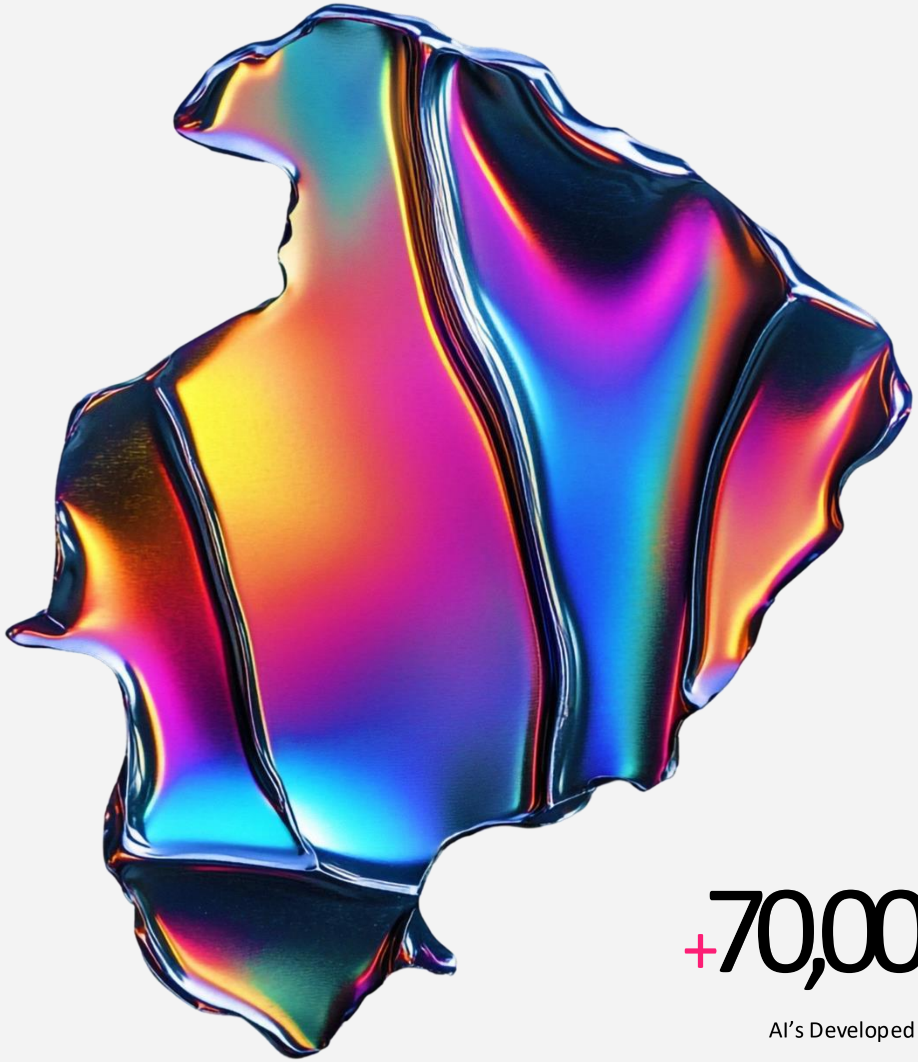


Automated out

Automated Out: Forecasting Which Careers Face the Greatest AI Disruption

Tsepo Tsolo-mohlakwana | DT480 – Research Project in Data Science

+70,000

AI’s Developed

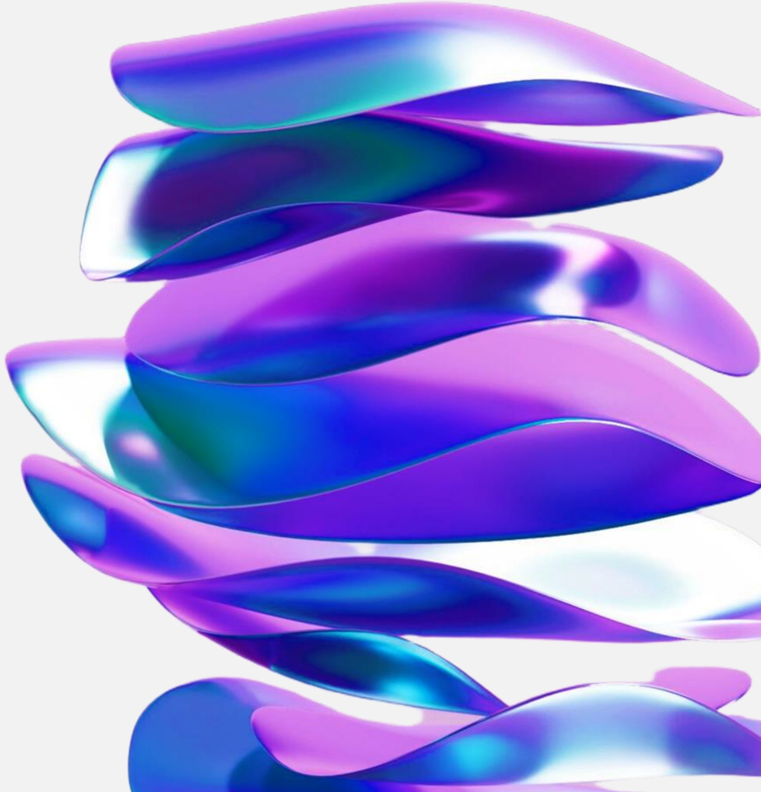
Let’s Begin

Why this matters

Context – Why This Matters

- AI is now deeply embedded in business
- 80% of executives report AI use (Gartner, 2023)
- 75% of knowledge workers use AI tools (Microsoft, 2024)

300+



Problem statement

- Existing tools are too broad or qualitative
- No job-title-level forecasts
- Static data overlooks change

Broad

Qualitative

Objectives

- Forecast job-specific automation risk
- Visualize industry trends
- Support upskilling and career planning
- Contribute to AI risk research

Goals+

The project had four main goals. First, build a classification model to assess job-level automation risk. Second, map these risks across industries. Third, offer insights to guide reskilling and education. And finally, contribute meaningfully to academic and applied discussions on AI's workforce impact.

Literature Review – Global Trends

AI is advancing fast in the U.S.
Healthcare, finance, manufacturing heavily impacted

"Routine and repeatable work" most at risk

DEEP EXPLANATION

DATA DRIVEN: THE POWER OF AI

Literature Gaps

- Reliance on surveys (OECD, McKinsey)
- No predictive modeling
- Lack of real-time insights

Methodology Overview

- Merged 3 datasets
- Applied SMOTE for balance
- Trained 3 ML classifiers
- Evaluated with F1-score, AUC

My process followed a robust machine learning pipeline: data gathering, cleaning, feature engineering, balancing, training, and evaluation. Three models were tested Decision Tree, Random Forest, and XGBoost—with XGBoost delivering the best results

Task load
Adoption
Automation Risk

Data sources

- AI task load (AI_risk.csv)
- Industry-level AI adoption
- Automation risk metrics by job

$328,548 B

US investment into AI

Preprocessing & Cleaning

Removed

Null Columns

•Removed 100% null columns

Filled

Missing values

•Filled missing values

Encoded

Labels

•Encoded labels

Converted

Percentage to decimals

•Converted percentages to decimals

FEATURE ENGINEERING

- AI_Impact_Percentage = AI Models / Tasks
- Industry risk mapping: High (1), Medium (0), Low (-1)

I added calculated fields like AI_Impact_Percentage to represent how automation-heavy a job is. I also manually assigned industry risk levels, helping the model understand which sectors are more vulnerable to disruption.

Jupiter

Class Imbalance Challenge

- Few samples for rare job roles
- SMOTE improved model fairness
- Enabled minority class recognition

SMOT

Synthetic Minority Oversampling Technique which generates new examples for underrepresented classes.

MORE

As a result, the model could better detect high-risk but rare jobs like data entry clerks.

Model Comparison

```
XGBoost Accuracy: 78.54 %

Classification Report:
              precision    recall  f1-score   support

           0       0.55      0.54      0.54       144
           1       0.87      0.94      0.90       144
           2       0.66      0.55      0.60       165
           3       0.92      0.97      0.95       165
           4       0.79      0.87      0.83       134
           5       0.59      0.32      0.41       149
           6       0.85      0.92      0.88       151
           7       0.91      0.97      0.94       130
           8       0.66      0.55      0.60       154
           9       0.80      0.85      0.83       131
          10       0.47      0.43      0.45       130
          11       0.91      0.94      0.92       134
          12       0.80      0.91      0.85       149
          13       0.57      0.55      0.56       134
          14       0.95      0.98      0.96       139
          15       0.64      0.71      0.67       133
          16       0.95      0.98      0.97       130
          17       0.95      0.97      0.96       148
          18       0.78      0.78      0.78       133
          19       0.90      0.91      0.90       170
          20       0.69      0.61      0.65       161
          21       0.66      0.54      0.59       142
          22       0.73      0.81      0.77       152
          23       0.72      0.63      0.67       156
          24       0.78      0.76      0.77       152
          25       0.95      0.99      0.97       142
          26       0.76      0.83      0.79       153
          27       0.75      0.85      0.80       135
          28       0.89      0.95      0.92       122
          29       0.81      0.80      0.80       136
          30       0.80      0.92      0.85       143
          31       0.83      0.89      0.86       147

    accuracy                           0.79      4608
   macro avg       0.78      0.79      0.78      4608
weighted avg       0.78      0.79      0.78      4608
```

```
Decision Tree Model
Best Params: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2}
Accuracy: 35.07 %
Feature Importances: {'Tasks': np.float64(0.2251534420977151), 'AI models': np.float64(0.3152389483080803), 'AI_Wo
rkload_Ratio': np.float64(0.196732640480446), 'AI_Impact_Percentage': np.float64(0.2628749691137586), 'Industry_Ri
sk_Level': np.float64(0.0)}

Random Forest Model
Best Params: {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}
Accuracy: 40.72 %
Feature Importances: {'Tasks': np.float64(0.25081478032460686), 'AI models': np.float64(0.26302337357711625), 'AI_
Workload_Ratio': np.float64(0.24021025144766653), 'AI_Impact_Percentage': np.float64(0.24595159465061048), 'Indust
ry_Risk_Level': np.float64(0.0)}
```

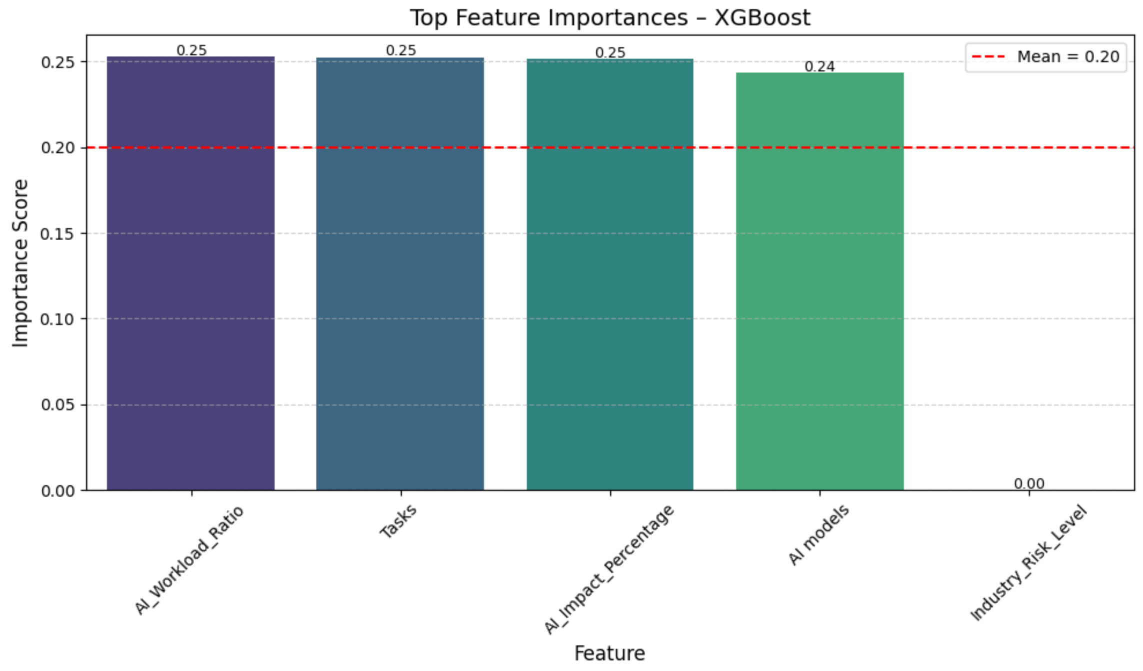
- XGBoost performed best (F1 = 0.78)
- Decision Tree, Random Forest less generalizable
- AUC scores ranged 0.78–0.92

AI-DRIVEN

Feature importance

- Top predictors: AI_Workload_Ratio, Task Count
- Human traits like creativity missing

AI_Workload_Ratio was the most influential feature, followed by Task Count. Unfortunately, the dataset lacked variables like creativity or problem-solving, which are key in automation resistance. Future versions of this model will aim to include those.



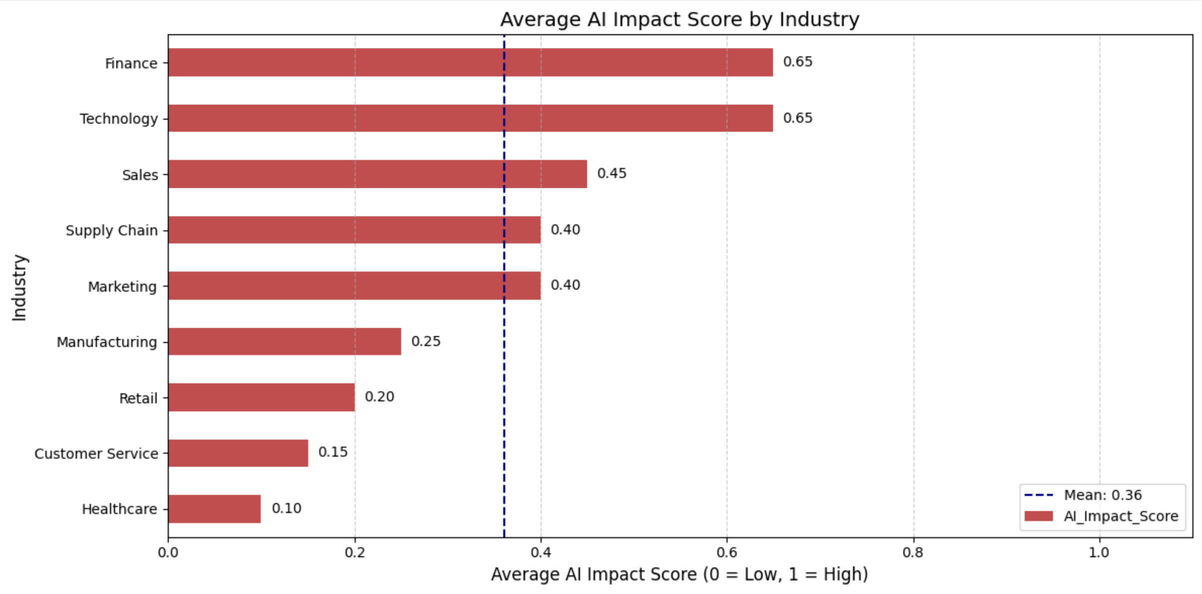
Industry-Level Risk Patterns

- Healthcare, Customer Service= Low Risk
- Supply chain, Marketing = Medium risk
- Technology, Finance = High Risk

By mapping jobs to industries, I found clear risk trends. As expected, repetitive, rule-based jobs in supply chain and customer service showed high risk. More cognitive roles in marketing or IT were safer. This supports claims by the BLS and Stanford HAI

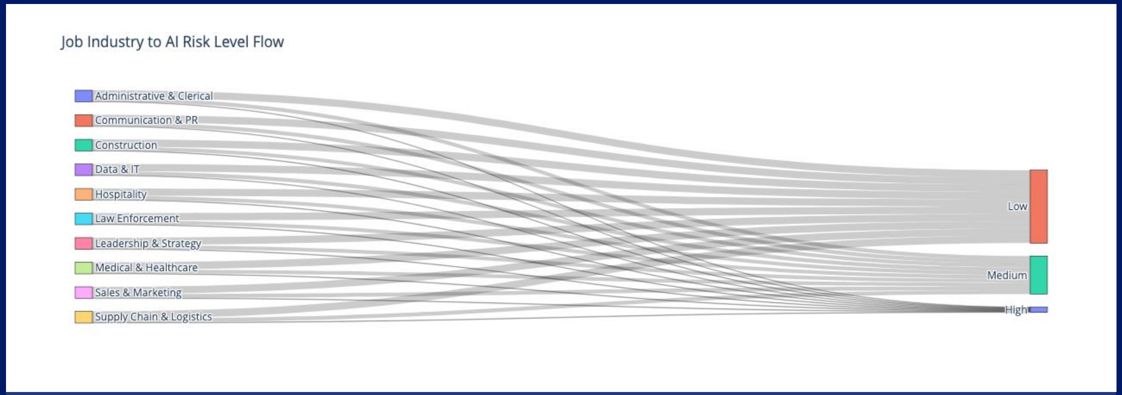
400-800M

Jobs could be displaced by 2030(McKinsey and Company)

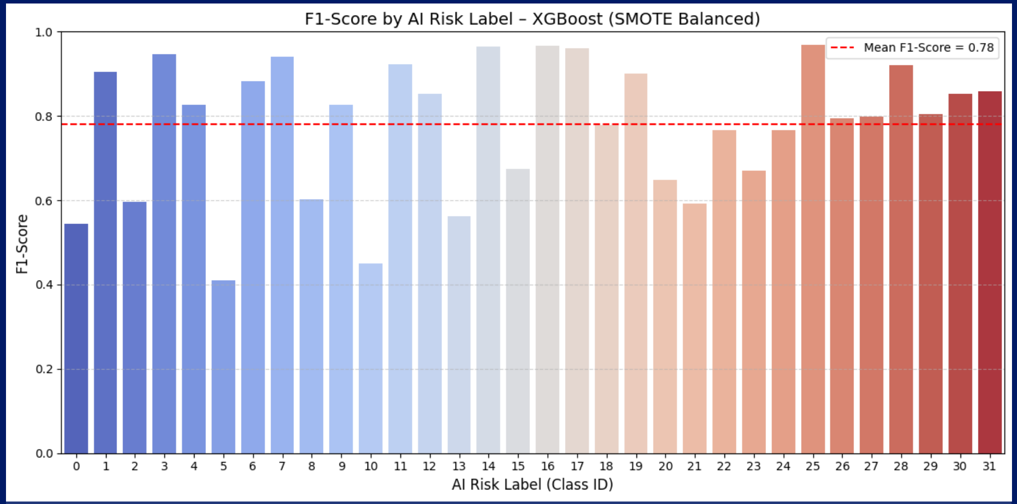


Diagrams

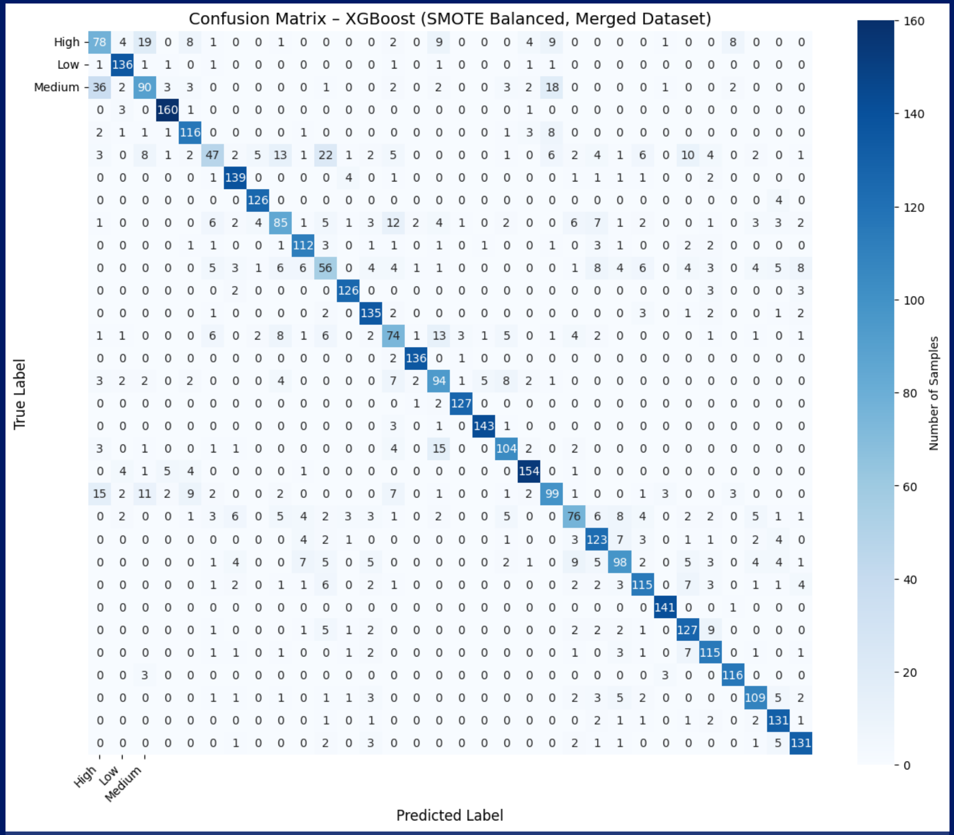
Sankey: Job → Industry → Risk



F1-scores highlight class precision



Confusion matrix shows misclassifications



visualization

Real World Implications

- Guides workforce development
- Informs policy and HR strategy
- Useful for students and career changers

YEARLY-JOB MARKET SHIFT

This isn't just academic. Career coaches could use this tool to help students choose resilient careers. HR leaders can assess where to invest in AI augmentation versus human skill development. And governments can use the findings to design reskilling policies.

5+

- Outdated datasets
- Missing soft skills
- Model interpretability (black-box issue)

Challenges & Limitations

There were challenges. First, data timeliness: most sources lack market Variety. Second, the absence of soft skills made it hard to model creativity-driven roles. Lastly, XGBoost is powerful but hard to explain. Tools like SHAP and LIME could address that in future versions.

Future Work

54+

I plan to build a real-time, interactive version of this tool accessible to job seekers, HR leaders, and educators. It would offer custom risk scores based on up-to-date data, job location, and skill profile.

2026 BEYOND

OUR HANDSOUR FUTURE IS IN OUR HANDS

1. Connect to live job APIs
2. Add cognitive and soft skill indicators
3. Build a web-based risk calculator
4. Regional customization

- AI disruption is real and unequal
- Predictive tools are needed
- XGBoost-based model performs reliably

Conclusion

Q&A / Closing

Thank you

The END