National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name: MANSI SHARMA
………………………………………………………………………………………

Student ID: x23410396@student.ncirl.ie
………………………………………………………………………………………

Programme: MSc Data Analytics ……………………………………… **Year:** 2025 -26………………

Module: Deep Learning and Generative AI

Lecturer: VIKAS SAHANI
……………………………………………………………………………

Submission Due Date: …20.11.2025…………………………………………………………………………
……

Project Title: DOMAIN APPLICATIONS
………………………………………………………………………………

Word Count: …………………………2817…………………………………………

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Mansi sharma
………………………………………………………………………………………
……

Date: …………20.08.2025……………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

I. [INSERT MODULE NAME]

II. [INSERT TITLE OF YOUR ASSIGNMENT]

Your Name/Student Number	Course	Date

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

III. AI ACKNOWLEDGMENT

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool

IV. DESCRIPTION OF AI USAGE

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

V. Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

VI. Additional Evidence:

[Place evidence here]

VII. Additional Evidence:

[Place evidence here]

Automation Risk in Jobs

MSc Data Analytics
Mansi Sharma
x23410396@student.ncirl.ie
National College of Ireland

Abstract—There is rising concern for job security not only because of increase in supply but, also decrease in demand for this reason there are people who are getting more into IR and computers but, there is another high risk of decrease of need because of automation. This project contains investigation which addresses this issue by integrating data which is based on salary and automation-risk. At times due to decrease in salary one may state that it causes less people to indulge and aim for that goal but, there is another factor all around which says AI is taking jobs, because of which people can be at risk of job loss, lay-offs and unemployment for long run of time. This project involves integrated data set which brings together many insights which can actually make difference in studying the market and preparing the individual for upcoming change all around the work. It uses random forests in order to determine the prediction with help of python language entire preprocessing and coding has been done.

Keywords— Random Forest, Machine Learning, Salary, automation risk, Python, feature engineering, Exploratory data analysis

INTRODUCTION

Automation was meant to make things easy and on the other hand it brings a bane to industry with elimination of jobs and involvement of machine learning and simulation with help of technology. Studying labour and job market is totally different when we compare the automation risk, but on the other hand it is aligned with the risk factor which appears to individuals who are aiming to learn and bring changes in a large domain area such as retail, logistics, transportation, manufacturing and administrative area. This project involves analysis and deep driving into findings based on real data sets which were merged. The data set involves 'automation' risk in different occupations it involves 702 rows and 54 columns in this. This data set involves automation probability in different occupations along with employment estimates in different US states. In this project the main target variable is Probability. Another data set in this project which is merged in order to deep drive into various parameters and study the main factor to risk jobs for upcoming individuals. The salary dataset involves salary and employment character of wide range of occupations. It involves 1394 rows and 20 columns. The main character in this data set is OCC, which is occupation code, titles, total employment estimates in it, hour and annual wage in it. The variables in this dataset serve as independent variable to study the prediction in automation risk. In the salary dataset it involves percentile wages which are 10th,25th, 75th percentile wages and these columns were removed because it may cause redundancy, contain more missing value, play less role in prediction of risk analysis and can cause noise in dataset and reduce the model performance as well.

Background and Scope

This project involves risk analysis in jobs due to automation; it may impact on jobs and related factors such as salary as well.

Artificial intelligence can be used in order to deal with repetitive patterns or some routine which is regular and contains algorithmic work, but if the work is irregular, inconsistent it remains a study to perform it with artificial intelligence. The scope of this project involves analysis of US occupations and market which involves automation in that market and key titles which are impacted with it. Integrated study of wages and salary with job category, develops a predictive model which involves automative analysis and deep diving into policy making and integration of system in order to make difference in job market.

GOALS AND BUSINESS VALUE

The goal is to find the main reason for automation risk in job. In this project it estimates and predicts the automation vulnerability with the help of machine learning models and examines how job characters such as salary and occupational titles. It involves numerical data, categorical data and textual dataset as well, it trains and evaluates the continuous variable with model performance which is been measured by room mean square which is RMSE. There is feature importance analysis as well, it contributes to technical aspect.

There are several business values which this project addresses. It includes strategic and policy making for workforce planning and future key development and scaling the salary as a main target in order to bring employment and understand the need of new roles and responsibilities with automation affect and involvement in business growth. There must be relief staffing and contribution of educated individuals who can deal with advent of emergency situations and break down of servers or power, there must be enough policy and laws to safeguard business as it is a huge investment and lots of scope and hopes are aligned with every business idea and actions in it. There can be a need of reskill of individual with retraining and upskilling with introduction of new technologies and investing in the right direction rather than just focusing on salary and qualification secured by any individual. There must be long-term hiring strategies for different roles and need for future stability. It also helps individuals with career goal decisions and long-term job security with advent of upcoming upgraded market and demand; it helps institutes as well to reshape the need of hour and build their curriculum which may help every individual for long term career and business goals.

Literature Review

In their seminal study, Frey & Osborne (2013) calculated that 47% of jobs in the U.S. were at high risk of automation, setting a marker for later work. This was followed by a more robust task-based analysis from the OECD, which reported that only 14% of jobs faced a high risk of automation, reasoning that jobs tended to offer a mixture of automatable tasks and those requiring a human touch. To deepen the understanding further, World Bank (2016) research pointed out that risk varies substantially across economies, as developing countries are at a higher risk with larger shares of routine-based jobs. In addition, Acemoglu & Restrepo (2018) introduced the term "so-so technologies," which emphasized that automation is not only technologically determined, but that economic and regulatory factors could partly slow displacement as well.

Although these studies yield essential macro-level information, a significant gap exists in forming granular predictive models that leverage a multitude of datasets for occupational-level risk estimation. This project directly addresses this gap. Unlike previous studies that rely solely on one-dimensional datasets, we integrate multiple datasets—including deep salary data and state-level employment data—into a single machine-learning model. This allows the model to move from a theoretical concept of exposure to a data-driven tool that identifies essential economic predictors, such as wage level, as part of their model and provides a better actionable, refined understanding of the vulnerability to automation for specific occupations.

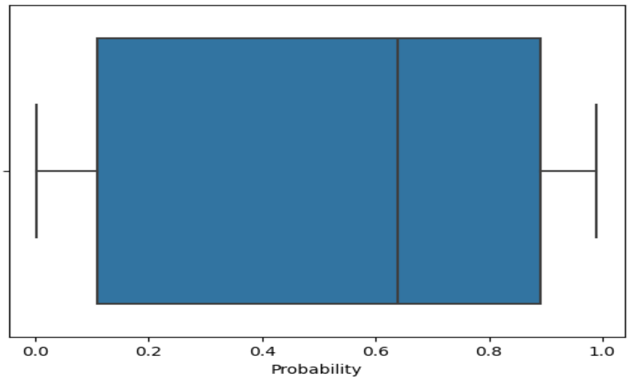
Ethical Concerns

The creation of predictive models that estimate automation risk in occupations involves complicated ethical considerations that outweigh technical accuracy. For example, one major consideration is the societal implications of job displacement and economic inequity, with low-skilled workers at the highest risk of unemployment and thus deepening our existing social divides if we do not have adequate policies in place, such as retraining and even social safety nets. In addition, we are interested in the technical integrity of predictive models, which may implicitly reflect societal injustices when trained on biased historical data, resulting in "algorithmic bias" toward certain demographic groups, as well as the need to collect and act on employment data must comply with data privacy regulatory frameworks, such as GDPR. And finally, we argue that acting responsibly on any predictions will be key; their ethical design will be to use them for strategic workforce planning and skills development never to justify mass layoffs without efforts or supports for those dislocated. Action that compromises human dignity or equal treatment (in the drive for technological efficiency) will undermine their positive use.

Methodology and Machine Learning Implementation

In this project the dataset is integrated analysis of dataset which are automation and salary of US state and involve probability as main factor in entire machine learning analysis. Data cleaning and preprocessing is done with help of python language, and the data has been loaded with help of libraries such as NumPy and pandas. The preprocessing involves removal of wage columns which may not contribute in findings but rather make the model complex and noisy and merging of dataset as well. It involves conversion of data type in code and analyzing with the help of random forest models in machine learning.

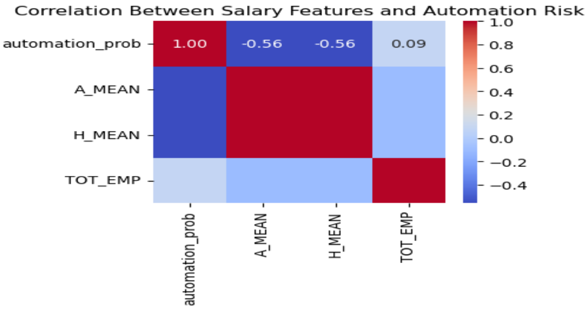
In order to proceed with the project, it was mandatory to study the main variable first which is Probability:



a) Boxplot – Probability Variable

The above is the boxplot which depicts automation probability across all occupations. The median probability lies around 0.65 and half of occupation have risk above 65% which state a large occupation may get affected by automation in this which shows need of reskill, work force planning and policy as well.

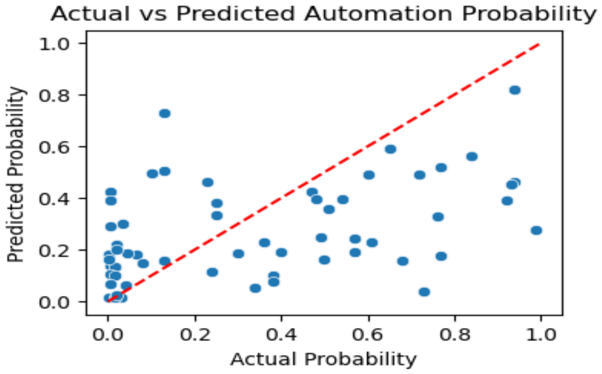
In this the one with 0.8 above can be drawn under mitigation strategies, median risk role for retraining and low risk training for stable areas for future work. It is to study skewness which is inclination of graph towards one side in order to learn the model. The state level employment column was dropped because it tells more about geographical condition rather than about the main occupation. Another is Heatmap:



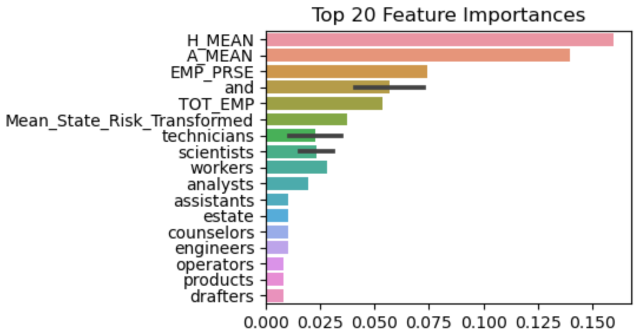
This is heatmap it shows relationship of salary and automation. Automation has inverse relations with salary; high paying jobs have low automation chances while total employees have no correlation with probability of automation.

Foe the machine learning model, there is random forest regression method which is been used in order to perform in this project. The reason of choosing this model in prediction is that:

Random forest regression handles complex relationships and non-linearity as well, it performs well with different types, a single decision tree easily overfits it reduce overfitting with many trees and averaging. It gave a strong baseline which shows better results. There is RMSE in this which is 0.2977 which is strong performance for probability performance.



In the above graph it shows actual probability, x-axis vs predicted on y-axis automation probability. The red line is perfect prediction line in the above graph, if it is linear it shows the predicted value is same as actual value. The scattered plots shows most points are in 0.0 and 0.4 on y-axis.this shows model predict low probability of automation even when actual probability is higher. Many points lies between the red line which show model is under predicting model risk.it shows that the model which is been performed is actually predicting low values only.



Top 20 influential features

Here in the above model high bar means more influence on model performance. In the above model it shows hourly mean wage has highest importance and generally high wage jobs are less automated as well as compared to routine jobs which have high automation risk with them. Also, jobs with fluctuations has more exposure to automation with them.

RMSE (Root Mean Squared Error) is a standard evaluation metric used in regression models. In this

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2}$$

This is formulae for RMSE in random forest regression, in this automation probability is 0 and 1 and the resultant is 0.2977 this means typical prediction error is 0.30 while in this it reflects the same moderately accurate. In the hope of improving predictive performance beyond the original Random Forest model (RMSE ≈ 0.2977); a tuned Random Forest Regressor was developed and subsequently assessed through 5-Fold Cross-Validation. K-Fold Cross-Validation divides the original dataset into five partitions or 'folds', and then the model is trained four times on four folds with the remaining fold used to test, ensuring that all observations can be used in both train and validation. K-Fold Cross-Validation provides more reliable and less biased insight into generalization performance than a single train-test split since it can account for randomness and dataset imbalance on sample variance. Here we have a mean cross validated RMSE of .307 with a low standard deviation of .026, suggesting performance was consistent across folds and that the model was able to generalize without overfitting any individual data subset. Compared to the original model, the tuned Random Forest provides better stability, more robust error estimation, and better interpretability through clearer feature importance rankings. In order to improve bushy pruning by increasing the number of trees, limiting the depth of trees, and tuning the minimum split and leaf, which restricted potential models, increasing estimation stability regarding predictive automation probability in a multidimensional labour-market dataset.

FINDINGS

In the above entire project, there are findings which are very crucial. The analysis revealed several important insights on how risk from automation differs across occupations and how different occupation characteristics affect that risk. First, exploration data analysis showed that, on average, probability of automation is high across many occupations, with a median risk level around 0.65--meaning more than half of jobs exposed in the dataset have a high risk of probability of automation. This suggests the need for preparation for workforce planning and training. The heatmap results reconfirmed the relationship between salary and probability of automation, demonstrating that lower-wage occupations are generally more automatable, while higher wages commonly requiring specialized and/or cognitive skills, were not automated or posed less risk of automation.

The Random Forest Regression model demonstrated that wage variables (H_MEAN and A_MEAN) had the highest predictive power for automation probability. This finding aligns with previous economic research that supports the idea that wage levels capture skill intensity and occupational complexity, both of which matter to automation vulnerability. Features related to employment, such as TOT_EMP and

EMP_PRSE, were also shown to make a meaningful contribution to the model; this suggests that occupations with high employment counts or less stable employment patterns are more sensitive to potential technological displacement. Finally, encoded occupational classifications (e.g., technicians, workers, engineers, scientists) contributed to predictive value, suggesting that specific job classifications and the structure of the tasks they perform help determine a worker's exposure to automation.

Model evaluation indicated reasonable prediction performance in the initial model, with an RMSE of 02977 on the test set. The scatter plot of actual vs predicted values indicates the model expressed an under-prediction of automation risk for some occupations- especially for occupations with a greater true probability. To enhance reliability, we evaluated a tuned Random Forest model using 5-Fold Cross-Validation, producing a mean RMSE of 0.307 and a mean standard deviation of 0.026. This indicates a relatively consistent level of error between folds, and shows this model predicts as opposed to overfitting the model to more specific subsets of the data. Overall, results confirm the multi-factorial nature of occupational automation risk, that is strongly associated with wage levels and job categories; predictability of automation risk using machine-learning techniques was moderate, yet meaningful.

Conclusion

This project effectively merged automation-risk data with salary and occupational characteristics, resulting in the creation of a machine-learning model capable of predicting automation risk at the occupation level. The Random Forest model was impressive in recognizing important patterns, supporting that wage indicators and task categorization are better predictors of automation risk. From exploratory data analysis, we confidently concluded there was considerable job exposure, namely with low-wage, routine-tasks; but higher-skilled jobs with greater pay exhibited some resilience to automation.

The Random Forest model also improved the validity of the predictive outputs. By embedding a tuned Random Forest model with standard 5-Fold Cross-Validation we provided stable and generalizable accuracy estimates of the predictiveness of the automation outputs. Overall, despite RMSE measures indicating that predicting probability of risk for automation inability is inherently difficult to predict accurately, the model is reliable and straightforwardly interpretable; thus still useful analytically. The outputs continued a recent surge of literature (i.e., studies) to help provide insights for policymakers and educators or employers to respond to expected and potential technological disruption by developing targeted reskilling, updating curriculum, and developing strategies to support an anticipated long-term workforce change. Ultimately, predicting automation risk at this unit-level empowers institutions and people to be able to assess any real or potential changes in the labor landscape, and the ongoing assessment of automation technology with workforces.

REFERENCES

[1] C. B. Frey and M. A. Osborne, “The future of employment: How susceptible are jobs to computerisation?” Technological Forecasting and Social Change, vol. 114, pp. 254–280, Jan. 2017.
[2] M. Arntz, T. Gregory, and U. Zierahn, “The risk of automation for jobs in OECD countries: A comparative analysis,” OECD Social, Employment and Migration Working Papers, no. 189, OECD Publishing, 2016.
[3] D. Acemoglu and P. Restrepo, “The race between man and machine: Implications of technology for growth, factor shares, and employment,” American Economic Review, vol. 108, no. 6, pp. 1488–1542, Jun. 2018.
[4]World Bank, “World Development Report 2016: Digital Dividends,” World Bank Group, Washington, DC, 2016.