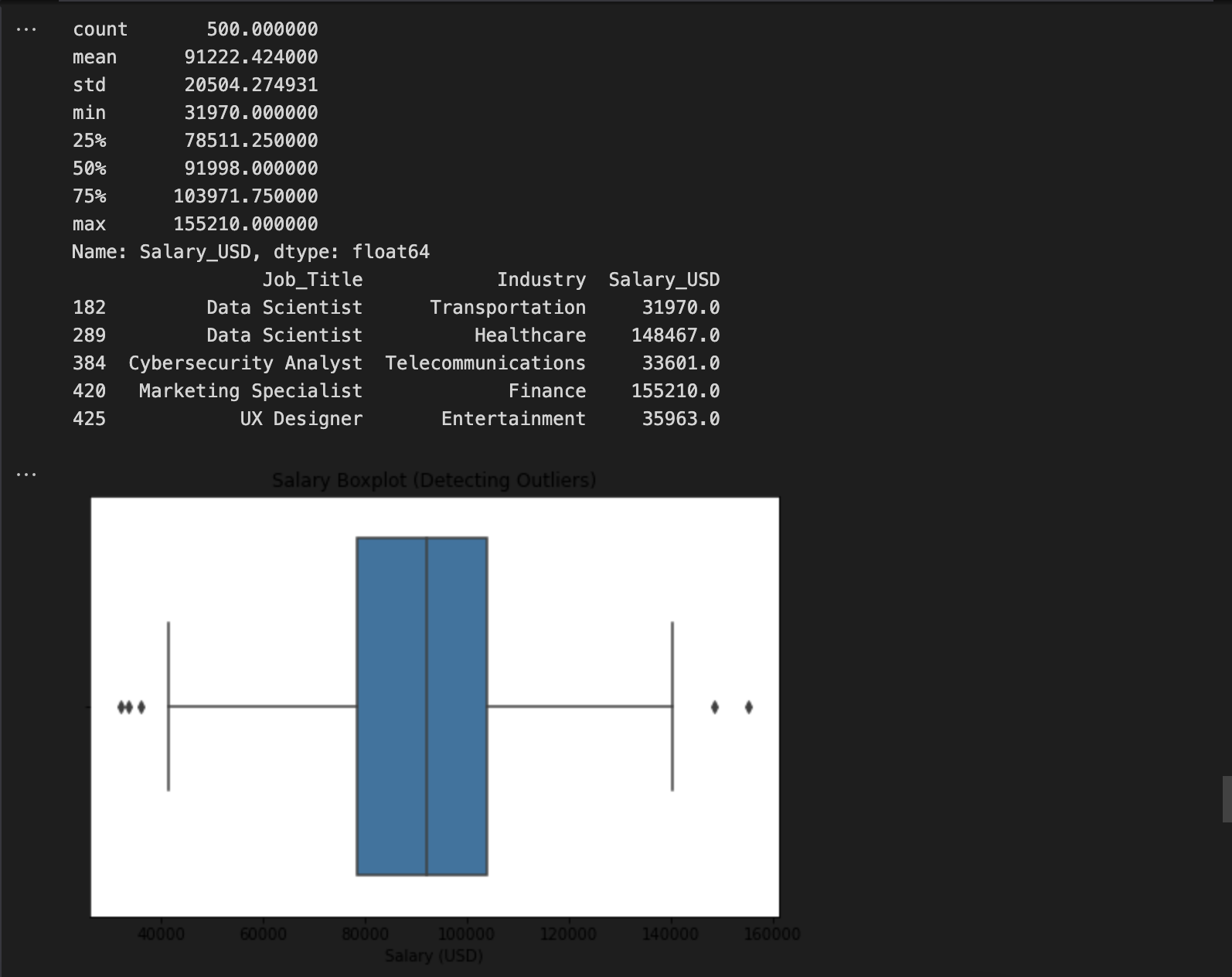
AI vs JOb Market

Anna Bottu

MSCS-634: Advanced Big Data and Data Mining

The dataset I used was the “AI-Powered Job Market Insights” dataset from Kaggle (https://www.kaggle.com/datasets/uom190346a/ai-powered-job-market-insights), which contains 500 records and 10 attributes. The attributes include Job Title, industry, company size, location, ai adoption level, automation risk, required skills, salary (USD), remote friendliness, and job growth projection. I chose this dataset because AI’s impact on the future of work is widely discussed topic, and I wanted to explore which jobs are most at risk of automation, how AI adoption caries across industries, and what skills are becoming more critical as AI continues to evolve.

During preprocessing, there were not a lot of missing values and if there were they were handled through imputation. Categorical variables such as Industry, Job\_Title, and AI Adoption Level were transformed using one-hot encoding, while numeric features were standardized with scaling to enhance model training. The exploratory data analysis revealed several key patterns such as salary distribution was right-skewed, with the highest salaries concentrated in the Technology and Finance industry, particularly in retail, showed higher automation risk, technology roles were more likely to offer remote work options compared to healthcare or retail, and larger companies demonstrated greater AI adoption levels than smaller firms. When it comes to feature engineering, new variables were created by grouping required skills into technical, managerial, and creative categories, and by constructing a composite risk index that combined automation risk with AI adoption.



A graph of different colored bars

AI-generated content may be incorrect.

The modeling phase produced several important insights across regression, classification, clustering, and association rule mining. For regression, the target variable was salary, and models such as Linear Regression, Ridge, and Lasso were applied. Results indicated that AI adoption level, industry, and job role were the strongest predictors of salary. In the classification task, the target variable was job growth projection (growth vs. decline), and both Logistic Regression and Random Forest were tested. The Random Forest model outperformed Logistic Regression with an accuracy of approximately 85%, highlighting that jobs projected for growth were strongly associated with AI-intensive roles in the technology sector. Clustering with K-means revealed three distinct job groupings: high-salary, high-growth technology roles such as AI Researchers and Cybersecurity specialists, mid-salary, moderate-growth roles like Sales Managers, and low-salary, high automation-risk positions such as those in retail. These clusters reinforced the growing polarization between high-skill and automatable jobs. Association rule mining using Apriori and FP-Growth showed patterns linking job characteristics to outcomes. One notable rule showed that when AI adoption was high and roles were remote-friendly, salaries were more likely to exceed $100,000. This analysis demonstrated that remote-friendly, AI-driven jobs are both more profitable and more resilient to automation.

A graph with a line drawn on it

AI-generated content may be incorrect.

Based on the findings, several recommendations can help different stakeholders prepare for the changing job market. Policymakers should prioritize retraining and reskilling programs for workers in at-risk positions, such as retail and clerical roles, to help them transition into more sustainable careers. Companies can play a role by investing in AI focused upskilling opportunities for their employees and adopting hybrid work models that keep them competitive while supporting flexibility. For job seekers the key will be developing both technical and cross disciplinary skills such as AI, data analytics, cybersecurity, and project management that are in high demand and harder to automate. Finally, educators should work to align their curricula with the changing workforce needs, placing greater emphasis on AI literacy, automation resilience, and digital collaboration skills that will prepare students for future opportunities.

When it comes to data privacy, this project used a man-made dataset, which means the information isn’t tied to real people and the risk is very low. In real-world situations it’s important to keep data anonymous and handle it securely. Fairness is another concern since some industries like healthcare, or the arts are not well represented in the data, which could affect the accuracy of the results. There’s also the risk of bias in the models themselves. For example, a classification model might label certain jobs as declining which could create negative perceptions. To address this, I made sure the training sets were balanced so that one class didn’t overpower the other, and I also reviewed which features had the most influence to check for hidden biases. Transparency was also a priority; I used visualizations and explainable models so that the results could be understood clearly by anyone reviewing the findings.