Decoding Income: Data-Driven Solutions for Economic Equity

Adel Hassen
January 30th, 2025

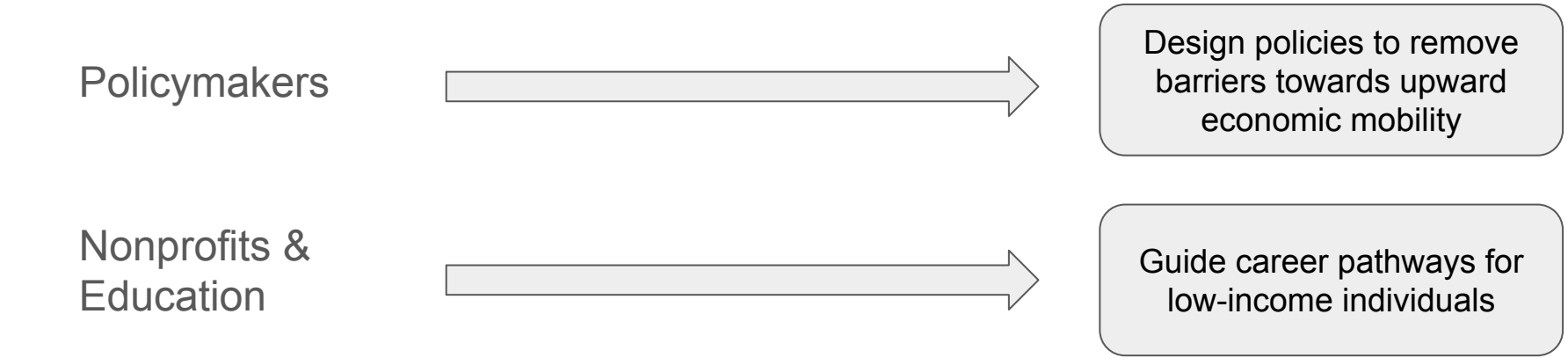
Executive Summary

- Successfully identified influential factors determining whether an individual earns above or below $50k
- Key takeaways
  - Education level, capital gains, and weeks worked during a year have a strong relationship with income
  - Demographic features like sex and age show disparities
  - Predictive modeling achieved 94.76% accuracy in classifying income groups

These insights can drive data-driven policies and economic mobility programs.

Why does this Matter?

Understanding characteristics associated with earning above or below $50K enable key stakeholders to take mitigative



Scope and Constraints

- Scope
  - Exploratory Data Analysis
  - Data Preparation
  - Modeling
  - Results
- Constraints
  - Time
  - Resources
  - Data
  - Human Capital

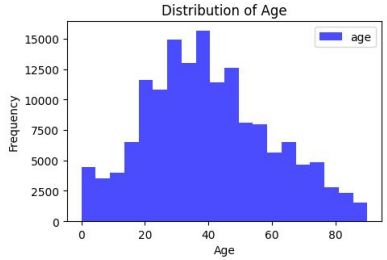
Data Overview

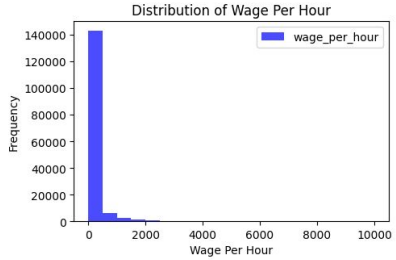
- US Census data from 1994-1995
- Roughly 300,000 records
- 40 features
  - 7 numerical
  - 33 categorical
- Bias and privacy considerations
  - Potential historical biases in census data
  - Not PII (Personally Identifiable Information) but must be careful
  - Demographic data requires special attention when used in modeling

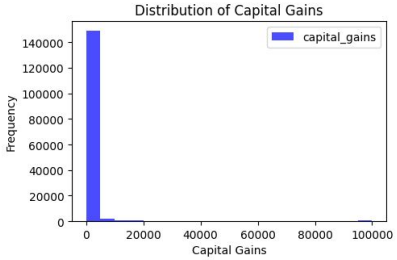
Exploratory Data Analysis

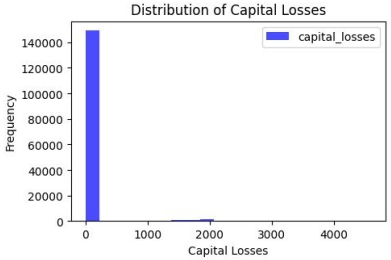
Distribution of Numerical Data

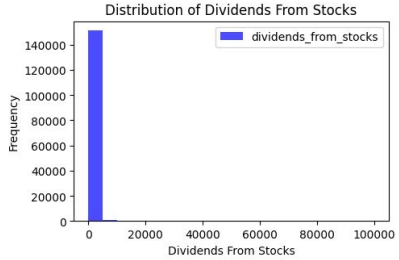
- Multiple features contain outliers and are skewed right
- Weeks worked in a year and number of person worked for employer have bimodal (two peaks) distributions
- Age is close to normally distributed but looks a bit skewed

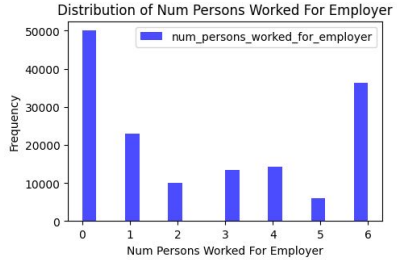


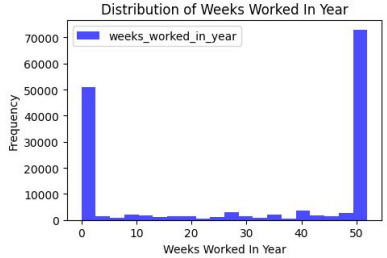






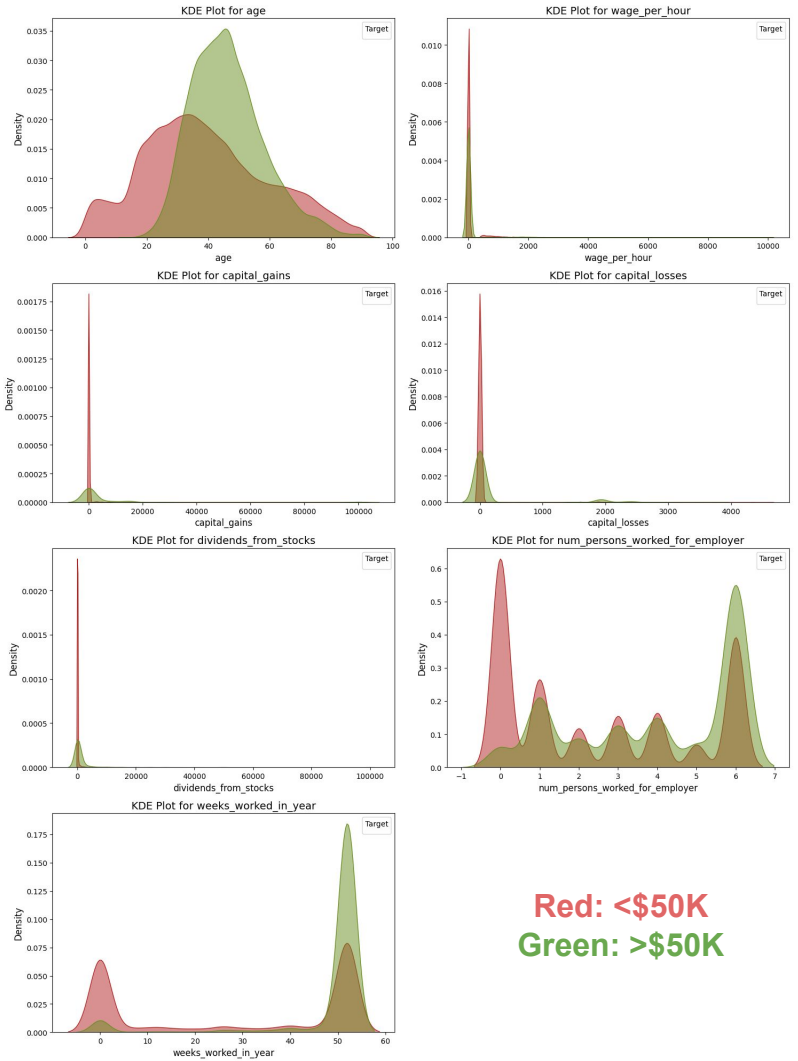






Distribution of Income Class

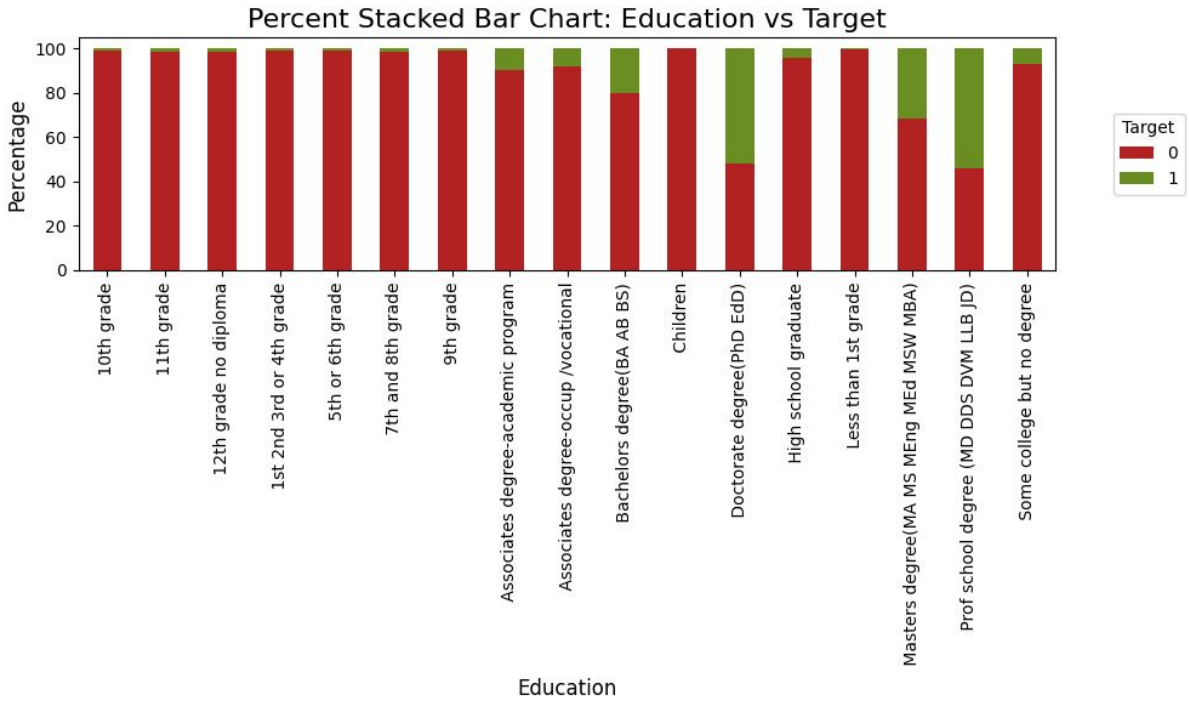
- Older individuals are more likely to earn >$50K than younger ones.
- When number of persons worked for employer is 6, it increases chances of >$50K. When 0, increases chances of <$50K.
- 50-52 weeks worked increases >$50K likelihood while 0 weeks strongly correlates with ≤$50K.



Red: <$50K
Green: >$50K

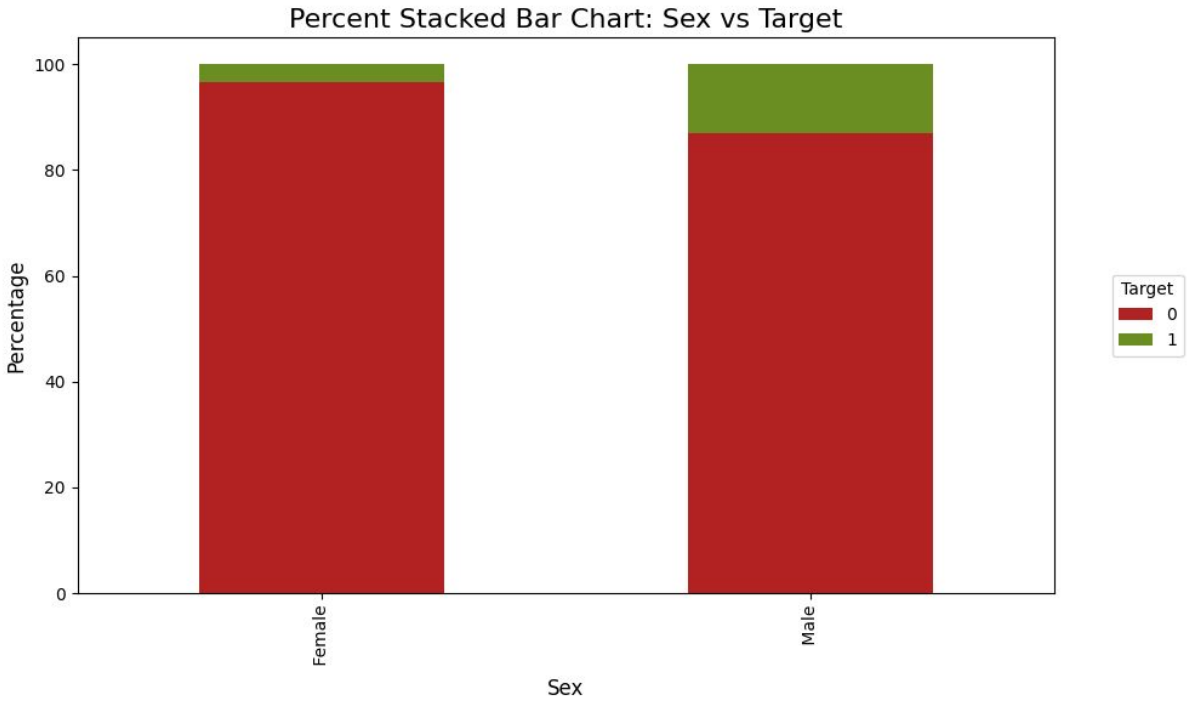
Education

- Advanced degrees show the highest percentage of income >$50K
  - PhD
  - Prof school degree
  - Masters degree



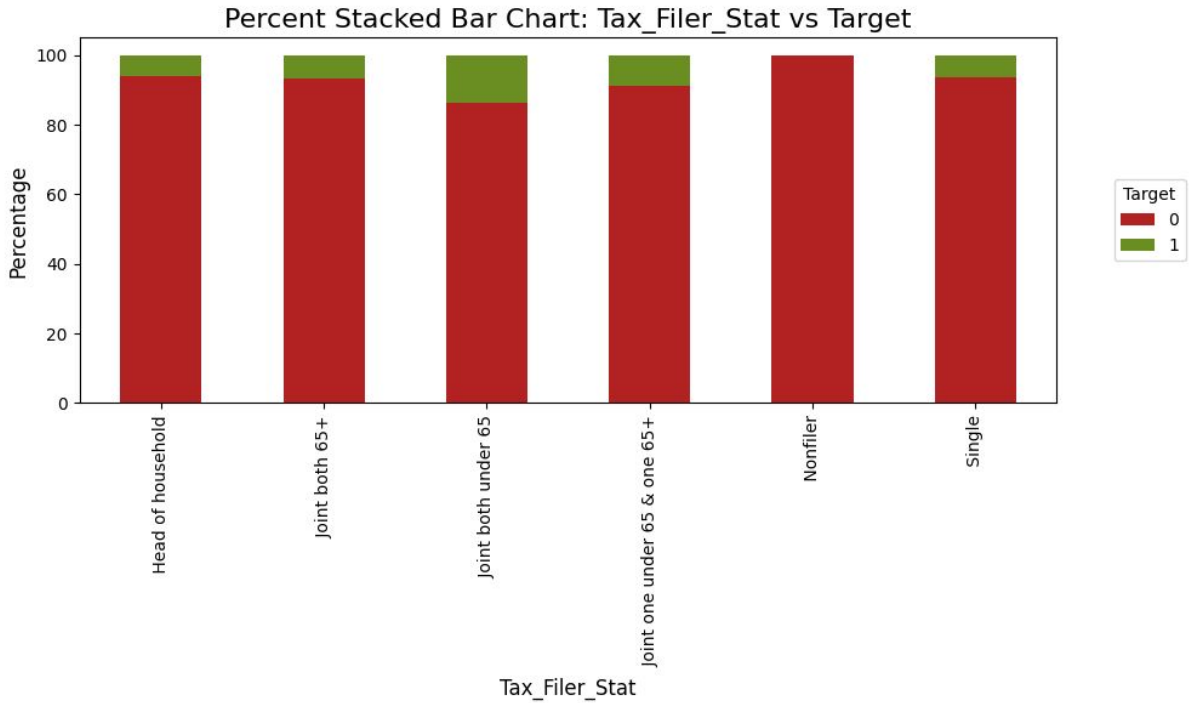
Sex

- Males have higher percentage of earning >$50K
- Highlights gender pay discrepancy



Tax Filer Status

- Joint both 65+ shows the highest percentage of income >$50K
- Nonfiler is nearly 100% <$50K



Data Preparation

Data Preparation Key Terms

- Ordinal feature: categorical data with a clear order
- Nominal feature: categorical data with no clear order
- Encoder: convert categorical data into numbers
  - Ordinal encoder: Preserves order of categories
  - One hot encoder: transform into binary representation
- Normalization: scales data to specific range, usually between 0 and 1

One Hot Encoding

Colour
Green
Red
Blue

Green	Red	Blue
0	1	1
1	1	1
1	0	1
0	0	0
0	1	0

Ordinal Encoding

Original Encoding	Ordinal Encoding
Poor	1
Good	2
Very Good	3
Excellent	4

Data Preprocessing

- Ordinal features (education feature)
  - Rank in order
  - Ordinal encoder to convert to integers
  - Normalize data between 0 and 1
- Nominal features
  - One hot encoder
- Numerical features
  - Normalize data between 0 and 1

Simple data pipeline that converts all features into numerical value with range 0 to 1

Modeling

Machine Learning Approach

- Problem type: Binary classification of income class
  - Below $50K (92%)
  - Above $50K (8%)
- Steps
  - Use lazypredict to test basic models
  - Select best model and perform hyperparameter tuning
  - Try improving model (feature selection, sampling, etc.)
  - Retrain on full training data
  - Return results

Best Model

- LightGBM Classifier was the best model
- Qualities of LightGBM
  - Based on decision trees
  - Boosted tree method
  - Handles missing values and categorical features
  - Optimized for speed
- Accuracy
  - Default hyperparameters: 94.28%
  - Hyperparameter tuning: 94.50%

Attempted Improvements

- Feature selection with recursive feature selection
- Oversampling
  - Random oversampling
  - SMOTE

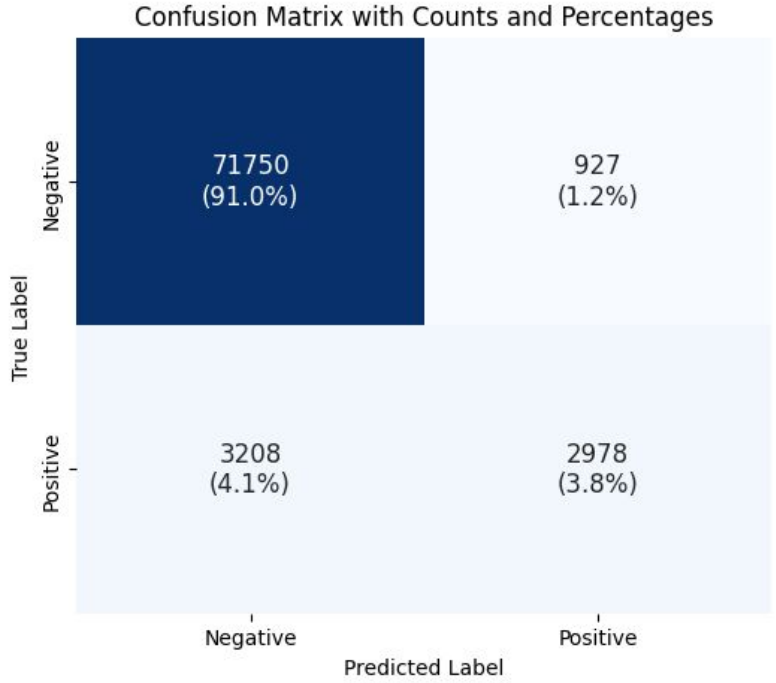
Model	Accuracy	F1 Score
Default hyperparameters	**94.28%**	**93.68%**
Feature selection	93.23%	92.20%
Random oversampling	85.99%	88.45%
SMOTE	93.25%	93.14%

Results

Model Evaluation and Metrics

After retraining on full training set:

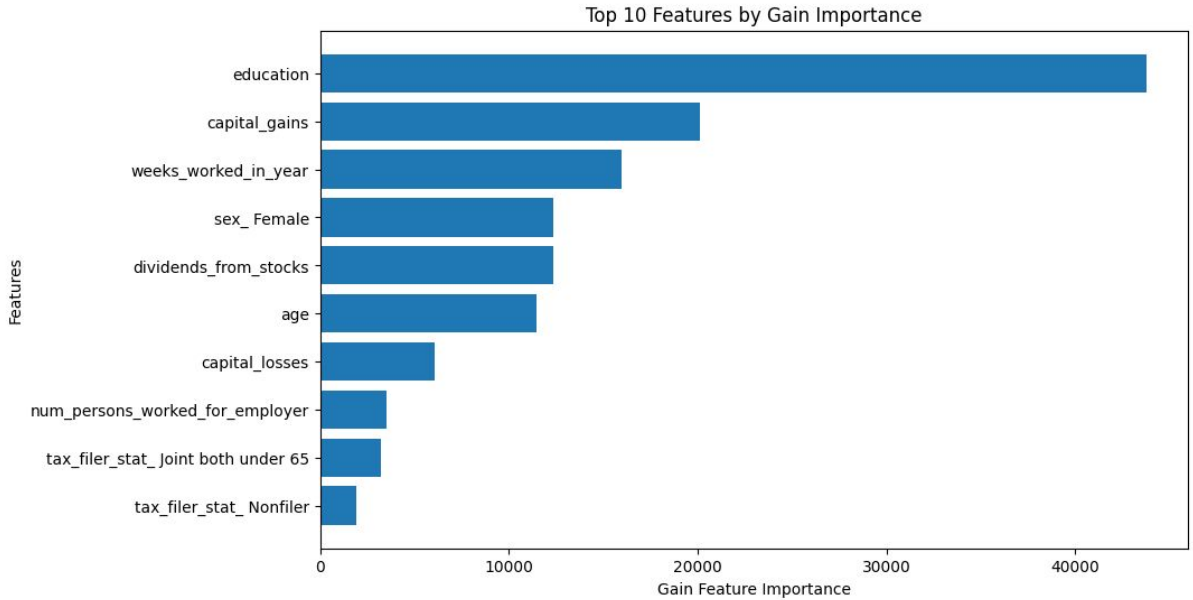
- Accuracy: 94.76%
- F1 Score (Weighted): 94.20%
- Precision: 94.19%
- Recall: 94.76%



Our model outperformed the 93.8% majority-class baseline.

Feature Importance

- Education is most important feature
- Capital gains ranks second
- Demographic features like sex and age rank in the top 10
- Tax filer status is in the top 10



Conclusion

Recommendations

- Education is the strongest predictor of income class
- Demographic features rank highly as predictors of income class highlighting disparities
- Practical applications
  - Policymakers: design programs and policies around upskilling, workforce development, and wage fairness
  - Nonprofits & educators: Develop career guidance programs and align educational curricula with economic opportunities

Drive Adoption

- Dataiku Data Science Studio (DSS)
  - End-to-end machine learning workflow in one platform
  - Streamlined data preparation and feature engineering
  - Collaboration and reproducibility
  - Scalability and deployment
- Expert support
  - Data science advisors
  - Learning modules for Dataiku DSS
  - Office hours
  - Masterclasses

Next Steps

- Deeper dive into each feature to build understanding
- Explore missing values and steps to impute them
- Impact of outliers and skewness
- Rebinning categories
- Trying more feature selection methods
- Creating more features and incorporating new data
- Error analysis of ML models
- Responsible ML practices
- Production-level ML pipeline with experiment tracking, orchestration, monitoring, CI/CD, etc.