# Introduction to Artificial Intelligence

## Problem: “Predicting software engineer salaries using regression and natural language processing and finding any significant patterns”

## Group members

UG-Rai, Neeraj [Neeraj.Rai@city.ac.uk](mailto:Neeraj.Rai@city.ac.uk" \t "_blank) [rai-n (github.com)](https://github.com/rai-n)

UG-Kolbuszewski, Krystian [Krystian.Kolbuszewski@city.ac.uk](mailto:Krystian.Kolbuszewski@city.ac.uk" \t "_blank) [krystiank-city (github.com)](https://github.com/krystiank-city)

UG-Malynovskyi, Maksym [maksym.malynovskyi@city.ac.uk](mailto:maksym.malynovskyi@city.ac.uk" \t "_blank) [Malyrion (github.com)](https://github.com/Malyrion)

## What is your dataset, problem domain?

Option A: We use an existing dataset such as:

* Example from Kaggle

Option B: We will scrap data from a site and clean it

We created our own dataset. This includes many job descriptions, mostly with salaries and some without. We can use the data without salaries as part of the testing to estimate salaries by predicting a salary for the job adverts which do not have salaries.

We scrapped the following websites using Python’s Beautiful Soup and Urlib packages. The code had be adjusted to correctly scrap data on the different websites.

|  |  |
| --- | --- |
| Totaljobs [Software Engineer Jobs in December 2021, Careers & Recruitment - totaljobs](https://www.totaljobs.com/jobs/software-engineer?s=header) | Graphical user interface, text, application  Description automatically generated |
| Glassdoor  [Software developer Jobs | Glassdoor](https://www.glassdoor.co.uk/Job/software-developer-jobs-SRCH_KO0,18.htm) | Graphical user interface, text, application, email  Description automatically generated |
| Indeed  [Job Search | Indeed](https://uk.indeed.com/?r=us) | Graphical user interface, text, application, Word  Description automatically generated |
|  |  |

For each job record, we saved the job description along with the information related to the job such as working hours, location, job title and who the job was advertised by.

We used Python’s Spacy package to see if the job description was in English. If so, we scrapped the data, we performed some checks on each of the job records against a range check to see if there were any anomalies, and then uploaded it to a PostreSQL table in AWS

## Is your model classification or regression?

The output of the problem will be a continuous value, that is the salary and any other numeric patterns we identify.

We will use the job adverts with salary as training data to predict salaries for those that do not.

## Did you have any missing, corrupt or misleading data? If so, how did you cope it?

* Yes
* One of the initial ideas to clean our data was to remove any null values from our dataset, or replace them with an average number. If we remove the nulls however our dataset would become significantly small. In our case we have 500 rows of data. Which can be considered a small dataset. And 337 rows where the Salary column is null.
* Text

  Description automatically generated
* However we have decided to use Natural Language processing which can use null entries for the training.
* Once we have cleaned our data, we would need to build our model and run it. If our model would not perform well, we need to see if we are undefining or overfitting our data. Therefore, depending on which model we would use we need to configure the hyperparameters to improve our model accuracy. Also, we would need to check if there any rows that would be giving noise to our model and investigate possibility to substitute it or remove it completely.

## Have you omitted some data? If so, why?

* Yes
* We have decided to create our own columns with some Boolean values to represent features such as (Remote Working) (Easy to Apply) and some others. There is also a possibility to create our own tokenisation of Job Titles e.g., Software Engineer =1, Data Scientist=2 and so on. This way our model doesn’t need to work with strings and go through Natural Language Processing.

## Did you apply techniques to understand your dataset?

* We run a confusion matrix to

## What models did you use?

Salary prediction:

* XGBoost
* Linear Regression
* Cluster analysis
* SVR
* Random Forest Regressor

## How did you encode the input variables?

## What are the criteria for selecting model performance evaluation tools?

## What were your outputs?

## Did you have any problems or difficulties working with the dataset?

Methodology:

Evaluation:

Conslusion:

Referencing: