation and viewing.
* Look into ML models if necessary, conducting advanced analysis on datasets of PGTAs to assist in making better-informed decisions.

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.

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

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.

Gantt Chart

**Introduction**