

AI-powered Job market insights

Problems of the future



September 28, 2024

Beka Daniel Bonsa

Contents

[One-pager 2](#_Toc178450430)

[Intent 2](#_Toc178450431)

[Desired outcome 2](#_Toc178450432)

[Deliverable 2](#_Toc178450433)

[Constraints 2](#_Toc178450434)

[1. Overview 3](#_Toc178450435)

[**2.** Motivation 3](#_Toc178450436)

[3. Success metrics 3](#_Toc178450437)

[4. Requirements & Constraints 4](#_Toc178450438)

[Requirements 4](#_Toc178450439)

[Constraints 4](#_Toc178450440)

[4.1 What's in-scope & out-of-scope? 4](#_Toc178450441)

[In-scope 4](#_Toc178450442)

[Out-of-scope 4](#_Toc178450443)

[5. Methodology 5](#_Toc178450444)

[5.1. Problem statement 5](#_Toc178450445)

[5.2. Data 5](#_Toc178450446)

[5.3. Techniques 6](#_Toc178450447)

[5.4. Experimentation & Validation 7](#_Toc178450448)

[5.5. Human-in-the-loop 7](#_Toc178450449)

[6. Implementation 8](#_Toc178450450)

[6.1. High-level design 8](#_Toc178450451)

[6.2. Infra 8](#_Toc178450452)

[6.3. Performance (Throughput, Latency) 8](#_Toc178450453)

[6.4. Security 8](#_Toc178450454)

[6.5. Data privacy 9](#_Toc178450455)

[6.6. Monitoring & Alarms 9](#_Toc178450456)

[6.7. Cost 9](#_Toc178450457)

[Estimated monthly costs 9](#_Toc178450458)

[6.8. Integration points 10](#_Toc178450459)

[6.9. Risks & Uncertainties 10](#_Toc178450460)

[7. Appendix 10](#_Toc178450461)

[7.1 References 10](#_Toc178450462)

# One-pager

## Intent

At the dawn of the AI era, we find ourselves navigating a rapidly evolving landscape of technological innovation and transformation. With this in mind, the aim of this this analysis is to identify current trends in the AI-job market and to study links between unemployment and the implementation of AI in different companies. The goal of this is to help raise awareness about how AI will affect the job market and the economy. Doing so, might give governments and individuals a head start in preparing for what is to come.

## Desired outcome

The desired outcome is to create a model that predicts the AI-powered job market for the coming years. Based on this model, governments could improve AI-regulation laws and prepare for a growing number of layoffs. This might prove challenging as the field of AI is experiencing a rapid pace of development, with constant advancements that could significantly impact job roles and industry demands.

## Deliverable

The key deliverable will be a comprehensive report outlining the findings from the analysis of current AI-job market trends and the links between AI implementation and unemployment rates. This report will include a predictive AI job market model, policy recommendations and an economic impact assessment. Together, these deliverables will create a comprehensive framework for addressing AI’s role in the new digital era.

## Constraints

Perhaps the most important constraint present for the model is the uncertainty of how AI will evolve in the coming years. For the past few years, the field of AI has been evolving rapidly. But this does not mean the growth in this field will not face challenges or slow down due to factors such as regulatory hurdles, ethical concerns, or the limitations of current technologies.

Additionally, the accuracy of the predictive model will heavily depend on the quality of data provided. Which is why it is vital to represent most companies in almost every sector in hopes of sampling data with less bias.

# 1. Overview

This document aims to develop a predictive model for understanding how AI will impact the job market in the coming years. The problem arises from the rapid adoption of AI technologies, which could lead to significant job displacement, regulatory challenges, and economic shifts. The proposed solution is to analyze current AI job trends and provide strategic recommendations to help governments and employees prepare for these changes. The desired outcome is a comprehensive forecast that guides policymaking, workforce adaptation, and economic planning to mitigate the negative effects of AI while maximizing its potential benefits.

# **2.** Motivation

As AI-driven automation increasingly replaces or alters traditional roles, there is growing concern about mass layoffs and job displacement among workers. Most of this is reminiscent to the Industrial Revolution. During that period, there was widespread fear and uncertainty as machines began replacing factory workers leading to job displacements and a profound restructuring of the economy. We find ourselves in a similar situation today and preparing for these problems will be key for prospering in this new era.

# 3. Success metrics

Usually framed as business goals, such as increased customer engagement (e.g., CTR, DAU), revenue, or reduced cost.

The success of this project will be measure by several outcomes. The first being accuracy of the predictive model. The productive model should demonstrate high predictive capabilities in forecasting employment trends and job displacements as new data becomes available.

Another outcome that can be measured is the extent to which the model’s policy recommendations based on the model are adopted by governments and regulatory bodies over time.

Just like the industrial revolution, the new age of AI could provide a massive boost for the economy. Living standards and general quality of life might reach levels never seen before. It also has the potential to do quite the opposite. Therefore, we should use success metrics that not only track economic growth for businesses but also the overall well-being and living standards for the people involved.

Finally, an important success metric for the general workforce is the extent of workforce adaptation initiatives by businesses and governments. This could be the number of new jobs created, the number of reskilling programs implemented, or the percentage of workers successfully transitioning into AI-related roles, reflecting how effectively businesses and governments are supporting workforce adaptation.

# 4. Requirements & Constraints

### Requirements

* Accurate job market forecasting: The model must provide reliable/reasonable predictions based on real world data from various job sectors. This includes which roles are likely to be displaced.
* Reskilling recommendations: Based on the model, it should be possible to deliver actionable reskilling recommendations for the workforce.
* Policy recommendations: The system should help generate policy recommendations for governments to adapt AI regulations.
* Data visualization: As a complimentary requirement for the previous three, it should also be possible to visualize the model’s output for non-technical users which can range from policy makers, business leaders and the average worker.

### Constraints

* Data availability: The system must work within the limitations of available datasets and adjust predictions as new data becomes available. Since the dataset has a lot of features with many different values, a huge amount of data is needed for properly training the model.
* Variability in regulations: Different regions may have different AI regulations and labor laws in place, which can impact the adoption rate of AI and its effects on the workforce.
* Technological Change: Given the rapid advancements in AI, the system must be flexible enough to incorporate future technological developments that could affect the job market.

# 4.1 What's in-scope & out-of-scope?

### In-scope

* Collection of data from businesses and readily available public datasets.
* Reskilling recommendations
* Policy recommendations
* Industry-specific analysis: The project will analyze how AI affects specific industries, providing a breakdown of the economic and employment impacts within each sector.

### Out-of-scope

* Close to real-time job market monitoring
* Global AI regulation framework: Since AI regulations varies widely across countries.
* Company-specific analysis: The project will not offer company-specific analysis of how AI will have an impact on a company.

# 5. Methodology

## 5.1. Problem statement

The problem of predicting the impact of AI on the job market can be framed as a supervised machine learning problem focusing on mostly on classification and regression. Classification can be used with the task of predicting which job roles are at high risk of automation, while regression can be used with the task of predicting the rate at which AI adoption will affect the job market over time.

Recommendations for reskilling programs will be generated based on these results framed as a recommendation system that maps at-risk jobs to emerging job opportunities requiring similar skill sets. The core problem, therefore, is to predict and classify how AI will influence employment across sectors and forecast the resulting job market dynamics, allowing businesses and policymakers to take preemptive action.

## 5.2. Data

What data will you use to train your model? What input data is needed during serving?

The primary part of the data that will be used for training the model is ["AI-Powered Job Market Insights"](https://www.kaggle.com/datasets/uom190346a/ai-powered-job-market-insights).

The features of the data as described by the owner of the data, are the following:

Job\_Title:

Description: The title of the job role.

Type: Categorical

Example Values: "Data Scientist", "Software Engineer", "HR Manager"

Industry:

Description: The industry in which the job is located.

Type: Categorical

Example Values: "Technology", "Healthcare", "Finance"

Company\_Size:

Description: The size of the company offering the job.

Type: Categorical

Categories: "Small", "Medium", "Large"

Location:

Description: The geographic location of the job.

Type: Categorical

Example Values: "New York", "San Francisco", "London"

AI\_Adoption\_Level:

Description: The extent to which the company has adopted AI in its operations.

Type: Categorical

Categories: "Low", "Medium", "High"

Automation\_Risk:

Description: The estimated risk that the job could be automated within the next 10 years.

Type: Categorical

Categories: "Low", "Medium", "High"

Required\_Skills:

Description: The key skills required for the job role.

Type: Categorical

Example Values: "Python", "Data Analysis", "Project Management"

Salary\_USD:

Description: The annual salary offered for the job in USD.

Type: Numerical

Value Range: $30,000 - $200,000

Remote\_Friendly:

Description: Indicates whether the job can be performed remotely.

Type: Categorical

Categories: "Yes", "No"

Job\_Growth\_Projection:

Description: The projected growth or decline of the job role over the next five years.

Type: Categorical

Categories: "Decline", "Stable", "Growth"

When serving the data, all these features will be made use of to generate predictions.

## 5.3. Techniques

For classification purposes, a Random Forest model will be used for classifying if individuals are at risk of losing their jobs in the future by considering the features mentioned earlier such as Job\_Title, Industry\_Size etc. Similarly it will also be used to classify which industry sectors are more likely to have worker lay-offs using these features.

A Random Forest regression model will be used for the purpose of predicting salary based on the features. For the regressor, categorical variables will be converted into numerical values.

If there are missing values in the dataset, K-Nearest Neighbors imputation will be used if necessary.

Since we are using a tree based regressor, in this case random forest regressor, we do not need to implement one-hot encoding or label encoding on features like Automation Risk etc.since each of the categorical values are considered equally different.

To ensure the model will not be skewed, outliers will either be excluded or capped using statistical tools like Z-score.

For the purposes of training the model the dataset will be split while training using k-fold cross-validation. This gives us a good basis on evaluating our model on new data.

## 5.4. Experimentation & Validation

Using k-fold cross-validation ensures that the model generalizes well to unseen data, and the average estimation provides close to realistic estimate of its performance.

For regression tasks RMSE (root mean squared error) will be used for penalizing wrong results.

For classification tasks, precision, recall, and f1-score will be used to assess results. A confusion matrix will also be used to analyze the models predictions.

If there are governments or businesses applying some of the recommendations that resulted from the model’s predictions, the aftermath of the recommendation system can be documented assess how well the recommendations translate to real life.

A customer-based assignment will be used for treatment and control where different companies and organizations can be assigned treatment and control groups. The metrics for the A/B testing will be the percentage of workers who successfully transition to new roles and the number of employees enrolling in reskilling programs based on the model’s suggestions.

## 5.5. Human-in-the-loop

How will you incorporate human intervention into your ML system (e.g., product/customer exclusion lists)?

Experts from specific industries or job market analysts can review the model’s predictions and provide insights on the model. Since the system will be used to provide regulatory recommendations, human experts will periodically review the model’s suggestions before they are shared with decision-makers, ensuring that policy implications align with regional laws, ethical considerations, and practical implementation.

To maximize the accuracy of the model, relevant data will be gathered from each business and individual that is in the control test group.

# 6. Implementation

## 6.1. High-level design

A diagram of a company

Description automatically generated

## 6.2. Infra

The system will be hosted in a cloud-based infrastructure. This choice is driven by the need for scalability, flexibility, and the ability to handle large datasets and perform machine learning tasks efficiently. The cloud platform will provide the necessary tools for storage, model training, inference, and security, all while being cost-effective and scalable to handle varying workloads.

## 6.3. Performance (Throughput, Latency)

How will your system meet the throughput and latency requirements? Will it scale vertically or horizontally?

The system will rely primarily on horizontal scaling because it allows the system to handle unpredictable surges in demand and provide higher resilience by distributing the load across multiple resources. This allows balancing throughput when the demand for the service is high.

Latency is less of a problem as this is a system designed to run once for a given instance and output a report.

## 6.4. Security

An authentication and authorization system will be set in place to provide a secure and standardized authentication process for the businesses and individuals that are partaking in this project. There will be firewall protection to block unauthorized external access to help mitigate data-breaches.

The system must also be resilient against DDoS attacks using protection provided by the cloud providers, in addition to setting a rate limit for every IP address.

## 6.5. Data privacy

This project will operate within the boundaries set by data privacy compliance regulations such as GDPR.

* Employees and companies will have the right to access the data that the system collects and processes about them.
* Users will have the ability to correct or update inaccurate data within the system. An interface or support channel will allow users to request corrections to their personal data.
* Users will have the option to request deletion of their personal data from the system
* Before collecting or processing any personal data, the system will obtain explicit user consent. This will include consent for collecting data, processing it for job predictions, and any reskilling recommendations. Consent can be withdrawn at any time, and the system will stop processing the user’s data upon request.

In cases where full anonymization is not possible, pseudonymization techniques will be used. Personal identifiers will be replaced with pseudonyms (e.g., user IDs) so that the data can still be analyzed without directly identifying individuals.

## 6.6. Monitoring & Alarms

All system events including API requests, model inferences and user interactions will be logged to a centralized logging system.

Throughput and latency will be monitored to ensure the system is handling the API calls without degradation. One alarm is triggered when the throughput is low and at the same time the latency is high.

CPU, memory and disk utilization should be constantly monitored along with network traffic to raise alarms if the load on the infrastructure is too high.

Failed login attempts, suspicious activity and every data access will be logged and if certain criteria are met, an alarm should be triggered.

The alerting mechanism in place should be able to send emails or SMS messages to system administrators.

## 6.7. Cost

### Estimated monthly costs

|  |  |
| --- | --- |
| Consulting experts | 8000 NOK |
| Cloud infrastracture | 10000 NOK |
| Backup and recovery | 2000 NOK |
| System security | 5000 NOK |
| Data collection team | 30000 NOK |

## 6.8. Integration points

For the upstream data, the system will integrate with external job market databases (e.g., LinkedIn, Glassdoor, government job boards) to ingest data related to job roles, descriptions, salaries, and company information.

For downstream integration, a web-based dashboard will provide insights into job roles at risk of automation, salary predictions, and reskilling recommendations. It will be possible to interact with the system, upload employee data, and generate reports.

## 6.9. Risks & Uncertainties

Risks are the known unknowns; uncertainties are the unknown unknows. What worries you and you would like others to review?

The risks of using data that cannot be easily verified is that there is no guarantee on the quality of the data collected. Guaranteeing the validity of the data collected will significantly contribute to creating models that provide beneficial recommendations.

Another uncertainty when it comes to collecting data is that some, if not most, companies prefer to keep things private. This makes the data collection process even harder. A comprehensive report generated by the model requires huge amounts of data to train on, if the dataset is not vast and huge, its predictive ability will be limited.

Digital data regulation laws have become stricter over the years. A detailed look of the data collected can be shown to external entities to avoid lawsuits or shutting down the project.

# 7. Appendix

# 7.1 References

Tharmalingam, L. (2024, August 26). *AI-Powered Job Market Insights*. Hentet fra Kaggle.com: https://www.kaggle.com/datasets/uom190346a/ai-powered-job-market-insights