Are certain job sectors statistically more at risk of replacing human jobs with AI across industries?

Hector Carvajal

Contribution: ½ of Introduction, Chi-Square test and Conclusion

hicarvajal@ucdavis.edu

Aditya Joshi

Contribution: Data Summary, Modeling, Evaluation and Confidence Intervals

adijoshi@ucdavis.edu

Department of Statistics, University of California, Davis

STA141: Fundamentals of Statistical Data Science

September 12th, 2024

Introduction

For the past decade, artificial intelligence (AI) has been replacing both blue-collar and white-collar jobs that involve straightforward and repetitive tasks. Some common occupations affected by AI taking over are in transportation services such as self-driving cars, self-driving delivery vehicles, and even truck driver jobs; warehouse jobs where AI now handles stocking, picking, and quality control; and in advanced fields such as marketing, programming and various other highly skilled areas. It is no secret that AI is flooding the job market, and taking over jobs once occupied by people. By now, most have at least heard of what AI is, and what it is capable of achieving, which has left workers and companies alike feeling uncertainty whether this is an essential resource in the workforce.

According to a research study done by the University of Michigan, analysts predict that AI could displace 85 million jobs by 2025, impacting both manual and highly skilled professions alike. This is why our goal is to identify whether a new job listing is at high risk of being automated in the next 10 years using classification methods from STA 141A. The scope of our research project is to create a model that can help us understand the impact AI is having on the job market today. The report will also explore the relationships between AI adoption, salary and automation risk across different industries using confidence intervals and the Chi-Squared test. The significance level for this project is $\alpha = 0.05$

Data Description and Summary

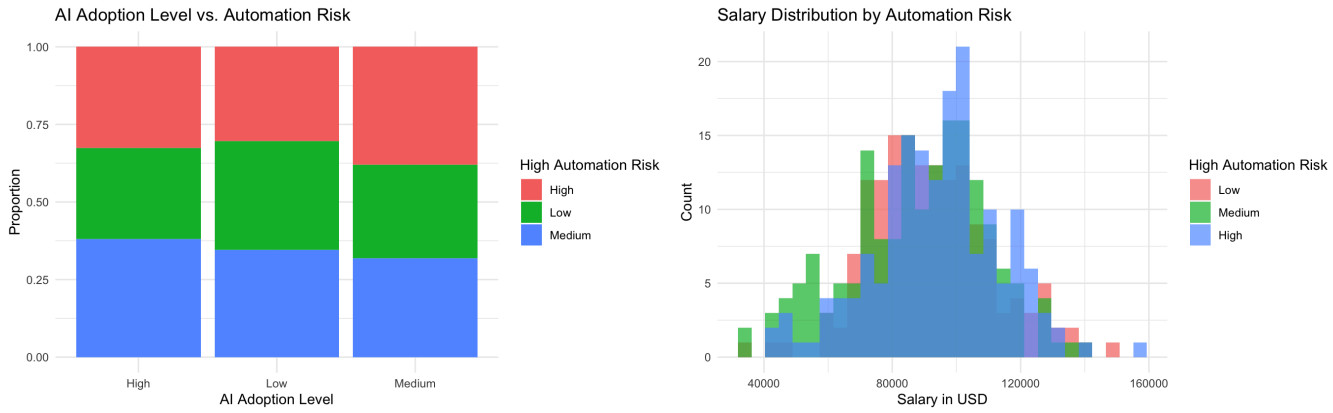
The dataset used for this study is titled, "AI-Powered Job Market Insights," and sourced from Kaggle, a platform that hosts different datasets for different questions. This dataset can be accessed at the following URL:

https://www.kaggle.com/datasets/uom190346a/ai-powered-job-market-insights/data

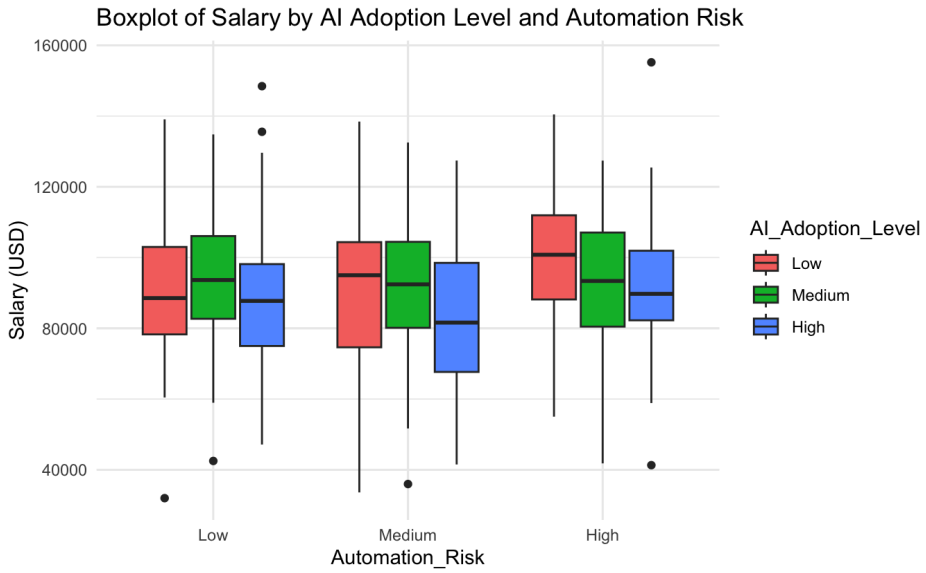
This dataset is a synthetic but realistic snapshot of the modern job market, particularly focusing on the role of artificial intelligence (AI) and automation across various industries. The dataset contains 10 columns:

- **Job_Title:** The title of the job role. Factor levels: Cybersecurity Analyst, Marketing Specialist, AI Researcher, Sales Manager, UX Designer, HR Manager, Product Manager, Software Engineer, Data Scientist, Operations Manager
- **Industry:** Categorical- The industry in which the job is located. Factor levels: Entertainment, Technology, Retail, Education, Finance, Transportation, Telecommunications, Manufacturing, Healthcare, Energy
- **Company_Size:** The size of the company offering the job. Factor levels: Small, Medium, Large
- **Location:** The geographic location of the job. Factor levels: Dubai, Singapore, Berlin, Tokyo, San Francisco, London, Paris, Sydney, New York, Toronto
- **AI_Adoption_Level:** The extent to which the company has adopted AI in its operations. Factor levels: Low, Medium, High
- **Automation_Risk:** The estimated risk that the job could be automated within the next 10 years. Factor levels: Low, Medium, High
- **Automation_Risk_High:** Binary variable indicating whether Automation_Risk is "High". This is the target column for the classification model.
- **Required_Skills:** The key skills required for the job role. Factor levels: UX/UI Design, Marketing, Project Management, JavaScript, Cybersecurity, Sales, Machine Learning, Python, Data Analysis, Communication

- **Salary_USD:** The annual salary offered for the job in USD. Range of $31969.52 to $155209.82
- **Remote_Friendly:** Indicates whether the job can be performed remotely. Factor levels: "Yes", "No"
- **Job_Growth_Projection:** The projected growth or decline of the job role over the next five years. Factor levels: "Decline", "Stable", "Growth"



The charts show that the dataset contains a relatively even distribution for the selected labels and most groups have a comparable number of entries which will help reduce the bias in a KNN model.



This boxplot shows a lot of variations in salaries for job listings at different levels of automation risk. One observation is that the median for job listings with high AI adoption level is always the lowest, regardless of automation risk. The subset of job listings with low AI adoption but high automation risk seem to have the highest salary.

Methodology

To prepare the dataset for analysis, first we created our target column: Automation_Risk_High which is a binary variable. That is, 1 if the observation is in fact at high risk of AI automation, and 0 otherwise. The classification fit will only predict whether a new observation is at high risk of automation or not instead of the “low”, “medium”, “high” factor levels in the original Automation_Risk column.

After first testing a logistic regression, we found that the model is only slightly better than randomly guessing, so instead of transforming the data, we decided to use the K-Nearest

Neighbor algorithm to fit the classification model. The analysis used a parameter tuning process that aimed to find the optimal number of neighbors (‘k’) for the model.

To validate the model, the dataset was split 70-30 for training and testing purposes. A range of ‘k’ values was experimented with, and were able to improve the model’s accuracy drastically reaching an accuracy rate of 85%.

Finally, we built and interpreted 95% confidence intervals to identify particular subgroups of interest with insights for Salary_USD.

Model Configuration

The KNN model is built using the following configuration:

- Number of Neighbors (k): 10
- Distance Metric: Euclidean distance
- Feature Normalization: Yes (Standardized using z-score)
- Training Set Distribution: 232 (Class 0 = Low or Medium Automation Risk), 119 (Class 1 = High Automation Risk)
- Cross-Validation: 10-fold cross-validation

Confusion Matrix

	Actual: 0	Actual: 1
Prediction: 0	94	15
Prediction: 1	5	35

Misclassification Analysis

Sensitivity	94.95%
Specificity	70.00%
Positive Predictive Value (Precision for Class 0)	86.24%
Negative Predictive Value (Precision for Class 1)	87.50%
Misclassification Rate	13.42%
95% CI for true accuracy	(0.8003, 0.916)

The selected KNN model demonstrates strong performance with a high overall accuracy (86.58%) and substantial agreement (Kappa = 0.6833) between predicted and actual classifications. The model's accuracy is significantly higher than the no information rate (NIR), which represents the accuracy that could be achieved by always predicting the most frequent class (66.44% in this case) and the baseline logistic regression model (with same number of predictors), which was only slightly more accurate than randomly guessing.

Confidence Intervals

The contrast confidence intervals for the salaries of certain subgroups are also worth investing in this dataset. Using R, the following confidence intervals can be calculated:

We are 95% confident that job listings with high levels of AI adoption in operations, on average, pay between $106.74 and $9004.70 less in annual salaries than job listings with medium levels of AI adoption in operations.

We are 95% confident that job listings with high levels of AI adoption in operations, on average, pay between $1152.20 and $10388.12 less in annual salaries than job listings with low levels of AI adoption in operations.

We are 95% confident that job listings with high risk of being automated within the next 10 years have a true average annual salary between $90890.03 and $96925.03

We are 95% confident that job listings with medium risk of being automated within the next 10 years have a true average annual salary between $84540.68 and $91032.21

We are 95% confident that job listings with low risk of being automated within the next 10 years have a true average annual salary between $89050.42 and $95174.49

Chi-Square Test

H_0**:** Adoption of AI does not significantly impact automation risk across industries.

H_A**:** Adoption of AI impacts the automation risk in at least one industry.

Data	X-Squared	Degrees of Freedom(df)	p-value
contingency_table	3.5046	4	0.4772

After performing a Chi-Square Test, our result revealed that there is no statistically significant relationship between AI adoption and automation risk across the different job sectors. Our results for p-value were 0.4772, which is greater than the standard significance level of 0.05. Because our p-value of 0.4772 is greater than the significance level, we fail to reject the null hypothesis and conclude that AI adoption does not significantly affect the chances of job automation.

The test statistic, X-Squared, was calculated as 3.5046, and this result measures the difference between the observed and expected frequencies in the contingency table. In our case, this value tells us that the observed frequencies do not deviate significantly from the expected frequencies. This further supports our conclusion that there is not a significant correlation between AI adoption and automation risk.

Conclusion

Our research aimed to answer the following question: "Are certain job sectors statistically more at risk of replacing human jobs with AI across industries?" We approached this question by using the K-nearest neighbor (KNN) algorithm to classify whether a job falls into the high automation risk category through binary classification. We used all columns in the dataset as predictors in order to identify the key factors that contributed to automation risk.

To better understand the differences in salary and automation risk, we analyzed the confidence intervals. Some of our findings suggest that jobs with higher AI adoption follow a pattern of offering lower salaries compared to those with medium or low AI adoption levels The overlap in the intervals suggest that different factors such as AI adoption and automation risk do not significantly change the salary ranges.

Finally we conducted a Chi-Square test to find other trends in our data, specifically to examine whether AI adoption had a significant impact on automation risk. From the results we gathered, we found strong evidence to conclude that there is no strong correlation between AI adoption and automation risk.

We conclude that while AI adoption is becoming more noticeable in many industries across multiple sectors, our findings suggest that the integration of AI in the workforce does not translate to an increased risk of job automation. This evidently shows that AI can be used in the workplace without posing the fear of being an immediate danger to job security.

R Appendix

```
set.seed(321)
library(dplyr)
library(glmnet)
library(nnet)
library(caret)
library(class)
library(ggplot2)
library(caTools)
data <- read.csv("data1.csv")
data <- data %>%
  mutate(Automation_Risk_High = ifelse(Automation_Risk == "High", 1, 0))

data$Automation_Risk <- factor(data$Automation_Risk)
data$Company_Size <- factor(data$Company_Size)
data$Location <- factor(data$Location)
data$AI_Adoption_Level <- factor(data$AI_Adoption_Level)
data$Required_Skills <- factor(data$Required_Skills)
data$Remote_Friendly <- factor(data$Remote_Friendly)
data$Job_Growth_Projection <- factor(data$Job_Growth_Projection)
```

```
data$Automation_Risk_High <- factor(data$Automation_Risk_High, levels = c(0, 1))

trainIndex <- createDataPartition(data$Automation_Risk_High,
                                  times=1,
                                  p = .7,
                                  list = FALSE)
train <- data[trainIndex, ]
test <- data[-trainIndex, ]

preProcValues <- preProcess(train, method = c("center", "scale"))
trainTransformed <- predict(preProcValues, train)
testTransformed <- predict(preProcValues, test)

knnModel <- train(
                  Automation_Risk_High ~ .,
                  data = trainTransformed,
                method = "knn",
                  trControl = trainControl(method = "cv"),
                  tuneGrid = data.frame(k = c(1,3,5,7,10))
                 )

best_model<- knn3(
            Automation_Risk_High ~ .,
```

```
                data = trainTransformed,
                k = knnModel$bestTune$k
               )

predictions <- predict(best_model, testTransformed,type = "class")
cm <- confusionMatrix(predictions, testTransformed$Automation_Risk_High)

ggplot(data, aes(x = Automation_Risk, fill = Automation_Risk)) +
  geom_bar() +
  labs(title = "Distribution of Automation Risk", x = "Automation Risk", y = "Count") +
  theme_minimal()

# Visualize the relationship between AI Adoption Level and Automation Risk
ggplot(data, aes(x = AI_Adoption_Level, fill = factor(Automation_Risk))) +
  geom_bar(position = "fill") +
  labs(title = "AI Adoption Level vs. Automation Risk", x = "AI Adoption Level", y =
"Proportion", fill = "High Automation Risk") +
  theme_minimal()
data$AI_Adoption_Level <- factor(data$AI_Adoption_Level, levels = c("Low", "Medium",
"High"))
data$Automation_Risk <- factor(data$Automation_Risk, levels = c("Low", "Medium",
"High"))
```

```
# Visualize the relationship between Salary and Automation Risk
ggplot(data, aes(x = Salary_USD, fill = factor(Automation_Risk_High))) +
  geom_histogram(bins = 30, alpha = 0.7, position = 'identity') +
  labs(title = "Salary Distribution by Automation Risk", x = "Salary in USD", y = "Count", fill
= "High Automation Risk") +
  theme_minimal()

ggplot(data, aes(x = AI_Adoption_Level, y = Salary_USD, fill = Automation_Risk)) +
  geom_boxplot() +
  labs(title = "Boxplot of Salary by AI Adoption Level and Automation Risk",
       x = "AI Adoption Level",
       y = "Salary (USD)") +
  theme_minimal()

ggplot(data, aes(x =  Automation_Risk, y = Salary_USD, fill = AI_Adoption_Level)) +
  geom_boxplot() +
  labs(title = "Boxplot of Salary by AI Adoption Level and Automation Risk",
       x = "Automation_Risk",
       y = "Salary (USD)") +
  theme_minimal()


data$Automation_Risk <- as.factor(data$Automation_Risk)
```

```
results <- data %>%
  group_by(Automation_Risk) %>%
  summarise(
    mean_salary = mean(Salary_USD, na.rm = TRUE),
    sd_salary = sd(Salary_USD, na.rm = TRUE),
    n = n()
  ) %>%
  mutate(
    error = qt(0.975, df = n - 1) * (sd_salary / sqrt(n)),
    lower_bound = mean_salary - error,
    upper_bound = mean_salary + error
  )

print(results)

high_ai_salaries <- data$Salary_USD[data$AI_Adoption_Level == "High"]
medium_ai_salaries <- data$Salary_USD[data$AI_Adoption_Level == "Medium"]
low_ai_salaries <- data$Salary_USD[data$AI_Adoption_Level == "Low"]

ci_high_medium <- t.test(high_ai_salaries, medium_ai_salaries,
                         alternative = "two.sided", conf.level = 0.95)

ci_high_low <- t.test(high_ai_salaries, low_ai_salaries,
```

```
                alternative = "two.sided", conf.level = 0.95)

cat("95% Confidence Interval for the difference in salary between High and Medium AI
adoption:\n")
print(ci_high_medium$conf.int)

cat("\n95% Confidence Interval for the difference in salary between High and Low AI
adoption:\n")
print(ci_high_low$conf.int)
summary(best_model)
cm
```

Sources

[1] L. Tharmalingam, “AI-powered job market insights,” Kaggle, https://www.kaggle.com/datasets/uom190346a/ai-powered-job-market-insights/data (accessed Sep. 12, 2024).

[2] L. Langreo, “What educators think about using AI in schools,” Education Week, https://www.edweek.org/technology/what-educators-think-about-using-ai-in-schools/2023/04 (accessed Sep. 12, 2024).