Salary Estimation App Development Report

Introduction

This report introduces a newly developed application by Andreas Wagner, Boris Tsonkov, and Yue Wang. The application was developed for job posters and job seekers to rectify a common problem encountered during job posting: an improved way of making estimates on salaries with a better degree of accuracy. The application leverages a Machine Learning (ML) model to give end-users an estimate of their probable salaries, leading towards better, more informative career choices.

Objective

Follow the steps provided in the project description here.

Data Sources and Collection

This application relies on job boards with public API, treating each of them as an essential source of salary information and job descriptions. Thus, the following sources have been selected:

Samples

- **arbeitnow.com**:

https://github.com/mygitbob/feb24_project_job-market/blob/boris/data/processed/arbeitnow_job_entry_0.json

- **jobicy.com**:

https://github.com/mygitbob/feb24_project_job-market/blob/andreas/data/raw/raw_sample_data/jobicy_raw_pretty.json

- **jooble.org**:

https://github.com/mygitbob/feb24_project_job-market/blob/andreas/data/processed/raw_sample_data/joodle_job_entry_0.json

- **okjob.io**:

https://github.com/mygitbob/feb24_project_job-market/blob/andreas/data/processed/raw_sample_data/okjob_raw.json

- **reed.co.uk**:

https://github.com/mygitbob/feb24_project_job-market/blob/andreas/data/processed/raw_sample_data/reed_raw.json

- **themuse.com**:

https://github.com/mygitbob/feb24_project_job-market/blob/main/data/processed/muse_job_entry_0.json

- **adzuna.com**:

https://github.com/mygitbob/feb24_project_job-market/blob/andreas/data/processed/raw_sample_data/adzuna_job_entry_0.json

Data Processing and Machine Learning Model

Data Transformation

Normalization of Job Categories/Titles

Identifying synonyms and variations of job categories in the extracted data. For example, “Software Engineer” and “Software Developer” could be considered synonyms.

Development of a list of standard designations for job categories to be used in by the salary calculator.

Transformation of all identified synonyms and variations into their corresponding standard designations.

Standardization of Salary Calculation

All salaries must be in a consistent unit, either monthly or annually.

If some salaries are in different units, they should be converted to the desired unit.

Normalization of Locations

Identifying synonyms, variations, and alternative spellings for locations in the extracted data.

Development of a list of standard designations or codes for locations to be used by the salary calculator.

Data Enrichment (optiopnal, if we have enough time & data)

Addition of additional features or information, which could be relevant for salary prediction, such as years of experience, qualifications, company size, etc.

Data Cleaning and Preprocessing

First, before it goes through any of the cleaning and preprocessing stages, the following steps should executed collected to prepare the data for training the ML model:

- Removing duplicate entries
- Handling missing values
- Normalizing job titles and company names - Extracting and standardizing salary information

Optional Preprocessing Steps: These may be utilized to provide more detail into the preprocess method, particularly techniques like text normalization and feature engineering, which allow improvement in the performance of the models.

Machine Learning Model Development

The core of the app is the development of an advanced Machine Learning model that can predict salaries based on job postings. We will cover this in Step 3.

Model Training

The training process will involve:

- Splitting the data into training and validation sets
- Selecting appropriate features for salary prediction
- Choosing and tuning a suitable ML algorithm (e.g., regression, random forest, neural networks)

Prediction and Validation

The model will then be validated by use of the datasets from Adzuna.com and Themuse.com, for example, to measure its accuracy and reliability. This is going to give a high degree of confidence that the app will produce useful results across a wide range of jobs and market conditions.

Final Delivery

The product to address the business problem should be formulated as an API that consumes a machine learning model.

This API should provide endpoints for sending input data and receiving predictions.