AI Impact on Job Market: Predicting Increasing vs. Decreasing Jobs (2024–2030)

ITCS 3156 Final Project Report

Haida Makouangou

Table of Contents

1. Introduction

a. Problem Statement

AI is changing the job market. Organizations adopting AI-driven tools are projecting certain occupations to grow, while others may decline due to automation, digital transformation, and shifting skill requirements.

This project investigates whether supervised machine learning models can accurately predict whether a job is expected to increase or decrease between 2024 and 2030, using a dataset containing over 30,000 job records with economic, demographic, and occupational features.

The central research question is:

Can machine learning models distinguish between jobs likely to grow versus those likely to decline under AI adoption?

b. Motivation

Understanding AI-induced labor transitions benefits multiple stakeholders:

- **Students** who want to enter fields aligned with future demand;
- **Students** who want to enter fields aligned with future demand;
- **Workers** planning reskilling strategies to stay competitive;
- **Organizations** forecasting talent and employment needs;
- **Policymakers** evaluating the broader economic impacts of automation.

Accurate predictions can help make better choices about education, workforce development, and economic policy.

c. Approach

My approach follows a complete machine learning pipeline:

- Data exploration and visualization
- Preprocessing (cleaning, encoding, scaling, splitting);
- Training two ML algorithms:
  - Logistic Regression
  - Random Forest;
- Evaluating performance using accuracy, F1-score, and confusion matrices;
- Analyzing feature importance to understand labor-market drivers;
- Choosing the best model and testing it on unseen data.

2. Data

a. Introducing the Dataset

The dataset includes 30,000 job records and 13 features, such as:

- Median Salary (USD)
- Automation Risk (%)
- Projected Openings (2030)
- Gender Diversity (%)
- Remote Work Ratio (%)
- Required Education Level
- Industry Sector
- Current Job Openings (2024)

A binary feature named Job Status Binary was created to represent:

- 1 = Increasing jobs
- 0 = Decreasing jobs

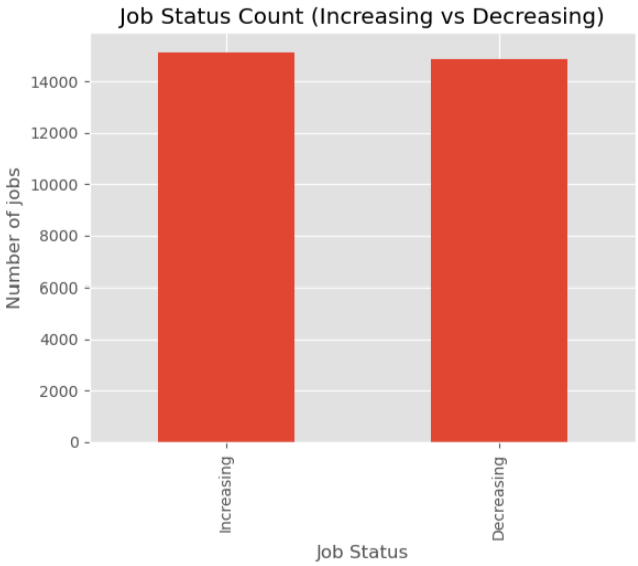
This structure meets the project's requirement of ≥10,000 samples and ≥5 features.

b. Visual Analysis of Data

The exploratory analysis provides several insights relevant to modeling.

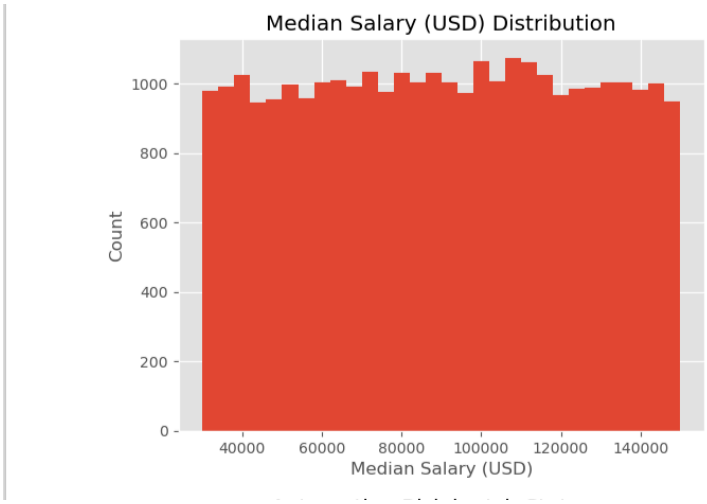
Class Distribution

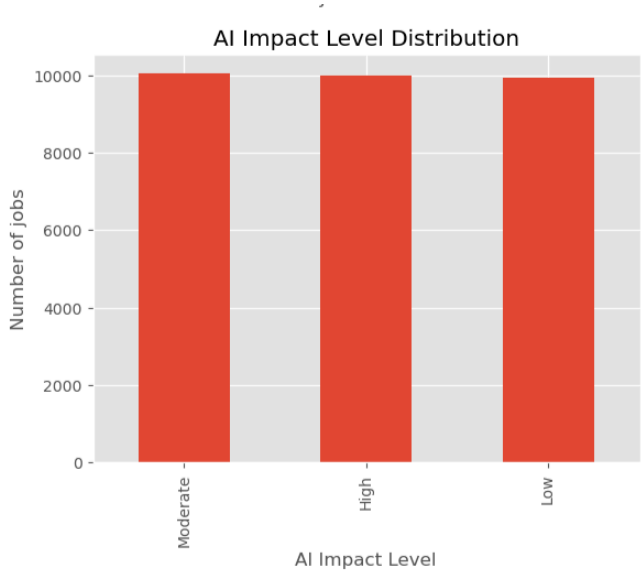
The dataset is balanced, with comparable counts of increasing and decreasing jobs. This balance is advantageous for training classification models without requiring special adjustments for class imbalances.



Feature Distributions

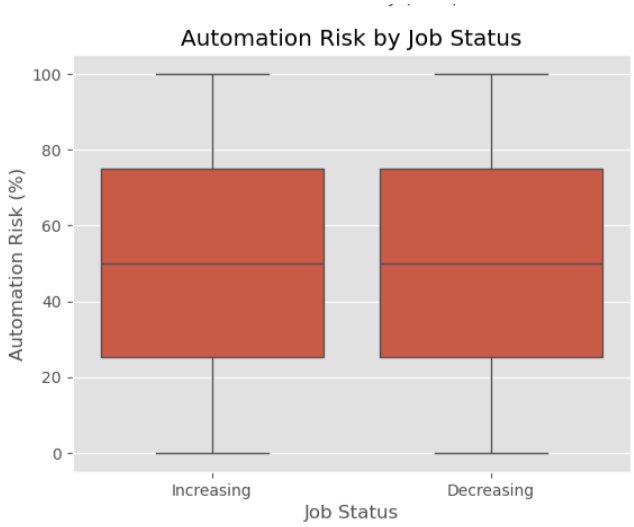
Plots of salary, AI impact level, automation risk, and projected openings reveal substantial variability across occupations, suggesting a rich feature space for predictive modeling.





Automation Risk Patterns

Higher automation risk is more frequently associated with jobs labeled as *decreasing*, a relationship consistent with existing labor market research. Conversely, jobs predicted to grow often exhibit favorable economic indicators, such as strong projected openings and higher median salaries.



These observations support the hypothesis that economic and structural job characteristics provide beneficial, but potentially noisy, signals for predicting labor trends.

c. Data Preprocessing

To prepare the dataset for modeling, the following steps were taken:

- **Target Encoding**

  Binary encoding of job status for classification.

- **Feature Selection**

  Numeric features included salary, openings, experience, automation risk, remote work ratio, and gender diversity.

  Categorical features included job title, industry, location, required education, and AI impact level.

- **Splitting Strategy**

  The data were divided into training (60%), validation (16%), and test (20%) subsets using stratified sampling to preserve class proportions.

- **Scaling and One-Hot Encoding**

  A StandardScaler standardized numeric features, while OneHotEncoder transformed categorical variables with handle_unknown="ignore" to prevent errors during inference.

- **Pipelines**

  All preprocessing and model training were encapsulated in Pipeline objects, ensuring clean, repeatable transformations that adhere to best practices.

3. Machine Learning Methods

a. Logistic Regression

Logistic Regression models the log-odds of the target class as a linear combination of the features. It is commonly used for binary classification and serves as a strong baseline model.

Strengths

- Highly interpretable
- Efficient on large datasets
- Performs well with linearly separable data

Limitations

- Struggles with nonlinear relationships
- Susceptible to underfitting when feature interactions are complex
- Sparse, high-dimensional, one-hot-encoded features degrade performance.

For this project, the model was trained with max_iter=1000 to ensure convergence.

b. Random Forest Classifier

Random Forest is an ensemble method that constructs multiple decision trees using bootstrap samples and random feature subsets.

Strengths

- Captures nonlinearities and interactions
- Robust to noise and outliers
- Provides interpretable feature importance rankings

Limitations

- Can overfit without tuning
- Less interpretable than linear models
- Sensitive to high-cardinality categorical input domains

This model used 200 estimators and a fixed random state for reproducibility.

4. Results

a. Experimental Setup

Both models were trained on identically preprocessed data and evaluated on validation and test sets. Standard classification metrics were used:

- Accuracy

- Precision, Recall, F1-score
- Confusion Matrix

For comparison, a majority-class baseline yielded an accuracy of approximately 0.50.

b. Model Accuracies

Logistic Regression

- Training Accuracy: 0.9759
- Validation Accuracy: ≈ 0.50

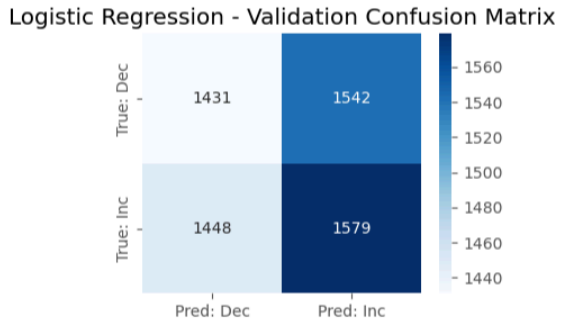
The substantial gap between training and validation accuracy indicates underfitting and suggests that the linear model cannot capture essential decision boundaries in the data.

```
===== Logistic Regression Evaluation =====
Train Accuracy: 0.5795555555555556
Validation Accuracy: 0.5016666666666667

Classification Report:
              precision    recall  f1-score   support

  Decreasing       0.50      0.48      0.49      2973
  Increasing       0.51      0.52      0.51      3027

    accuracy                           0.50      6000
   macro avg       0.50      0.50      0.50      6000
weighted avg       0.50      0.50      0.50      6000
```



Random Forest Classifier

- Training Accuracy: 1.00
- Validation Accuracy: ≈ 0.49
- Test Accuracy: 0.4983

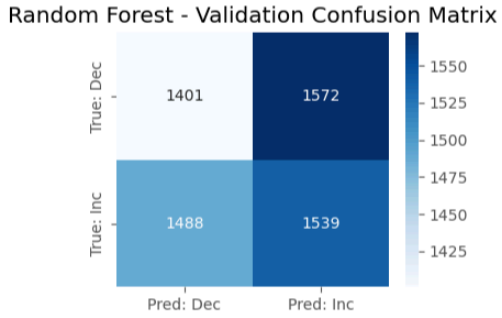
Random Forest achieves perfect training accuracy but performs no better than random guessing on unseen data. This pattern reflects overfitting and suggests limited predictive structure in the feature set.

```
===== Random Forest Evaluation =====
Train Accuracy: 1.0
Validation Accuracy: 0.49

Classification Report:
              precision    recall  f1-score   support

  Decreasing       0.48      0.47      0.48      2973
  Increasing       0.49      0.51      0.50      3027

    accuracy                           0.49      6000
   macro avg       0.49      0.49      0.49      6000
weighted avg       0.49      0.49      0.49      6000
```



c. Model Analysis

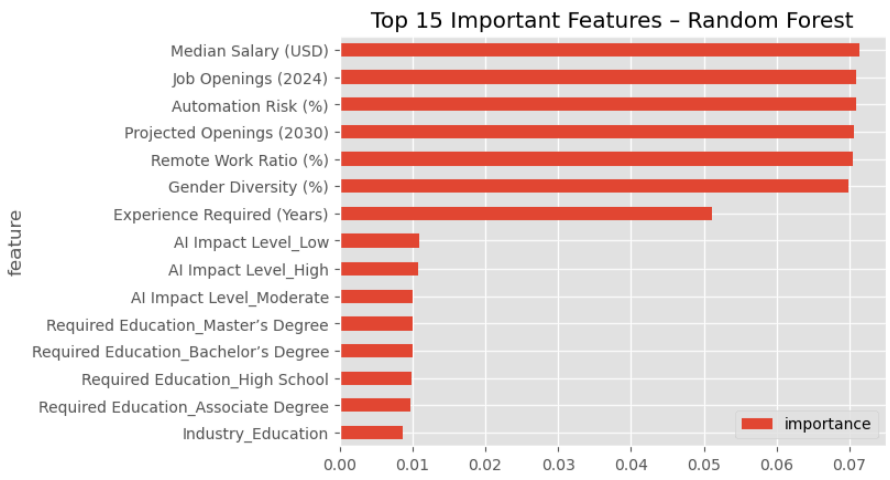
The confusion matrices for both models show nearly symmetric misclassification rates for increasing and decreasing jobs, consistent with an overall validation accuracy of ~50%. This indicates that neither model identified separable feature patterns strong enough to reliably differentiate the two classes.

Feature Importance

The Random Forest model highlights several features with comparatively higher influence:

- Median Salary
- Job Openings (2024)
- Automation Risk
- Projected Openings (2030)
- Remote Work Ratio
- Gender Diversity

Although economically intuitive, these features still lack sufficient signal to enable strong predictive performance.



5. Conclusion

a. Closure

This project explored whether supervised machine learning models could predict the direction of job growth amid increasing AI adoption. The project successfully illustrated the complete lifecycle of a predictive modeling pipeline, encompassing exploratory data analysis (EDA), preprocessing, model development, and evaluation, despite both logistic regression and random forest classifiers achieving only baseline-level accuracy.

The findings suggest that the existing feature set is insufficient to accurately forecast job market dynamics, implying that occupational trajectories influenced by AI require more detailed or temporally grounded data.

b. Challenges

The primary challenges encountered include:

- Weak predictive signal: Many occupations with similar characteristics have different growth trajectories.
- High-cardinality categorical variables, such as job titles and locations, result in sparse one-hot encodings.
- A tension between underfitting and overfitting: Logistic Regression tends to underfit complex patterns, while Random Forest tends to overfit yet still fails to generalize.

These challenges highlight the importance of feature engineering and richer datasets for modeling labor market trends.

c. Future Work

Future extensions of this project could pursue:

- Hyperparameter tuning (e.g., GridSearchCV) for both models;
- Feature engineering, such as industry grouping or occupational clustering;
- Incorporation of additional data sources, including historical employment trends, macroeconomic indicators, or time-series variables;
- Evaluation of advanced algorithms, such as XGBoost, Support Vector Machines, or Gradient Boosting;
- Dimensionality reduction techniques to mitigate the impact of high-cardinality categorical variables.

These enhancements may reveal deeper patterns governing job growth during the AI transition.

6. References/Acknowledgements

Dataset

Islam, Sahil. "AI Impact on Job Market: Increasing vs. Decreasing Jobs (2024–2030)." *Kaggle*, 2025, www.kaggle.com/datasets/sahilislam007/ai-impact-on-job-market-20242030.

Resources

"sklearn.linear_model.LogisticRegression." *Scikit-learn Documentation*, scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

"sklearn.ensemble.RandomForestClassifier." *Scikit-learn Documentation*, scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

"sklearn.compose.ColumnTransformer." *Scikit-learn Documentation*, scikit-learn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html.

"pandas.DataFrame." *Pandas Documentation*, pandas.pydata.org/docs/reference/frame.html.

Waskom, Michael. "Seaborn: Statistical Data Visualization." *Seaborn*, seaborn.pydata.org/.

Lee, Minwoo. *ITCS 3156 Course Materials: Intro to Machine Learning*. UNC Charlotte, 2025.

Acknowledgement:

I have reused portions of my previous coursework from ITCS 3156, specifically the machine learning pipelines and evaluation patterns from Module 2 – ML Basics, Module 8 – Logistic Regression, and Module 9 – Random Forest homework notebooks. Aside from these reused components, I hereby declare that I have not used any resources other than the ones listed in the References section of this report.

7. Source Code

https://github.com/HaidaMarese/ai-job-market-ml-project