Asian Institute of Technology

School of Engineering and Technology

AT82.03: Machine Learning

Progress Report: Job Automation Risk Prediction

Date of Submission: 14th November 2024

Submitted to: Dr. Chaklam Silpasuwanchai

Submitted by

Kaung Nyo Lwin- st125066

Nyein Chan Aung- st125553

Phone Myint Naing- st124973

Khin Yadanar Hlaing-st124959

Table of Contents

Table of Figures

ABSTRACT

The increasing trend of computerization poses significant risks to the global job market, particularly for roles that involve repetitive tasks. This project examines the impact of computerization on the Myanmar job market using data sponsored by **MyJob**, one of Myanmar's pioneer online job platforms, classifying job roles based on their susceptibility to computerization. The project is grounded in Frey and Osborne's framework [1], which assigns probabilities to occupations based on their likelihood of computerization. We develop a job recommender system to assist individuals in exploring low-risk career paths and offer insights for educational institutions and policymakers aiming to prepare the workforce for an AI-driven economy.

1 INTRODUCTION

1.1 Background of the Study

Globally, advancements in digitization and computerization are reshaping industries, economies, and labour markets. The rapid evolution of artificial intelligence (AI) and automation technology has led to increased productivity, innovation, and efficiency across sectors. However, these advancements also pose risks, as many occupations that involve routine, repetitive tasks are becoming increasingly susceptible to computerization. A widely cited study by Frey and Osborne (2013) [1] highlights that nearly half of the jobs in developed economies may be at risk of automation. This global shift towards digital economies is compelling countries to reevaluate workforce needs and prepare for a future where computerization impacts many traditional roles.

In Southeast Asia, countries like Singapore, Malaysia, and Thailand have taken significant strides toward building digital economies, investing heavily in ICT infrastructure, digital skills, and regulatory frameworks that support innovation. Similarly, Myanmar has embarked on its own digital transformation journey, aiming to position ICT as a driver of economic growth. Key initiatives, such as the Myanmar Digital Economy Roadmap (2018-2025) [2], aim to expand digital services, increase digital financial transactions, and attract foreign investment to boost the sector. Complementing this, the Myanmar e-Governance Master Plan (2016-2020) focused on digitizing government services to improve efficiency and Myanmar's e-governance ranking, reflecting the country's commitment to modernization.

As Myanmar's economy and job market continue to evolve, it becomes essential to assess which roles are most vulnerable to computerization. Jobs with routine, predictable tasks—often common in agriculture, manufacturing, and administrative services—are at higher risk, potentially leading to job displacement in these sectors. This project aims to address this issue by evaluating the computerization risks of various roles within Myanmar's job market. Using data from MyJob and the computerization probability framework established by Frey and Osborne [1], this project will identify high-risk roles and provide job recommendations suitable to their skills. In doing so, it offers a timely resource for workers, educators, and policymakers to support Myanmar's workforce in transitioning to a more resilient, digitally informed economy.

1.2 Problem Statement

With increasing digitalization and advancements in artificial intelligence, computerization poses a significant risk to Myanmar's workforce. Jobs that involve repetitive tasks or require limited skills are particularly susceptible to being replaced by automated systems. However, Myanmar currently lacks an established system to evaluate the computerization risk for various roles or to offer guidance for career transitions to less vulnerable occupations. Given the economic implications of such displacement, it is crucial to understand which roles are at the highest risk and to provide individuals with actionable insights on career planning and skill development.

This project aims to fill this gap by creating a computerization risk assessment model based on MyJob data. Using the framework developed by Frey and Osborne, which estimates computerization probabilities across occupations, this project will assign risk levels to various job roles within Myanmar. A job recommender system will complement this analysis, helping individuals explore alternative career paths with lower computerization risk and supporting policymakers in targeting skill development initiatives.

2 RELATED WORK

The basis for our automation risk model comes from Frey and Osborne's study, which assesses the computerization probabilities of 702 occupations. Other related studies include models predicting automation's impact across sectors, but most have limited applications in emerging markets like Myanmar. Our work contributes by adapting these models to a developing economy context with job data specific to Myanmar's job market.

3 DATASET

3.1 Description

We use job data sponsored by MyJob [3]; a pioneer of recruitment services in Myanmar; incorporating two main datasets: job listings and job skills. The job listings dataset provides details on various roles, including job descriptions and industry information, while the job skills dataset links specific skills to each job. Together, these datasets enable us to examine Myanmar-specific job roles and their required skills, supporting the assessment of each role's automation risk through comparisons with Frey and Osborne’s occupation classifications.

3.2 Features

From the Job Listings Dataset:

- **ID**: A unique identifier for each job listing.
- **Title**: The specific job title (e.g., "Data Analyst"), used to categorize and analyse different roles.
- **Description**: Detailed tasks and responsibilities associated with each role, aiding in the identification of skill requirements.
- **Job Summary**: A brief overview of the job role, summarizing key duties or requirements.
- **Created At**: The date the job listing was initially posted, is useful for analysing job posting trends over time.
- **Updated At**: The most recent date the job listing was modified, providing insights into evolving job roles.
- **Search**: Likely an internal indexing feature for locating job listings within the platform.
- **Anonymous**: Indicates whether the job listing is anonymous, possibly affecting the visibility of the employer information.
- **Slug**: A unique, user-friendly identifier used in URLs for each job listing.
- **Industries**: Specifies the industry sector (e.g., IT, Finance) that the job belongs to, helping to categorize roles by industry.

From the Job Skills Dataset:

- **Job ID**: This serves as an identifier that links each job listing to its associated skills. It connects with the Job ID in the job listings dataset, allowing us to see which skills are required for each particular job role.
- **Job Skill ID**: A unique identifier for each skill associated with a job, which allows for precise tracking of required skills.

- **Job Skill**: Lists the specific skills associated with each job (e.g., "General management," "Data analysis"). This field provides a direct mapping between job roles and their required skills, allowing for detailed skill analysis for each job.

4 METHODOLOGY

4.1 Risk Assessment

We assign automation risk probabilities based on Frey and Osborne's automation likelihood for similar occupations:

- **Low Risk**: Probability < 0.3
- **Moderate Risk**: 0.3 ≤ Probability < 0.7
- **High Risk**: Probability ≥ 0.7 Each job role in Myanmar is categorised by matching its skills and responsibilities with those in Frey and Osborne's classifications. This probability score allows job seekers to assess their risk level based on current automation trends.

4.1.1 Frey and Osborne's Method for Computerization Probability

In Frey and Osborne's method , there are three main indicators, which are the engineering bottlenecks for computerization, that need to be assessed in order to predict the probability of computerisation. They are perception and manipulation, social intelligence and creativity. In their research, these qualities are hard to replace by computers. The probability of computerization will vary on how much these three qualities are needed in a job, as illustrated in Figure 1.

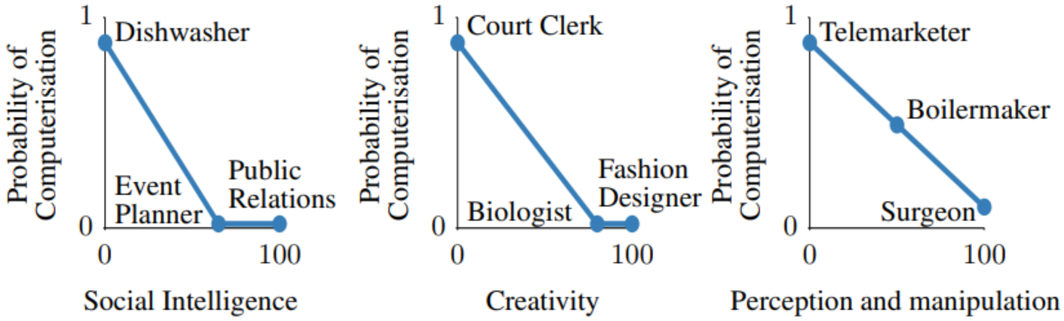

Figure 1 - [1] A sketch of how the probability of computerization might vary as a function of bottleneck variables

Furthermore, they used nine variables of abilities that constitute these qualities. The variables are called O*Net Variables, derived from O*Net data and shown in the table below.

Computerisation bottleneck	O∗NET Variable	O∗NET Description
Perception and Manipulation	Finger Dexterity	The ability to make precisely coordinated movements of the fingers of one or both hands to grasp, manipulate, or assemble very small objects.
	Manual Dexterity	The ability to quickly move your hand, your hand together with your arm, or your two hands to grasp, manipulate, or assemble objects.
	Cramped Work Space, Awkward Positions	How often does this job require working in cramped work spaces that requires getting into awkward positions?
Creative Intelligence	Originality	The ability to come up with unusual or clever ideas about a given topic or situation, or to develop creative ways to solve a problem.
	Fine Arts	Knowledge of theory and techniques required to compose, produce, and perform works of music, dance, visual arts, drama, and sculpture.
Social Intelligence	Social Perceptiveness	Being aware of others' reactions and understanding why they react as they do.
	Negotiation	Bringing others together and trying to reconcile differences.
	Persuasion	Persuading others to change their minds or behavior.
	Assisting and Caring for Others	Providing personal assistance, medical attention, emotional support, or other personal care to others such as coworkers, customers, or patients.

Figure 2 - [1] O∗NET variables that serve as indicators of bottlenecks to computerization.

Finally, they train an ML model with 70 hand-labelled jobs using these nine variables as predictors, producing the probability of 702 occupations according to the predictions of this model. The distribution of these variables is shown in Figure 3.

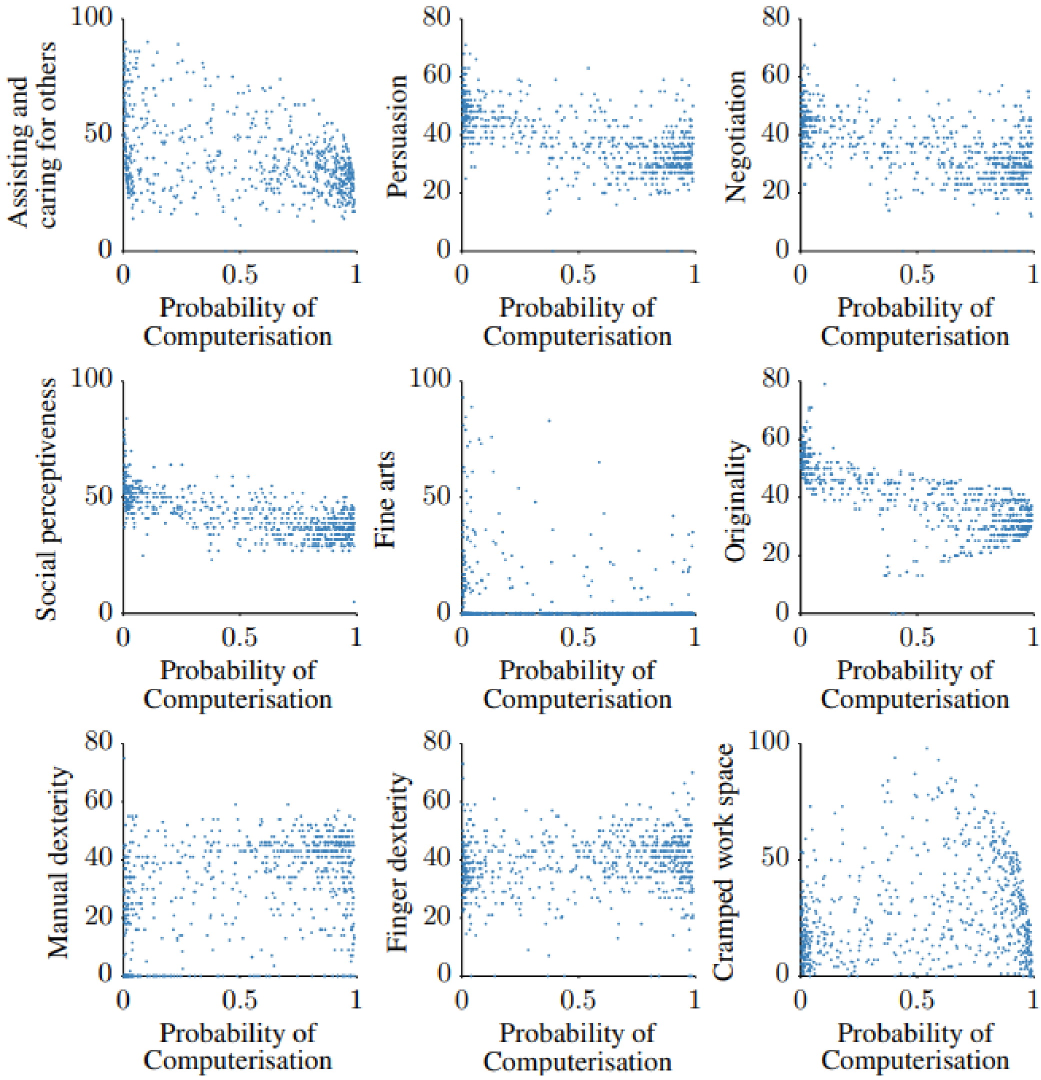

Figure 3 [1]- The distribution of occupational variables as a function of the probability of computerization; each occupation is a unique point.

4.1.2 Adapting Frey and Osborne's Method for Myanmar

We have decided to use Frey and Osborne's method to predict automation risk using the job data from Myanmar. As Myanmar is a developing country, cutting-edge technologies like AI do not impact the labour market yet. The current state of various industries in Myanmar is still in the process of digital transformation, which is fairly similar to the situations in the period of Frey and Osborne research over a decade ago. Therefore, we assume that the predictions based on the O*Net variables are still suitable for the labour market in Myanmar, which may not be the case in developed countries anymore. Moreover, Frey and Osborne's Method relies on the basic skills necessary for an occupation. As these basic skills are fairly similar regardless of the geolocation, we have justified this method as a suitable approach.

We will reference the probability of 702 occupations to calculate the risk. We will use a sentence transformer model to check the similarity between these 702 occupations and the input job title. Since the job titles which will be imported by users will be different given the labor market are different in Myanmar, taking only one probability score with the most similarity score is more prone to error. So, to be more robust, we should use the average of the topmost similar scores. However, the challenge to find how many top scores should be used arose as the probability highly varies if the number of top scores to be average changes. We have decided to use k-means clustering to overcome this. We will average the probability scores of each sample in the predicted clusters.

We also consider that the list of skills required for the same job title will be different in the context of Myanmar. So, we attempt to calculate the probability of automation based on the skills imported by users. Since there are nine bottlenecks variables defined in the Frey and Osborne paper, we will find the similarity scores between each skill and these variables using a transformer encoder model. The average of these scores will be taken into account in the risk calculation. The calculation can be summarized as follows:

$$P_{skill} = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} (1 - S_{ij}) \tag{1}$$

$$P_{title} = \frac{1}{m} \sum_{i=1}^{m} P_i \tag{2}$$

$$Risk = \frac{P_{title} + P_{skill}}{2} \tag{3}$$

where P is the probability of computerization, m is the number of samples belonging to the predicted group cluster, n is the skills, k is the O*Net variables and S is the similarity score between skills and O*Net variables.

4.2 Job Recommender System

For high-risk roles, our recommender system suggests alternative careers with lower automation risks. Using job similarity analysis, the system identifies viable transitions by comparing skill sets across jobs. This system is particularly valuable for workers in high-risk roles seeking to transition into more secure career paths without significant skill overhauls.

Our job recommender system is based on a sentence transformer and a K-nearest neighbours algorithm. The sentence transformer will embed the list of skills to the numerical data so that the nearest neighbours algorithm can find the nearest neighbours. The diagram is shown in Figure 4.

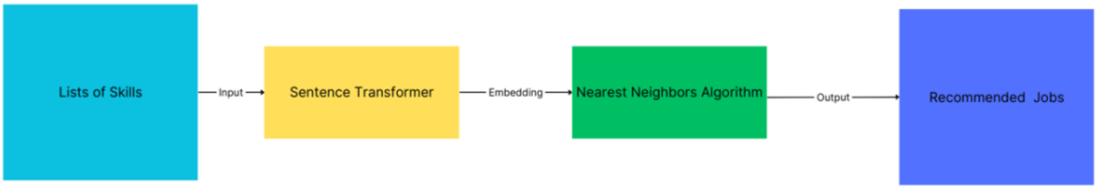

Figure 4 - Job Recommender System

4.3 Risk Prediction System

In order to predict the risk, we need to find the computerization probability of the input job title and the estimated similarity between the list of skill input and O Net variables. We use the k-means clustering algorithm to predict which group the input job title belongs to so that the average probability of the predicted can be calculated. On the other hand, we use the cosine similarity score for estimating O Net variable scores. The overview of the risk prediction system is shown in figure 5.

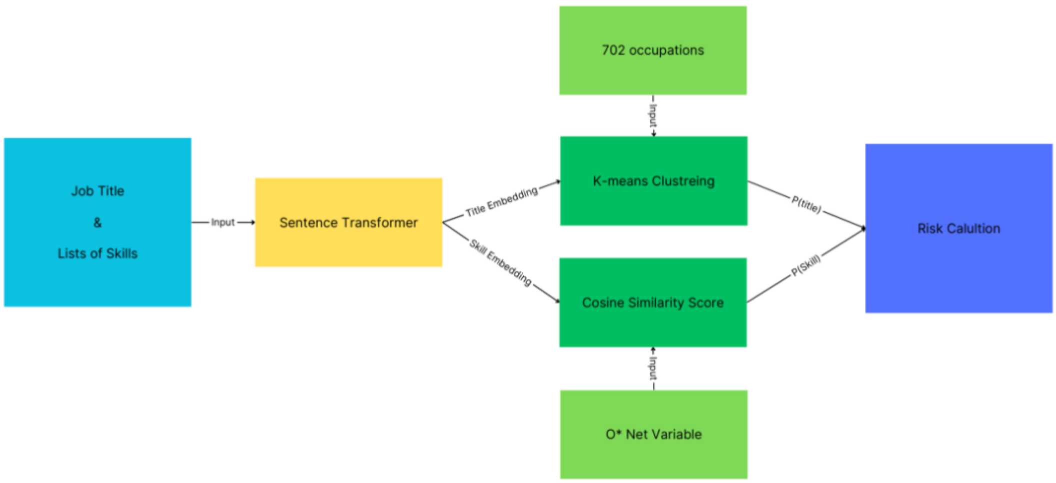

Figure 5 - Risk Prediction System

5 Model Development

5.1 Data Preparation

As our datasets are from a company in Myanmar, the raw data is not clean enough to use. There are many steps to do in order to get clean and usable data for our project such as merging, cleaning and feature extractions. The raw data contains two parts: job posts and job skills. These two datasets can be combined into one using Job ID and then after following imputation, the main part is to extract features from the combined dataset. Our purpose is to build a dataset containing job titles, industries and job skills so we aggregate and make a list of industries and job skills according to job titles. Here is the result of our final data.

	Title	Industries	Job Skills
1	A Level English Teacher Full Time Part Time	[Education Training]	[Teaching sharing Knowledge]
2	A Level Maths Teacher Full Time Part Time	[Education Training, Education Training, Educa...	[Teaching sharing Knowledge, Maths, Good at Math]
3	Academic Coordinator Native Only	[Education Training, Education Training, Educa...	[MS Office Applications, Handling Pressure, Tr...
4	Account	[Medical Healthcare]	[Accounting and Customer service]
5	Account Executive Finance Account	[Advertising Media Communications, Advertising...	[Interpersonal skill, Positive Personality Att...
...	...	...	...
874	property sale marketing supervisor	[Property Real Estate, Property Real Estate, P...	[Supervisory skills, Customer Management, Sale...
875	sale admin	[Retail Fashion FMCG, Retail Fashion FMCG]	[Sales admin, Microsoft s]
876	stock Controller and Purchaser	[Retail Fashion FMCG, Retail Fashion FMCG]	[Proactive and teamwork skills, Quality Assura...
877	test	[Government Public Relations]	[Proactive and teamwork skills]
878	testing	[Manufacturing Warehousing]	[Banking and Finance Skill]

878 rows × 3 columns

Figure 6 - Final dataset for the recommender model.

5.2 EDA Results

According to our dataset, there are some descriptive results after the exploratory data analysis and predictive results after testing our prototype models. EDA plays a very important role in our project to consider what is happening to our data. We learned some important facts from the analysis results and made our decision right in choosing a model.

5.2.1 Job Skills

After the EDA step, we discovered the top job skills demand for the last two years. Among them, Teamwork skill is the most required skill for jobs in Myanmar. Some skills like IT, Finance, Accounting and Marketing skills follow behind but it can be seen that there are almost half of the top ones. There are 1,041 unique skills in our dataset and here are the top ten skills to visualize the frequency.

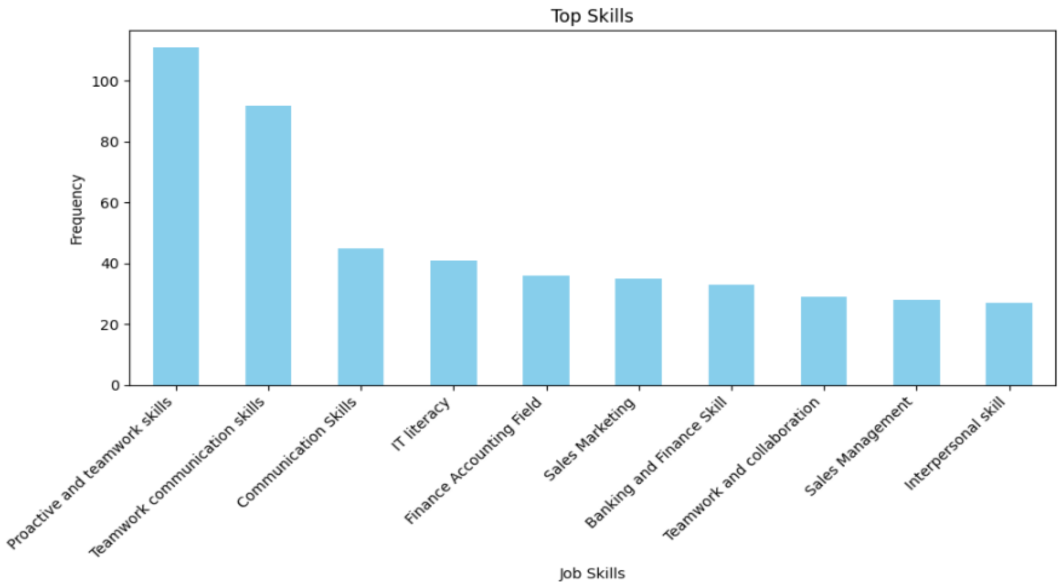

Figure 7- EDA result on job skills

5.2.2 Industries

There are 26 industries in total, according to our dataset and manufacturing and warehousing stand at the top of them. IT Telecom, Construction Engineering, Educational Training, Advertising Media and Retail Fashion related jobs seem to be equally standing behind the top ones.

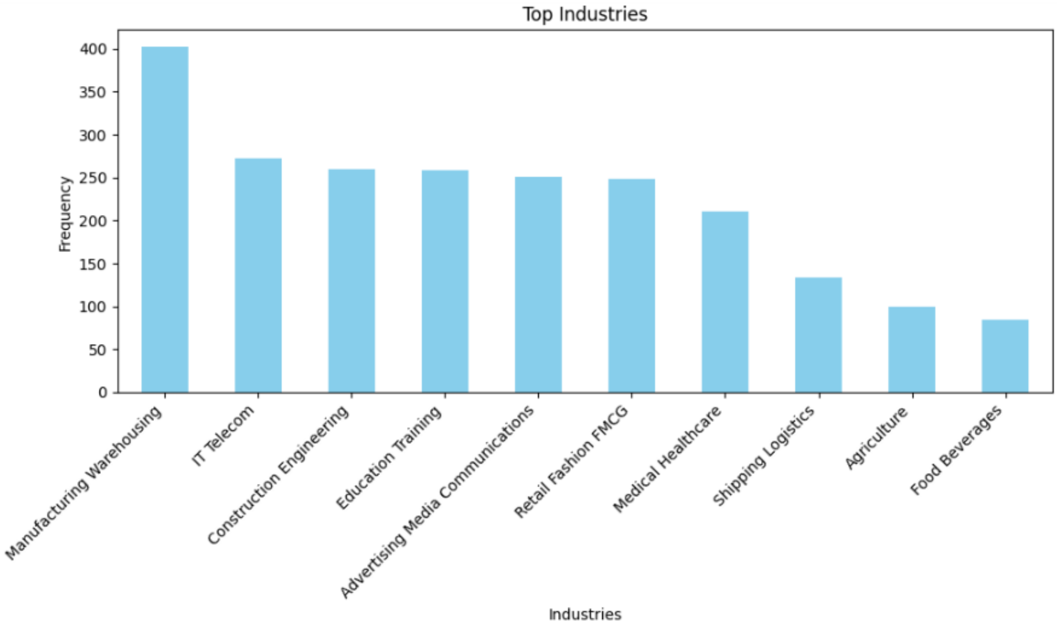

Figure 8 - EDA result on industries

5.2.3 Job Titles

In this section, Sales and Accountant job titles are mostly in the top ten list. Graphic designers differ from the other job titles and stand in second place.

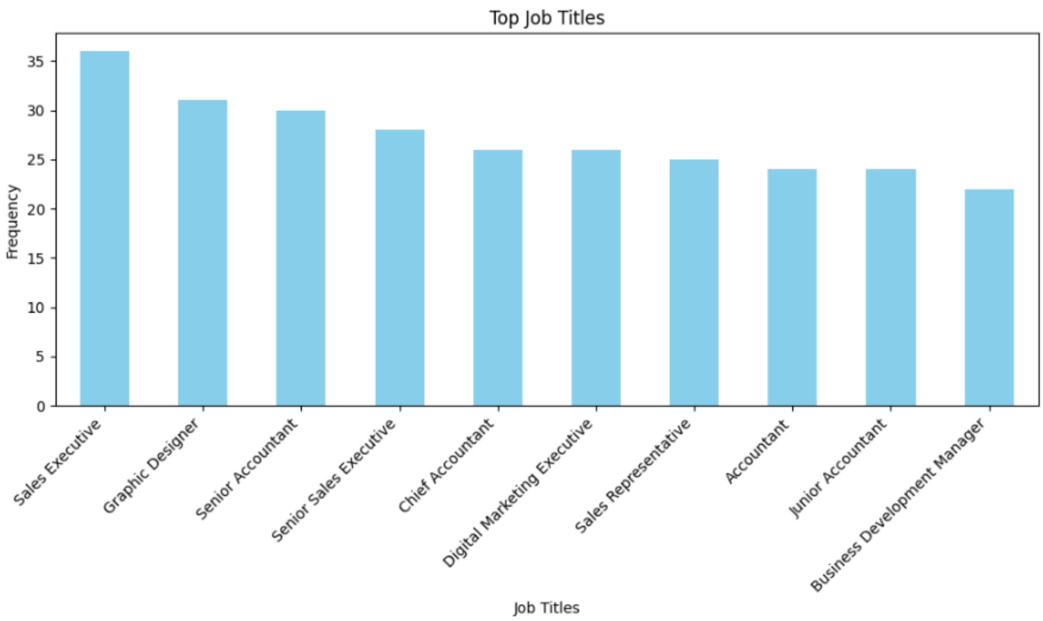

Figure 9 - EDA result on job titles

5.3 Modelling

5.3.1 Text Encoding

There are two prototypes for our Job Recommender System: Text Vectorizer + KNN and Sentence Transformer + KNN. We initially evaluate the results of these two models, and later, we will evaluate them on a test set. The number of neighbours in KNN is fixed as five and then we calculate to find which combination gets the best result. We consider the average distance between the nearest neighbours as a metric. The following figure is the evaluation result of our two prototype models.

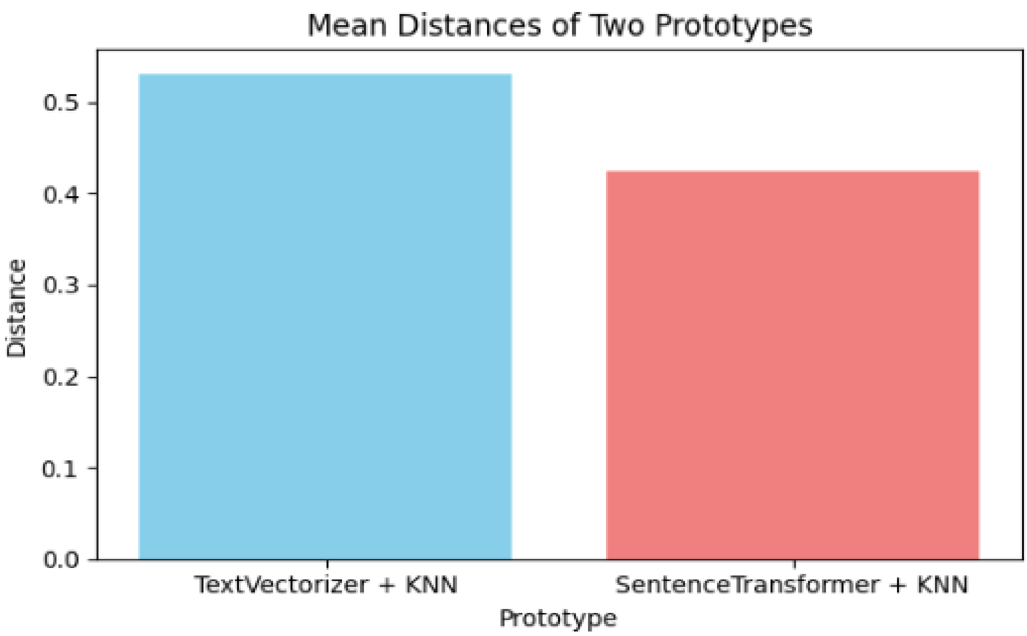

Figure 10 - Text Encoder Evaluation

5.3.2 Prototype 1 for job recommender

In the first prototype, Text Vectorizer is used to make the words in the skills into vectors and then use the KNN model to find the nearest job skills. In the figure, the Programming and Networking skills as the input and what we get are the job titles and industries for recommendation.

```
In [58]: query_vec = tf.transform(['Programming and Networking'])
         distances, indices = knn.kneighbors(query_vec)
         indices[0]

Out[58]: array([ 54, 352,  50, 520, 692], dtype=int64)

In [59]: distances

Out[59]: array([[0.32343944, 0.37157009, 0.4410423 , 0.67345084, 0.84553429]])

In [60]: distances.mean()

Out[60]: 0.5310073912643919

In [41]: df.iloc[indices[0]]

Out[41]:
```

	Title	Industries	Job Skills
55	Area Sales Executive Post	[Medical Healthcare, Medical Healthcare]	Networking Specializezd and Networking Special...
353	IT Support Associate EPS	[Travel Tourism Transportation]	Networking basic knowledge
51	Apprentice Technician	[IT Telecom]	PC Networking A Advanced
521	Presale Engineer	[IT Telecom, IT Telecom, IT Telecom, IT Teleco...	IT literacy and Teamwork communication skills ...
693	Sales Representative	[Shipping Logistics, Shipping Logistics, Manuf...	Sales Marketing and Logistics and Shipping and...

Figure 11 - Result of prototype 1

5.3.3 Prototype 2 for job recommender

For the second one, we use a pre-trained Sentence Transformer model from the Hugging Face platform to produce word embeddings and then as in the previous one, we use the KNN model to find the nearest job skills.

```
n [50]: query_emb = model.encode(['Programming and Networking'])
        distances, indices = knn.kneighbors(query_emb)
        indices[0]

ut[50]: array([352,  50, 601,  88, 798], dtype=int64)

n [52]: distances

ut[52]: array([[0.3191659 , 0.37865835, 0.45635784, 0.47077298, 0.49801326]],
              dtype=float32)

n [53]: distances.mean()

ut[53]: 0.42459363

n [51]: df.iloc[indices[0]]

ut[51]:
```

	Title	Industries	Job Skills
353	IT Support Associate EPS	[Travel Tourism Transportation]	Networking basic knowledge
51	Apprentice Technician	[IT Telecom]	PC Networking A Advanced
602	SCM Department Procurement Executive	[IT Telecom]	Telecom engineering
89	Assistant Service Engineer	[Medical Healthcare, Medical Healthcare]	Computing Skills and Engineering
799	Superintendent Assistant Engineer Fiber Mandalay	[IT Telecom]	Telecommunication Industry

Figure 12 - Result of Prototype 2

5.3.4 Risk Prediction

In order to calculate the risk, we need to predict two computerization probabilities of both job titles and a list of skills. We use a k-mean clustering model to cluster the list of occupations, that are the predicted results of our referenced paper so that we can calculate the probability of job title based on the predicted cluster. So, we search for the optimal clusters using the elbow method. We found that the optimal cluster number is 20, which can be seen in the following figure.

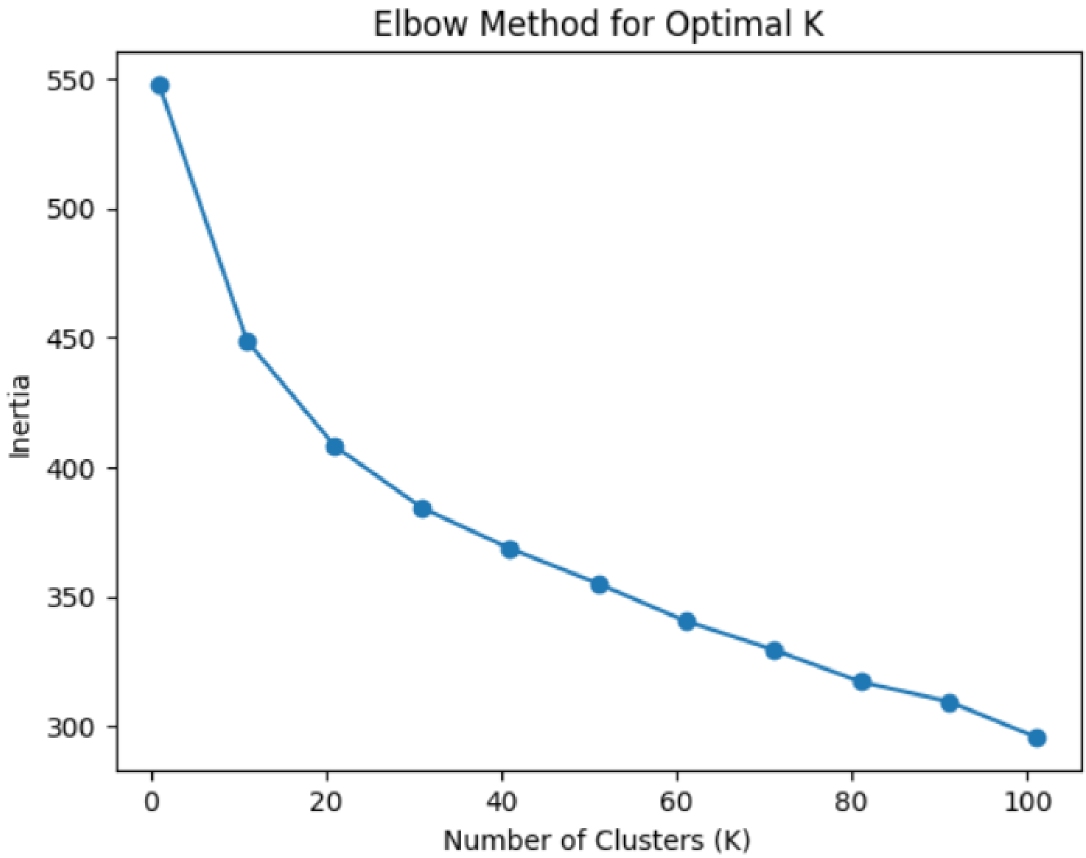

Figure 13 - Optimal cluster number

5.4 Model Evaluation

The evaluation process uses data from the BBC Business Insider dataset, which originally contains 365 job titles and their corresponding risk scores. To align with our model's approach, which calculates risk based on both job titles and associated skills, we enriched the dataset by adding relevant skills. This created a comprehensive test set that matches the model's requirements.

The evaluation metrics include:

- **Mean Squared Error (MSE):** Measures the average squared difference between the true and predicted risks, providing insight into prediction accuracy.
- **Mean Absolute Percentage Error (MAPE):** Evaluates the percentage error, offering a relative measure of prediction performance.

This evaluation ensures that the model's predictions are rigorously compared against real-world data, validating its effectiveness in estimating automation risks.

```
mean_squared_error(y_true, y_pred)
```

```
9.147299250785093
```

```
mean_absolute_percentage_error(y_true, y_pred)
```

```
3.631028576493364
```

Figure 14 - Result of model evaluation

The results of model evaluation can be seen in the figure 11. For 365 test samples, the mean square error is approximately 9. In order to see how much percentage the predictions vary from the test set data; we also measure the mean absolute error. It is around three per cent. Since we have taken into account the list of skills that should belong to the job title in risk calculation, the calculated risk from our should not be too close to the test set. On the other hand, the predicted results should not also be too far from the test data, which would indicate that the method of calculating the risk is not effective.

Based on the results, we have justified that the risk calculation method and our model predictions are between these two extremes.

5.5 Application Development

5.5.1 Backend

We are developing a backend system with Django, which will manage the database interactions, API endpoints, and business logic for calculating job risk assessments.

5.5.2 Database

SQLite is currently used, and it is suitable for development and prototyping. The following database diagram is designed to support functionalities like storing user profiles, tracking user skills, recording interactions with posts, and storing individual risk assessment results.

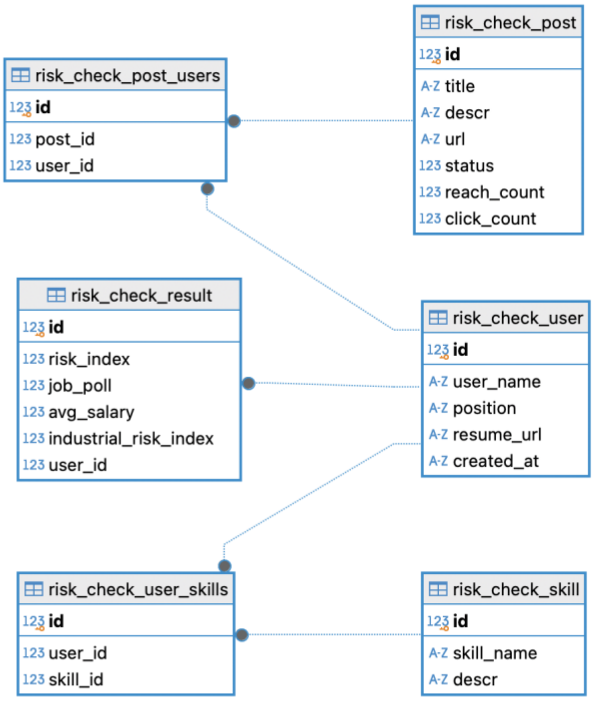

Figure 15 - ERD Diagram

- **risk_check_user**: Stores user information, including their name, job position, resume URL, and account creation date.
- **risk_check_skill:** Contains a list of skills, with each skill having a name and description.
- **risk_check_user_skills:** Acts as a bridge table connecting users and their skills, associating each user with various skills by user_id and skill_id.

- **risk_check_post**: Holds information about risk-related posts, including title, description, URL, status, reach count, and click count, possibly to display job risk insights or updates.
- **risk_check_post_users**: Connects posts with users, possibly to track which users interacted with specific posts.
- **risk_check_result:** Stores risk assessment results for each user, with fields for risk index, job poll, average salary, industrial risk index, and a reference to user_id.

5.5.3 Frontend

The Django web app uses Django's default templating system to render HTML pages. This setup allows you to integrate server-side rendering for dynamic content with minimal JavaScript, ensuring a straightforward and cohesive structure. Using Django templates simplifies development by embedding Python code directly in the HTML, making it easier to dynamically display job risk assessments and recommendations to users without needing a separate frontend framework.

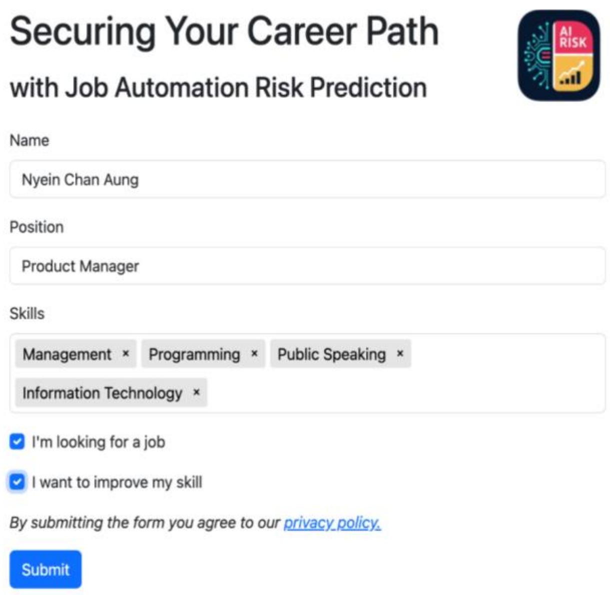

Figure 16 - User Input Page (Home)

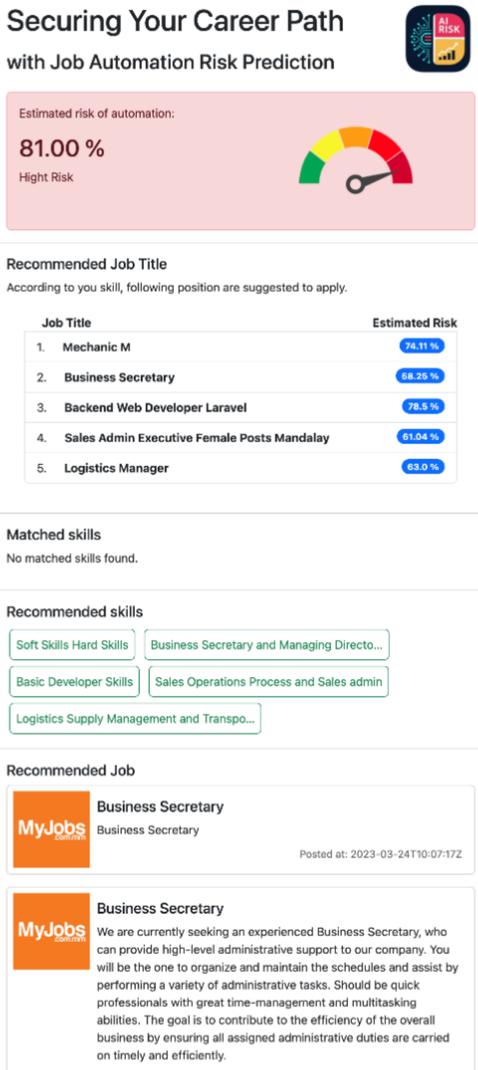

Figure 17 - Result Page

5.5.4 CICD Pipeline

The web application was developed using **Django**, a high-level Python web framework known for its robust architecture and scalability. Django provided a structured environment for creating the backend system, enabling seamless handling of database interactions, API endpoints, and integration with the machine learning model used for automation risk assessment. Its built-in features, such as the ORM (Object-Relational Mapping), authentication modules, and templating engine, facilitated a streamlined development process.

To ensure efficient deployment and updates, a **CI/CD pipeline** was implemented using **GitHub Actions**. The pipeline automates key processes like:

- **Code Integration:** Every change pushed to the main branch is automatically tested to ensure it doesn't break the application.
- **Build Process:** A Docker container or similar environment is created, encapsulating all dependencies and configurations for consistent deployment.
- **Deployment Automation:** Upon successful testing and building, the application is deployed directly to the production environment.

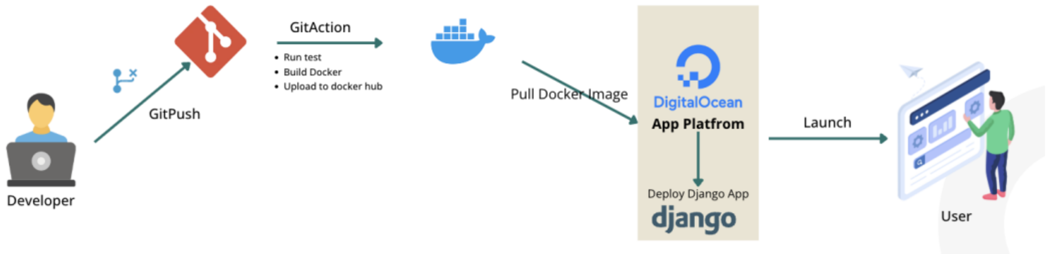

Figure 18 - CI/CD pipeline

5.5.5 Application Development and Deployment

The application was deployed on **DigitalOcean's App Platform**, a serverless PaaS (Platform as a Service) offering. By utilizing DigitalOcean's serverless architecture, the deployment eliminates the need for managing underlying infrastructure, enabling the team to focus solely on application functionality and user experience. The App Platform dynamically scales resources based on traffic, ensuring high availability and performance while optimizing costs.

This combination of Django for development, GitHub Actions for CI/CD, and DigitalOcean App Platform for deployment results in a highly efficient, scalable, and maintainable web application that delivers real-time automation risk assessments to users.

Application Demo Link: https://shark-app-ylemj.ondigitalocean.app

6 Discussion

6.1. Limitation

While this study presents a novel approach to job recommendation, several limitations must be acknowledged:

1. **Temporal Scope**: The dataset used in this study is limited to a specific time range, which may result in recommendations that do not adequately reflect jobs emerging outside this period.
2. **Data Source Homogeneity**: The dataset is derived from a single source, potentially limiting its ability to generalize across diverse job types and contexts.
3. **Model Dependence on Context**: The Sentence Transformer and K-Nearest Neighbors (KNN) approach relies heavily on the contextual relevance of job skills and titles, which may affect its robustness in scenarios with insufficient or ambiguous context.
4. **Reference Dependency**: The variables from the O*NET database used for training and evaluation are based on a specific research paper, creating a reliance on the validity and scope of these predefined variables.

6.2. Future Work

Several avenues for future work could enhance the scope and performance of the proposed system:

1. **Dynamic Data Collection**: Developing a data pipeline to crawl job postings daily and update the model monthly could ensure the inclusion of new and emerging job opportunities, improving the relevance of recommendations.
2. **Integration with External Platforms**: Collaborating with job hiring platforms, such as LinkedIn, could enable system integration and access to a broader dataset, enhancing the model's diversity and generalization capabilities.
3. **Model Fine-Tuning**: Fine-tuning the Sentence Transformer model with domain-specific job posting data could expand its knowledge base, leading to more accurate and context-aware predictions.
4. **Risk Prediction Model**: Building a risk prediction model using data such as AI advancement trends and job market statistics could provide insights into potential job market disruptions and enable proactive system enhancements.

7 References

[1] C. B. F. &. M. Osborne, “The Future of Employment: How susceptible are jobs to computerisation?,” 01 September 2013. [Online]. Available: https://www.oxfordmartin.ox.ac.uk/publications/the-future-of-employment. [Accessed November 2024].

[2] “ICT: From The Report: Myanmar 2020,” 2020. [Online]. Available: https://oxfordbusinessgroup.com/online-reader?id=183162. [Accessed November 2024].

[3] MyJob, “MyJob,” MyJob, [Online]. Available: https://www.myjobs.com.mm/. [Accessed 01 November 2024].

[4] “Will a robot take your job?,” 11 September 2015. [Online]. Available: https://www.bbc.com/news/technology-34066941. [Accessed 11 November 2024].