Name: Dafydd James Thomas
Student ID: 249004160

Title

Forecasting AI Skill Demand using Innovation Signals from Open-Source Software Material

Background

The demand for AI-related skills is evolving rapidly in response to technological innovation. It is important that such changes in demand are anticipated, so that resources can be allocated appropriately.

This project essentially concerns itself with the following research question: Can AI-related innovation signal data improve the accuracy of the relevant skill demand forecasts derived from job posting trends?

Demonstrated for a few AI-related skills, this project develops a system for forecasting such skill demand (as represented by job advertisements) by analysing innovation signals from Open-Source Software (OSS) contributions. As a broader goal, the project aims to develop a pipeline that integrates evidence-based & modular components, thereby composing a methodology that is extensible (e.g. with other innovation signals, evolving skill domains, etc.).

Practically, the project therefore entails gathering Job Advertisements and OSS data, relevant to 2-3 AI-related skills. Thereafter, these data are processed using established NLP and topic modelling techniques, then aligned temporally to explore trends and time-lagged correlations. Finally, forecasting models are developed to predict skill demand based on only job ads (baseline) and another that adds innovation signals as predictors. An evaluation follows to dictate whether innovation-informed models outperform a baseline forecast that is based solely on job posting data.

Through this approach, the study aims to contribute a novel but grounded framework for enhancing labour market intelligence and anticipatory workforce planning. Importantly, this framework in its modular setting is highly adaptable, and is readily applicable to data created in real-time.

Aims and Objectives

Aims:

To develop and evaluate an AI-related skill demand forecasting methodology, integrating innovation signals (from Open-Source Software in particular) and using job ads as a proxy for skill demand.

Objectives:

1. Identify and collect high-frequency innovation data from open-source repositories (GitHub).
2. Extract and represent skill-related signals over time from both innovation and job ad corpora.
3. Build baseline time-series forecasting models based solely on job ad trends for selected AI-related skills.
4. Integrate innovation-signal features in a richer forecasting model.
5. Evaluate forecasting performance with and without innovation signals to assess added value.
6. Discuss implications for skill gap anticipation.

Tasks

With Week 0 starting Monday 16th June (Friday of Week 11 is 5th September)

Literature Review & Dataset Finalisation (weeks 0-1)

* Select AI-related target skills/topics.
* Finalise datasets for job ads and innovation sources (Lightcast, ADZURA or LinkedIn datasets; GitHub).
* Review and select tools for text extraction (e.g., BERTopic, TF-IDF, SciBERT).

Data Processing & Representation (Weeks 2-4)

* Implement text extraction and representation pipeline.
* Generate time-series representations of skills in job ads and OSS signals.
* Align temporal indices and explore correlation patterns.

Forecasting Model Development (Weeks 5-8)

* Build and evaluate baseline forecasts using job ads (Either traditional (ARIMA) or Deep-Learning (LSTM))
* Develop augmented models with innovation features (e.g., lagged innovation trends).
* Run backtesting evaluations, comparing to actual values for skill mentions in job ads.

Evaluation & Comparative Analysis (Weeks 9-10)

* Evaluate model performance (NRMSE/RMSE, MAPE, directional accuracy).
* Compare forecasting performance with and without innovation signals.

Finally, each element of the end-to-end process is to be evidenced, reasoned & critically evaluated in the report, as well as the relevant visualisations & findings relating to the forecasting models.

Full write-up is to be incrementally produced along the 12 weeks, with finalizations in week 11.

Deliverables

* Fully modular code pipeline (Python/Jupyter, documented).
* Forecasting performance reports for baseline and augmented models **(Measurable Test of Hypothesis)**
* Dataset repository (processed and raw).
* Visual timeline and trends of skill mentions across sources.

The success & validity of this project hinges on the capacity for the developed innovation-signal features to *improve* existing forecasts. One strategy for measuring the success of a forecast is *backtesting*, wherein past-data is split into train/test segments: e.g. Pre-2023 data is used to inform the forecasts, whose predictions are evaluated against the *actual* outcomes in 2023-2025. In doing so, we can directly test how a forecasting solution *would* have performed previously.

Resources

- Access to job ad datasets (Lightcast, LinkedIn/Indeed, Kaggle or ADZURA; paid/free access available via APIs or downloadable/scrapeable)
- (Optional) Access to O*NET data for occupation-skill mapping (open & downloadable)
- Access to GitHub data (can be web-scraped or downloaded via APIs).
- Python stack: pandas, scikit-learn, statsmodels, spaCy/transformers (for NLP), matplotlib/seaborn.
- Literature access (ArXiv, Elsevier, IEEE)