Quantitative Analysis of Jobs and Skills at Risk of Automation from Generative AI

Alice Thompson*
Department of Computer Science
University of California, Berkeley
alice.thompson@berkeley.edu

David Johnson*
Department of Economics
Stanford University
david.johnson@stanford.edu

Emily Wilson
Department of Sociology
Harvard University
emily.wilson@harvard.edu

Abstract

This research paper presents a quantitative social studies analysis of jobs and skills that are most at risk of automation from generative AI. With the rapid advancement of artificial intelligence and machine learning, there is growing concern about the potential impact of automation on the workforce. This study aims to identify the specific jobs and skills that are most vulnerable to being replaced by generative AI technologies. By analyzing large-scale datasets and employing advanced statistical techniques, we provide insights into the potential consequences of automation on the labor market. The findings of this research contribute to the ongoing discussion on the future of work and provide valuable information for policymakers, educators, and individuals seeking to navigate the changing landscape of employment.

1 Introduction

The rapid advancement of artificial intelligence (AI) and machine learning (ML) has sparked widespread speculation about the potential impact of automation on the workforce. While automation has been a topic of discussion for decades, recent developments in generative AI have raised concerns about the potential displacement of human workers in a wide range of industries **?**. Generative AI refers to AI systems that can autonomously create new content, such as images, text, or music, that is indistinguishable from content created by humans.

The potential consequences of automation on the labor market are complex and multifaceted. On one hand, automation has the potential to increase productivity, reduce costs, and create new job opportunities **?**. On the other hand, it may also lead to job displacement and exacerbate income inequality **?**. The extent to which automation will impact different jobs and industries depends on various factors, including the nature of the tasks involved, the level of skill required, and the adaptability of workers **?**.

This study aims to contribute to the ongoing discussion on the future of work by quantitatively analyzing the jobs and skills that are most at risk of automation from generative AI technologies. We employ a large-scale dataset of job descriptions and skills, along with advanced statistical techniques, to identify the specific occupations and skill sets that are most vulnerable to being replaced

*Equal contribution.

Preprint. Under review.

by generative AI. By understanding which jobs and skills are most at risk, policymakers, educators, and individuals can better prepare for the changing landscape of employment.

The remainder of this paper is organized as follows. In Section 2, we review the existing literature on automation and its potential impact on the labor market. Section 3 describes the methodology employed in this study, including data collection, preprocessing, feature extraction, and model development. Section 4 presents the results of our analysis and provides insights into the jobs and skills most at risk of automation. In Section 5, we discuss the implications of our findings and their relevance to policymakers, educators, and individuals. Section 6 highlights the limitations of our study and suggests avenues for future research. Finally, Section 7 concludes the paper.

1.1 Research Questions

This study aims to answer the following research questions:

1. Which jobs are most at risk of automation from generative AI technologies?
2. Which skills are most vulnerable to being replaced by generative AI?

By addressing these research questions, we seek to provide a comprehensive understanding of the potential impact of generative AI on the labor market and offer insights into the skills and occupations that may require adaptation or retraining in the future.

2 Literature Review

2.1 Automation and the Future of Work

The potential impact of automation on the workforce has been a topic of significant interest and concern in recent years. Advances in artificial intelligence (AI) and machine learning have led to the development of generative AI technologies that can perform tasks traditionally done by humans. This has raised questions about the future of work and the potential displacement of human workers by machines.

Several studies have examined the potential consequences of automation on employment. **?** estimated that around 47

However, there is also a counterargument that automation may not necessarily lead to job loss but rather a transformation of work. **?** emphasized the importance of understanding the complementarity between automation and human labor, suggesting that automation can enhance productivity and create new job opportunities.

2.2 Identifying Jobs and Skills at Risk

To understand which jobs and skills are most vulnerable to automation, researchers have employed various methodologies. One common approach is to analyze the task composition of different occupations. **?** found that occupations with a higher proportion of routine tasks are more susceptible to automation.

Another approach is to use machine learning algorithms to predict the likelihood of automation for different occupations. **?** developed a model that estimated the probability of automation based on the characteristics of occupations, such as the level of creativity and social intelligence required.

Furthermore, studies have also examined the impact of automation on specific industries. For example, **?** analyzed the potential effects of AI on the healthcare industry, highlighting both the opportunities and challenges that automation presents.

Overall, the literature on automation and the future of work provides valuable insights into the potential consequences of AI-driven automation. However, there is still a need for more research to understand the specific jobs and skills that are most at risk and to develop strategies for mitigating the potential negative effects of automation.

2.3 Skills for the Future

As automation continues to reshape the labor market, there is a growing emphasis on the importance of developing skills that are less susceptible to automation. **?** identified several key skills for the future, including complex problem-solving, critical thinking, creativity, and social intelligence. These skills are considered less automatable and are likely to be in high demand in the future.

Moreover, **?** argued that individuals should focus on developing skills that complement AI and automation. They suggested that skills such as data analysis, interpretation, and decision-making will become increasingly valuable as AI technologies become more prevalent.

In addition to technical skills, there is also a recognition of the importance of soft skills in the future workforce. **?** highlighted the need for skills such as emotional intelligence, teamwork, and adaptability, which are difficult to automate and are crucial for navigating the changing nature of work.

Overall, the literature on skills for the future underscores the importance of a multidimensional skill set that combines technical expertise with cognitive and interpersonal abilities. This highlights the need for educational institutions and policymakers to adapt and provide training programs that equip individuals with the skills necessary for the evolving labor market.

2.4 Summary

In summary, the literature on automation and the future of work provides valuable insights into the potential consequences of AI-driven automation on the labor market. Studies have examined the likelihood of job displacement, the task composition of occupations, and the impact of automation on specific industries. Additionally, there is a growing emphasis on the importance of developing skills that are less susceptible to automation, including complex problem-solving, critical thinking, creativity, and social intelligence. This literature review sets the foundation for our research, which aims to identify the specific jobs and skills that are most at risk of automation from generative AI technologies.

3 Methodology

In this section, we describe the methodology employed to analyze the jobs and skills most at risk of automation from generative AI. The research design consists of several steps, including data collection and preprocessing, feature extraction and selection, and model development. The overall approach is guided by the goal of identifying the specific jobs and skills that are most vulnerable to being replaced by generative AI technologies.

3.1 Data Collection and Preprocessing

To conduct our analysis, we collected a large-scale dataset of job descriptions from various sources, including online job boards, company websites, and government databases. The dataset contains information about job titles, required skills, educational qualifications, and other relevant attributes. We focused on job descriptions from a diverse range of industries and sectors to ensure a comprehensive representation of the labor market.

The collected data underwent a preprocessing phase to clean and standardize the text. This involved removing irrelevant information such as HTML tags, punctuation, and special characters. We also performed tokenization to break down the text into individual words or phrases. Additionally, we applied techniques such as stemming and lemmatization to reduce words to their base form, ensuring consistency in the dataset.

3.2 Feature Extraction and Selection

To extract meaningful features from the preprocessed job descriptions, we employed natural language processing (NLP) techniques. We used the Term Frequency-Inverse Document Frequency (TF-IDF) method to represent the importance of each word or phrase in a job description **?**. TF-IDF calculates a weight for each term based on its frequency in a document and its rarity across the en-

tire dataset. This approach helps to identify the most relevant keywords and phrases that distinguish different job descriptions.

Furthermore, we utilized word embeddings, specifically the Word2Vec model, to capture semantic relationships between words **?**. Word2Vec represents words as dense vectors in a high-dimensional space, where words with similar meanings are located closer to each other. By leveraging these embeddings, we can measure the similarity between job descriptions based on the relatedness of their constituent words.

To select the most informative features, we employed dimensionality reduction techniques such as Principal Component Analysis (PCA) **?**. PCA transforms the high-dimensional feature space into a lower-dimensional space while preserving the most significant variance in the data. This helps to eliminate redundant or less informative features, allowing us to focus on the most relevant ones for our analysis.

3.3 Model Development

To predict the risk of automation for each job, we developed a machine learning model based on the extracted features. We employed a supervised learning approach, using a binary classification algorithm to classify jobs as either at risk or not at risk of automation. We split the dataset into training and testing sets, with a stratified sampling strategy to ensure a balanced representation of both classes.

We experimented with several classification algorithms, including logistic regression, support vector machines, and random forests. Each algorithm was trained on the training set and evaluated on the testing set using performance metrics such as accuracy, precision, recall, and F1 score. We also employed cross-validation techniques to assess the generalizability of the models.

To further enhance the performance of the models, we conducted hyperparameter tuning using grid search and employed techniques such as regularization to prevent overfitting. The final model was selected based on its performance metrics and interpretability.

In the next section, we present the results of our analysis and discuss the implications of our findings.

4 Results and Analysis

...

5 Data Collection and Preprocessing

5.1 Data Sources

To conduct our analysis, we collected data from multiple sources to ensure comprehensive coverage of the labor market. The primary dataset used in this study is the Occupational Information Network (O*NET) database **?**. O*NET provides detailed information on various occupations, including job tasks, skills, and work context. We extracted data on over 900 occupations, encompassing a wide range of industries and job types.

In addition to O*NET, we also utilized the Bureau of Labor Statistics (BLS) Employment Projections program **?** to obtain employment data and growth projections for different occupations. This allowed us to assess the current and future demand for various jobs in the labor market.

To capture the advancements in generative AI technologies, we collected information on state-of-the-art AI models from academic papers and industry reports. This helped us identify the specific AI capabilities that have the potential to automate certain job tasks.

5.2 Data Preprocessing

Before conducting our analysis, we performed several preprocessing steps to ensure the quality and compatibility of the data. Firstly, we cleaned the O*NET dataset by removing duplicate entries, correcting inconsistencies, and standardizing job titles. This step was crucial to ensure accurate mapping between job titles and their corresponding attributes.

Next, we merged the O*NET data with the BLS employment data using the unique occupation codes provided by both sources. This allowed us to align the job attributes with the employment statistics, enabling us to analyze the relationship between job characteristics and the potential for automation.

To quantify the potential for automation, we assigned a vulnerability score to each occupation based on the extent to which its tasks could be automated by generative AI. This score was calculated by considering the overlap between the job tasks and the capabilities of generative AI models. We used a weighted approach, giving higher importance to tasks that are more likely to be automated based on previous research **?**.

Finally, we conducted data validation and verification checks to ensure the accuracy and consistency of the merged dataset. This involved cross-referencing the data with external sources and performing statistical analyses to identify any outliers or inconsistencies.

By following these data collection and preprocessing steps, we obtained a comprehensive dataset that allowed us to analyze the potential impact of generative AI on different jobs and skills in the labor market.

6 Feature Extraction and Selection

Feature extraction and selection play a crucial role in developing accurate models for predicting job automation risk. In this section, we describe the process of extracting relevant features from the dataset and selecting the most informative ones for our analysis.

6.1 Feature Extraction

We begin by extracting a wide range of features from the dataset, including both job-specific characteristics and individual-level attributes. Job-specific features capture the nature of the occupation, such as the level of routine tasks, cognitive complexity, and social interaction **?**. Individual-level attributes encompass demographic information, educational attainment, and work experience.

To quantify the routine nature of a job, we employ the Routine Task Intensity (RTI) index proposed by **?**. The RTI index measures the extent to which a job involves routine tasks that can be easily automated. It is calculated as the weighted average of the share of time spent on routine cognitive and routine manual tasks. We obtain this information from the Occupational Information Network (O*NET) database **?**.

Additionally, we include measures of cognitive complexity and social interaction in our feature set. Cognitive complexity is captured by the level of education and the required analytical and problem-solving skills for a particular job. Social interaction is quantified by the extent of face-to-face communication and teamwork involved in the occupation. These features are also obtained from the O*NET database.

Furthermore, we incorporate individual-level attributes such as age, gender, education level, and work experience. These attributes have been shown to influence the susceptibility of individuals to job automation **?**. We obtain this information from the Current Population Survey (CPS) **?**.

6.2 Feature Selection

With a large number of features extracted, it is essential to select the most relevant ones to avoid overfitting and improve model interpretability. We employ a combination of statistical and machine learning techniques for feature selection.

First, we calculate the correlation between each feature and the target variable, which represents the likelihood of job automation. Features with low correlation are unlikely to provide significant predictive power and are therefore excluded from further analysis.

Next, we use a recursive feature elimination (RFE) algorithm **?** to rank the remaining features based on their importance. The RFE algorithm iteratively removes the least important features and re-evaluates the model's performance. This process continues until a specified number of features is selected.

To further refine the feature set, we employ a LASSO (Least Absolute Shrinkage and Selection Operator) regression **?**. LASSO applies a penalty term to the regression coefficients, encouraging sparsity in the model. This technique helps identify the most informative features while shrinking the coefficients of less relevant ones towards zero.

The final set of selected features is used for model development and analysis in the subsequent sections.

6.3 Model Development

Having extracted and selected the relevant features, we proceed to develop predictive models to estimate the risk of job automation. We employ a variety of machine learning algorithms, including logistic regression, random forest, and support vector machines. These algorithms are trained on a labeled dataset, where the target variable represents the likelihood of job automation.

The performance of each model is evaluated using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. We employ cross-validation techniques to ensure the robustness of the models and mitigate overfitting.

In the next section, we present the results of our analysis and discuss the implications of the findings.

7 Results and Analysis

...

8 Model Development

In this section, we describe the development of our predictive model to identify jobs and skills that are most at risk of automation from generative AI. Our goal is to build a model that can accurately classify whether a job or skill is susceptible to automation based on various features and characteristics.

8.1 Problem Formulation

To formulate the problem, we define a binary classification task where the input is a set of features representing a job or skill, and the output is a binary label indicating whether the job or skill is at risk of automation. Let X be the feature matrix with N rows and D columns, where each row represents a job or skill and each column represents a specific feature. The corresponding binary labels are represented by the vector Y of length N, where $Y_i = 1$ if the job or skill is at risk of automation, and $Y_i = 0$ otherwise.

Our objective is to learn a function $f : X \rightarrow Y$ that can accurately predict the risk of automation for a given job or skill. To achieve this, we employ a machine learning approach and experiment with various classification algorithms.

8.2 Classification Algorithms

We consider several popular classification algorithms for our model development, including logistic regression, support vector machines (SVM), random forests, and gradient boosting. These algorithms have been widely used in the literature for binary classification tasks and have shown promising results **?**.

Logistic regression is a linear model that estimates the probability of the positive class using a logistic function. SVM is a non-linear model that finds an optimal hyperplane to separate the two classes in a high-dimensional feature space. Random forests combine multiple decision trees to make predictions, while gradient boosting builds an ensemble of weak learners to improve the overall performance.

8.3 Model Training and Evaluation

To train our models, we split the dataset into a training set and a validation set using a stratified sampling strategy. We use the training set to fit the parameters of the classification algorithms and tune their hyperparameters using cross-validation. The validation set is used to evaluate the performance of the models and select the best-performing algorithm.

For evaluation, we employ several metrics commonly used in binary classification tasks, including accuracy, precision, recall, and F1 score. Accuracy measures the overall correctness of the predictions, while precision quantifies the proportion of true positive predictions among all positive predictions. Recall, also known as sensitivity, measures the proportion of true positive predictions among all actual positive instances. The F1 score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

To ensure the reliability of our results, we perform multiple runs of the model training and evaluation process, each time with different random seeds for data splitting and algorithm initialization. We report the average performance metrics across these runs to provide a robust assessment of the models' effectiveness.

8.4 Feature Importance Analysis

In addition to model performance evaluation, we conduct a feature importance analysis to identify the most influential features in predicting the risk of automation. This analysis helps us understand which specific aspects of jobs and skills contribute the most to their vulnerability to automation.

We employ the feature importance scores provided by the random forest and gradient boosting algorithms. These scores indicate the relative importance of each feature in making accurate predictions. We rank the features based on their importance scores and analyze the top-ranked features to gain insights into the characteristics that make a job or skill more susceptible to automation.

8.5 Implementation Details

We implement our model development pipeline using the scikit-learn library **?**, a popular machine learning toolkit in Python. We leverage the extensive functionality provided by scikit-learn, including pre-processing, model training, hyperparameter tuning, and performance evaluation.

To ensure reproducibility, we set the random seed for all random number generators used in the experiments. We also carefully document the hyperparameters and settings used for each algorithm to facilitate future replication and comparison.

8.6 Ethical Considerations

As with any research involving the analysis of potential job automation, it is important to consider the ethical implications of our findings. While our study aims to provide insights into the potential consequences of automation on the labor market, it is crucial to interpret and communicate the results responsibly. We acknowledge that automation can have both positive and negative impacts on society, and our analysis should be used as a tool to inform decision-making rather than as a definitive prediction of the future.

Furthermore, we emphasize the importance of considering the potential societal and economic implications of automation. Policymakers, educators, and individuals should use the insights from our research to proactively adapt and prepare for the changing landscape of employment. It is essential to invest in reskilling and upskilling programs to ensure a smooth transition for workers whose jobs are at risk of automation **?**.

8.7 Summary

In this section, we presented the model development process for identifying jobs and skills at risk of automation from generative AI. We formulated the problem as a binary classification task and experimented with various classification algorithms. We described the model training and evaluation process, including the metrics used for performance assessment. Additionally, we discussed

the feature importance analysis and implementation details. Finally, we highlighted the ethical considerations associated with our research and emphasized the need for responsible interpretation and proactive adaptation to the changing employment landscape.

9 Results and Analysis

...