Predicting Automation Risks in the AI Job Market

Soo Jin Jung

Pennsylvania State University
University Park, PA, USA
szj5489@psu.edu

Erika Ho
Pennsylvania State University
University Park, PA, USA
ezh5463@psu.edu

ABSTRACT

As Artificial Intelligence technology advances the transformation in the labor market is as fast as the innovations, understanding the skills that relate to job automation risk is a crucial advantage. This paper focuses on developing machine learning models to accurately predict the likelihood of job automation based on job characteristics and skill requirements in the AI job market. We used many different Machine Learning algorithms to highlight the importance of specific skills and job titles in reducing automation risk, offering awareness to plan the job market and policy in an increasingly automated job field. Throughout the different models used, we found that Random Forest Classifier gave the highest accuracy of 44%. We attempted to hyper tune the parameters, but did not see an improvement of accuracy within the model. Ultimately, Random Forest Classifier gave some insights, but not significant enough, to predict automation risk within certain jobs. The accuracy of the model is not high enough to make meaningful and reliable inferences about the potential of automation in jobs.

I. INTRODUCTION

How is AI taking over our jobs? The advancement of AI is key to reshaping the job market, as it introduces both chances and challenges to the table of jobs. As AI-driven automation becomes more widespread, questions about job security and the skill sets that can safeguard against automation are becoming key. Identifying the skills that reduce a job's risk of automation is crucial for developing a flexible workforce and expanding policies that support sustainable employment in an evolving economy and technological space.

This paper focuses on the development of machine learning models for the prediction of automation risk across occupations as a means of providing a tool that can be used for anticipating changes in the labor market. We will test various algorithms by integrating different features related to requirements of skills, job complexity, and other sector-specific factors that help us to select the most accurate predictive model. We hope this will help inform employers to develop workforce strategies to become better prepared for the future impact of automation. Pedagogically speaking, teaching students throughout their years of education to become proficient in skills that help mitigate the risk of automation in occupations is also beneficial.

Research Question: *How accurately can machine learning models predict the likelihood of job automation based on job characteristics and skill requirements across various occupations?*

II. Literature Review

A. *Parent Paper: AI meets labor market: Exploring the link between automation and skills*

In AI Meets the Labor Market: Exploring the Link Between Automation and Skills, Colombo, Mercorio, and Mezzanzanica (2019) investigates how AI-driven automation is altering skill requirements and affecting the structure of the labor market, specifically in Italy. By applying ML techniques to web-based job vacancies, the researchers develop a variety of skills linked to automation risk. This taxonomy integrates the European Skills, Competences, and Occupations (ESCO) classification system that shows skill requirements by occupation.

The study identifies two primary effects of AI on the labor market: (1) an extensive effect, which refers to the creation of new jobs and the displacement of existing ones, specifically in jobs that require more routine work (e.g., logistics, clerical work); and (2) an intensive effect, which involves the transformation of job roles through new skill requirements. The researchers conclude that occupations involving routine, manual tasks are at the highest risk of automation. On the other hand, roles that frequently use cognitive and interpersonal skills, such as decision-making and teamwork, are the least likely to get automated. The study's findings show that the demand for digital and soft skills is inversely related to automation risk. This suggests that non-routine tasks that are dependent on human interaction are difficult to automate.

B. *1st research Paper: Demand for AI skills in jobs: Evidence from online job postings*

The OECD's Demand for AI Skills in Jobs report (2021) by Mariagrazia Squicciarini and Heike Nachtigall presents a comprehensive analysis of the growing demand for AI-related skills across various sectors, using data from online job postings between 2012 and 2018. Based on the study, there are many more job postings that require AI skills especially in finance and information and communication technology. This reveals that AI-related jobs no longer need only technical skills, but both soft skills such as teamwork and communication as well as hard skills like machine learning and coding languages.

Squicciarini and Nachtigall write that AI-related skills that mostly relate to software engineering, big data analytics, and deep learning are the most sought after by employers. Additionally, problem solving and creativity are soft skills that are mentioned in job postings. This suggests that the integration of AI into the workplace is not just technical, but it also requires human-centered skills to adapt to the evolving landscape of AI. The paper shows that many sectors are incorporating AI to enhance efficiency, but the demand for soft skills reflects an understanding that interpersonal capabilities are just as important to implement artificial intelligence in the workplace.

C. *2nd research Paper: Dissection of AI Job Advertisements: A Text Mining-based Analysis of Employee Skills in the Disciplines Computer Vision and Natural Language Processing*

While many studies analyze AI skill requirements in general, Kortum, Rebstadt, and Thomas (2022) emphasize the importance of specialization within AI subfields in their paper Dissection of AI Job Advertisements: A Text Mining-Based Analysis of Employee Skills in the Disciplines of Computer Vision

and Natural Language Processing. This paper uses text mining techniques to extract and analyze computer vision (CV) and natural language processing (NLP) job postings, which are two popular subfields of AI. There are many more subfields of AI such as deep learning, neural networks, etc., but the paper delves into the two most popular subfields. By examining job advertisements on Indeed.com, this study reveals that each subfield requires a unique set of skills for each subfield.

For computer vision roles, the skills most frequently cited include image classification, object detection, and semantic segmentation, skills that are often applied in industries such as manufacturing, medicine, and retail. On the other hand, NLP roles are more likely to need skills in language modeling, sentiment analysis, and chatbot development, especially relevant in finance and customer service applications. This segmentation within AI roles underscores that a "one-size-fits-all" skill profile is inadequate for AI professions. Furthermore, this suggests that the risk of automation will vary significantly based on the specific requirements of each subfield. As AI-related job roles become more specialized, there is an increasing need for targeted training for their field and the ability to adapt to dynamic industry requirements.

D. Application

The insights from these three studies provide a foundation for our research, which seeks to develop and evaluate machine learning models that predict the likelihood of job automation. Through all this, we also had research gaps and challenges we had to adapt to. Each study highlights critical aspects that inform model design and feature selection for automation prediction:

1. **Hybrid Skill Demand and Model Features**:

   The OECD study reveals a trend toward hybrid skillsets combining technical and interpersonal skills. For our models, this suggests the need to incorporate skill-related features that account for both technical AI skills and soft skills. By capturing this balance, we aim to refine model accuracy in predicting automation susceptibility, particularly in jobs that integrate both cognitive and manual tasks.

2. **Task Complexity and Routine Predictors**:

   Colombo et al. (2019) demonstrate that routine-based occupations are more vulnerable to automation, whereas non-routine, cognitive-intensive roles are more resilient. This insight will guide feature engineering within our models, where we will assign task complexity scores to differentiate between routine and complex tasks. Additionally, leveraging job classification schemas, such as ESCO, allows our models to group similar tasks and gauge automation probabilities more precisely based on historical automation trends in those occupations.

3. **Subfield-Specific Skillsets and Specialized Models**:

   Kortum et al. (2022) show that distinct AI subfields (CV and NLP) require specialized skills, which could impact their automation risk differently. Drawing from this finding, our research will explore the development of

subfield-specific predictive models, particularly where job roles are tied to specialized skillsets. By tailoring models to subfield-specific features, we anticipate improved predictions for job automation across a widespread number and diverse occupations.

We encountered several research gaps that highlight areas for further exploration. Initially for our research, we wanted to predict artificial intelligence skills in jobs and our original parent paper was *Demand for AI skills in jobs: Evidence from Online Job Postings* [C] and through our presentation and peer reviews, our paper lacked machine learning methods and contained solely EDA. The data used in [C] was from Burning Glass Technologies (now Lightcast), an analytics software company that has comprehensive labor market data and analytics globally. We could not access this data due to having to pay for the data and the samples were too small to attempt to gain any insights from.

Hence, throughout our other research papers we landed on finding the automation risk in jobs from our new parent paper. We looked to Kaggle.com to find our data and it included automation risk; However, our key limitation was the reliance of synthetic data, which, while realistic, does not fully capture the intricacies and nature of a dynamic, real-world job market. This raises the need for high-quality, up-to-date datasets that include nuanced features such as specific task descriptions of jobs, skill hierarchies, and industry trends. Additionally, the models struggled to generalize across diverse industries.

Another gap emerged when attempting to understand how dynamic factors external from the labor market such as technological advancements or policy changes can influence automation risks over time. Yet again, the nature of attempting to predict policy change and technological advancements is difficult in itself.

Our research builds on these studies by creating predictive models that capture a range of job characteristics and skill demands, ultimately enhancing accuracy in predicting job automation risk. By integrating hybrid skill demands, task complexity, and subfield specialization, we aim to offer a comprehensive approach to modeling job automation that aligns closely with the realities of the evolving AI labor market.

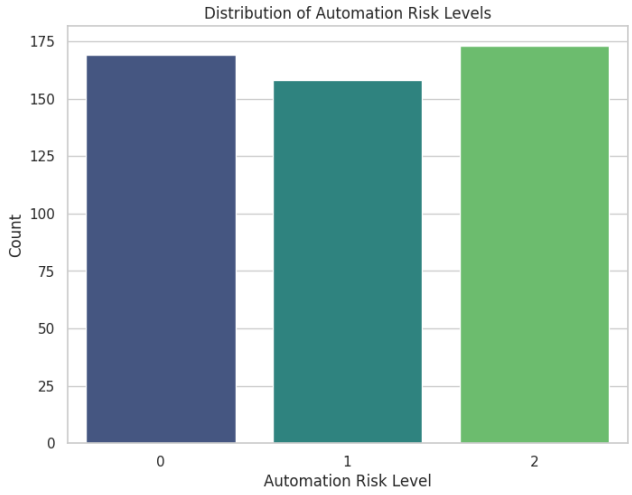
III. CONTRIBUTIONS

Throughout our own work of viewing the three papers aforementioned, we have decided to conduct our own proposition and findings that are inspired by the papers. We aim to find out the risk of automation within jobs across the world and in different sectors of the economy by building different machine learning models.

IV. METHODOLOGY

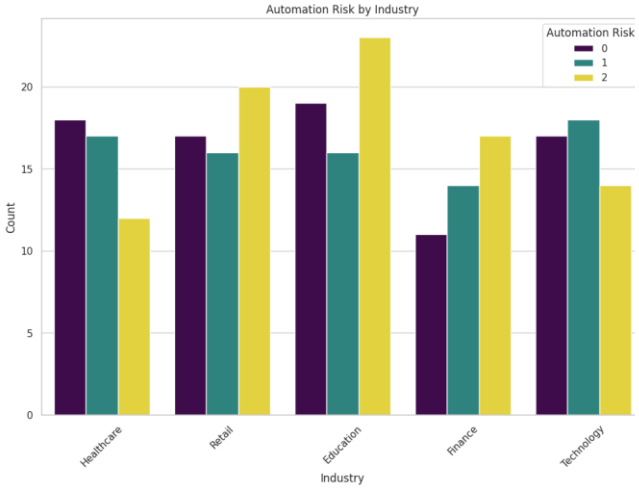
A. *Data pre-processing:*

Our data is synthetic from Kaggle, but it is stated to be a realistic snapshot of the job market. We do understand that using synthetic data does not adequately represent the job market in real-time, but this is the closest data set we found that also discusses the risk of automation. There were no missing values within our data to fix but below is a picture of the distribution of automation risk from 0-2 across all industries. In the original data automation risk was divided into low, medium, and high which we converted to numerically, 0 being low risk, 1 being medium risk, and 2 is high risk of automation (Graph 1).



Graph 1

We also separated the automation risk by industry (Graph 2)



Graph 2

To continue our pre-processing, we encoded our categorical variables to be able to use them in our analyses and split the data into train and test data to 80% and 20%, respectively.

B. *Method:*

Throughout our process, we have tried many different models as well as hyperparameter tuning them as well. However, there are some pros and cons to hyperparameter tuning that can lead to better or worse accuracy than before.

Hyperparameter tuning is essential for optimizing machine learning models, as it helps identify the best parameter settings that maximize model performance. One of the advantages of hyperparameter tuning is its potential to significantly increase accuracy. By adjusting parameters such as learning rate, depth of decision trees, or regularization terms, tuning allows the model to better capture underlying patterns in data, reducing bias, and improving accuracy. Additionally, effective tuning can lead to more efficient models that balance complexity and performance, reducing the risk of overfitting by identifying optimal parameter settings.

However, hyperparameter tuning has some drawbacks. It can be computationally intensive, especially with complex models like neural networks or ensemble methods, as each combination of parameters requires training and validation. This process can be time-consuming and may require substantial computational resources, which can be a limitation in certain environments. Moreover, poor tuning can lead to overfitting if parameters are set to fit the training data too closely, resulting in lower accuracy on new, unseen data. Also, insufficient tuning or the use of default parameters may cause the model to underfit, where it fails to capture important patterns. Thus, while hyperparameter tuning is powerful for boosting accuracy, achieving optimal results depends on a balance of experimentation, efficiency, and model validation techniques.

To preface, throughout all our methods listed we used every attribute from our data when training and did not drop or leave out specific columns. Each time we used GridSearchCV() our parameters used 5 folds and scoring = ‘accuracy’.

C. *Algorithm:*

To perform our methods, we used many Python packages mainly from sklearn. Initially, we started with logistic regression to test the accuracy of automation of our data using all the attributes, since we converted categorical values.

We received an accuracy of 35%, leading us to use GridSearchCV() with 5 folds in the sklearn.model_selection package to hypertune the parameters and it gave us a lower accuracy of 33%. We presume that the model overfits the data when hypertuning.

Secondly, we attempted to use DecisionTreeClassifier() with all attributes and then also used GridSearchCV() again.

Before tuning we had an accuracy of 38% and after we had an accuracy of 33% which we also attest to overfitting.

Next, we used RandomForestClassifier() and received a better score of 41% accuracy and then with tuning, we got a better accuracy of 44%. This is the first time we have used tuning and attained better accuracy than before. This also occurred when using KNeighborsClassifier() when before tuning the accuracy was 30% and after it was 36%. Support Vector Machine accuracy was 43% and after tuning it was blank%. We also tried XGBoost() which gave an accuracy of 41% and after tuning it gave an accuracy of 37%.

In all models, accuracy is calculated as the number of correct predictions divided by the total predictions made.

V. Results and Discussions

A. *Results*

The primary goal of this study was to develop machine learning models to predict automation risk for various job roles based on their characteristics and skill requirements. The key results from different algorithms are summarized as follows:

- Logistic Regression: Achieved 35% accuracy, which dropped to 33% after hyperparameter tuning.
- Decision Tree Classifier: Initial accuracy of 38%, reduced to 33% post-tuning, suggesting overfitting.
- Random Forest Classifier: The highest-performing model with an accuracy of 44% pre-tuning and 41% after tuning.
- K-Nearest Neighbors (KNN): Improved from 30% to 36% accuracy after tuning.
- Support Vector Machine (SVM): Reached 43% accuracy.
- XGBoost: Dropped from 41% to 37% after hyperparameter tuning.

Despite employing various techniques and models, the highest accuracy achieved was 44% using the Random Forest Classifier after tuning.

B. *Discussion*

Our results highlight the complexities when attempting to accurately predict job automation risks using machine learning models. While the Random Forest Classifier achieved the highest accuracy of 44%, this low percentage underscores the limitations of data and modeling approaches when trying to capture the job automation risk in such a dynamic environment. Factors such as synthetic data and generalization across many industries are the likely why our models performed poorly.

The variability in the model accuracy suggests that automation prediction is highly sensitive to feature selection and quality of data. By including categorical variables such as job titles, industries, and skill requirements provided a foundational framework, but the lack of real-world data limited the depth and relevance of these categorical variables. Moreover, models like Logisitic Regression and Decision Trees struggled to generalize, with tuning often resulting in overfitting. This reveals that more sophisticated feature engineering or domain-specific adjustments may be necessary to point towards a higher accuracy.

The performance of Random Forest and Support Vector Machines after hyperparameter tuning indicates that ensemble and kernel-based methods may be better suited for predicting job automation. However, the inconsistency in accuracy improvements across all models after hyperparameter tuning suggests that more experimentation with alternative real-world data or more job-specific features could improve the predictive accuracy.

Our findings emphasize the need for targeted improvements in methodology and collecting of data. Integrating real-time labor market data or pairing machine learning models with feedback from professionals could yield more reliable insights. Additionally, extending the scope of the data to include more sector specific analyses or subfield specific predictions could give more results that are actionable because of the nature of the demands of different industries.

Ultimately, our research reveals limitations in its predictive accuracy, but it provides a foundational framework for understanding automation risk through data-driven approaches.

VI. Conclusion

This study demonstrates the possibility of using machine learning models to forecast automation risks in the job field. While the models were useful, the overall predictivity that it had was improvable, with the best results being Random Forest at 44%. This suggests that the approach taken, based on synthetic data and simple machine learning models, cannot deliver the precision required for reasonable insights.

VII. Future Work

Based on our improvable accuracy we got here are some insights for future exploration:

- Enriching Dataset:
  - → Real-world data that showcases information like industrial trends, regional difference, and skills needed would help improve the model's reliability.
- Advance Algorithms:
  - → The complexity of predicting job automation risks demands a more advanced approach than traditional machine learning models. Including deep learning models like Recurrent Neural Networks (RNNs), as it processes sequential data, capturing temporal patterns, or Convolutional Neural Networks (CNNs), as it analyzes grid-like data identifying dimensional features, which could reduce the limitations of this paper.

These changes in future studies would help reduce the lack of accuracy and help expand the project to a bigger scale.

VIII. Reference

[1] Colombo, Emilio, Fabio Mercorio, and Mario Mezzanzanica. "AI meets labor market: Exploring the link between automation and skills." Information Economics and Policy 47, no. 4 (2019): https://www.sciencedirect.com/science/article/pii/S0167624518301318

[2] Kortum, Henrik, Jonas Rebstadt, and Oliver Thomas. "Dissection of AI Job Advertisements: A Text Mining-based Analysis of Employee Skills in the Disciplines Computer Vision and Natural Language Processing." Paper presented at the 55th Hawaii International Conference on System Sciences, January 4–7, 2022, Virtual Event, https://scholarspace.manoa.hawaii.edu/bitstreams/d4581c13-2474-4e24-9640-efb4720015e4/download.

[3] Squicciarini, M. and H. Nachtigall (2021), "Demand for AI skills in jobs: Evidence from online job postings", OECD Science, Technology and Industry Working Papers, No. 2021/03, OECD Publishing, Paris, https://doi.org/10.1787/3ed32d94-en.