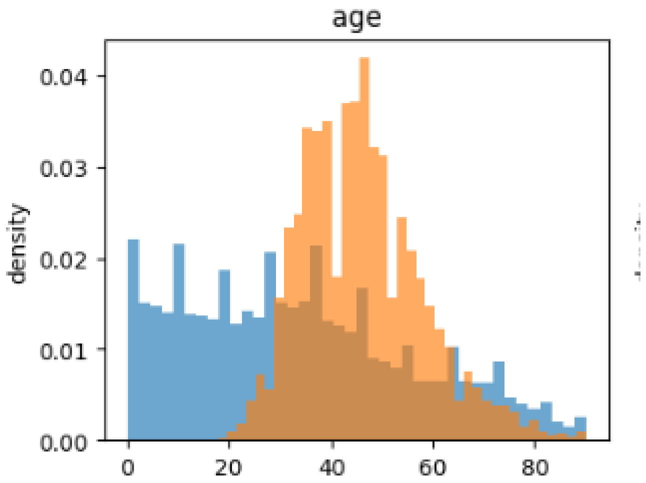
Dataiku Assessment

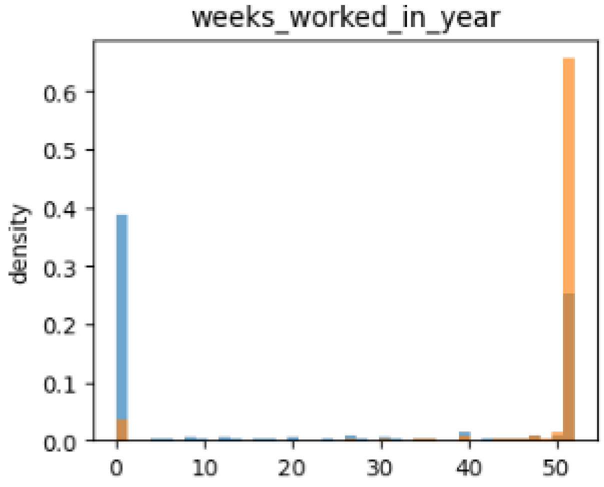
Abdul Moiz Amir

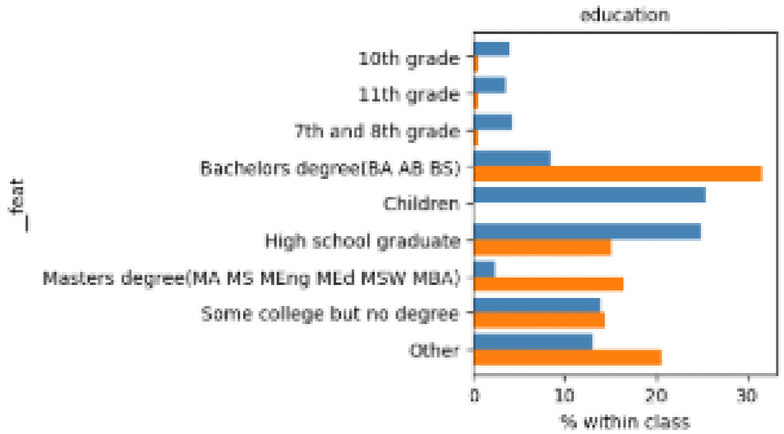
Census Income Analysis

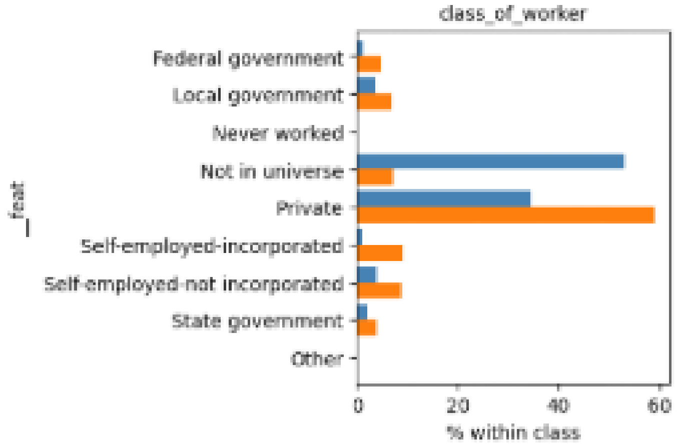
- Goal: Predict whether an individual earns more than $50K per year using U.S. Census data.
- Use Case: Government policy, labor economics, tax planning, or targeted outreach.
- Source: U.S. Census Bureau
    - 94% < $50,000
    - 6% > $50,000

Some Drivers of Income > $50,000









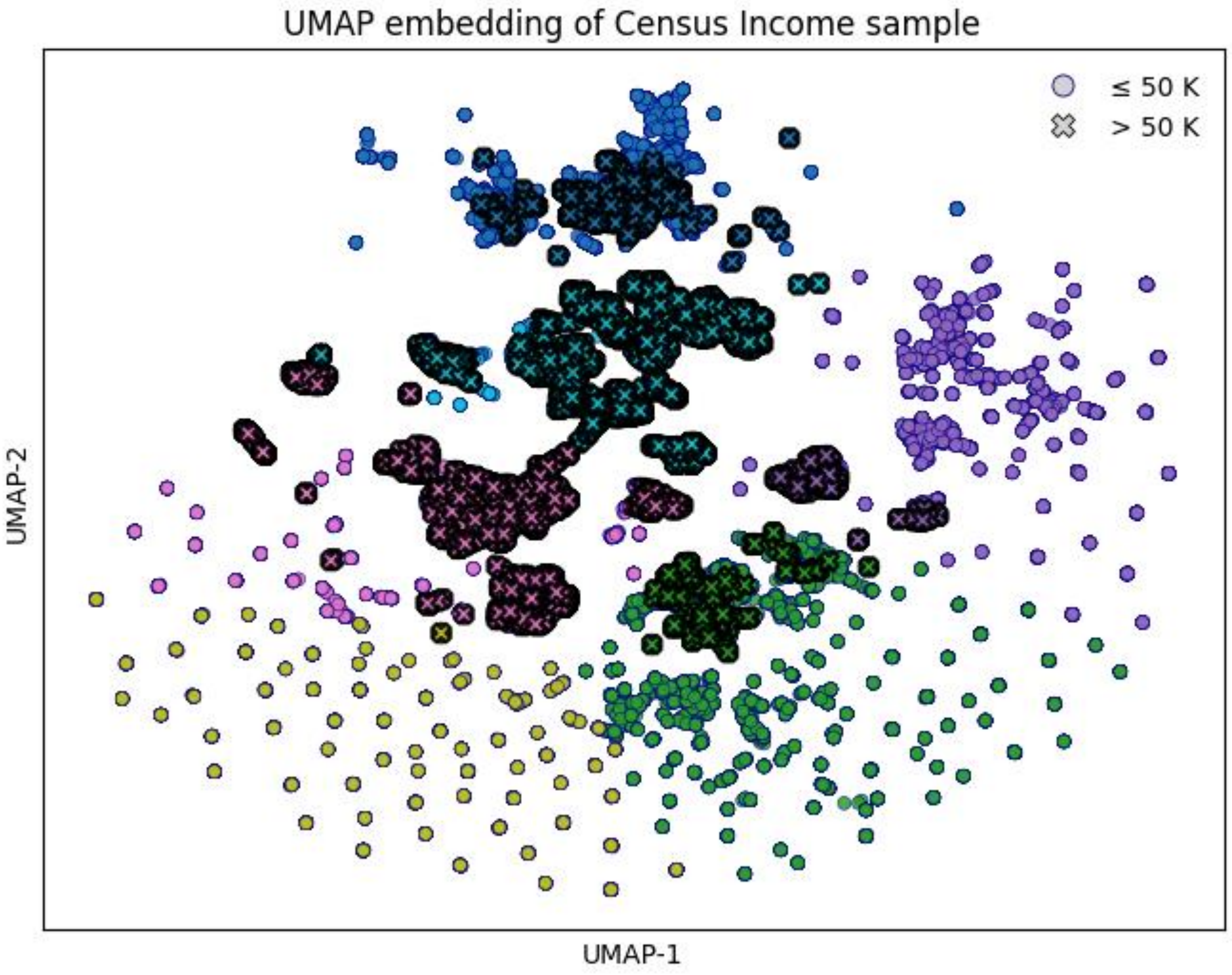
Some Drivers of Income > $50,000

- P(Income > 50k) = 0.06

- P(Income > 50k | Invests) = 0.32 (chances increases by **500%**)

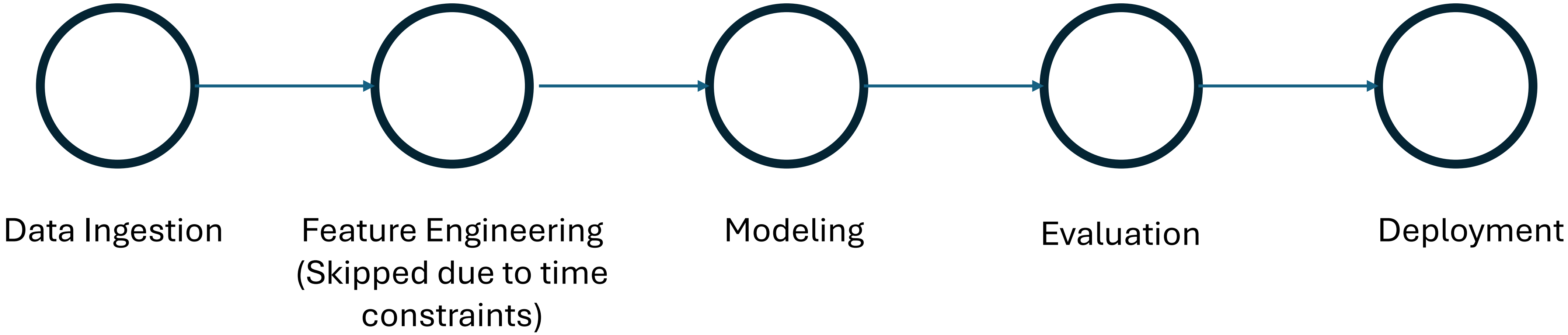
- P(Income > 50k | Masters) = 0.31 (chances increases by **500%**)

Demographic Cluster



- **Income > $50,000:**
  - white males with veteran benefits working in Private sector as a professional specialty
- **Income <= $50,000:**
  - Age groups not in the work force
  - Uneducated females

AI Classification Engine



Data Ingestion

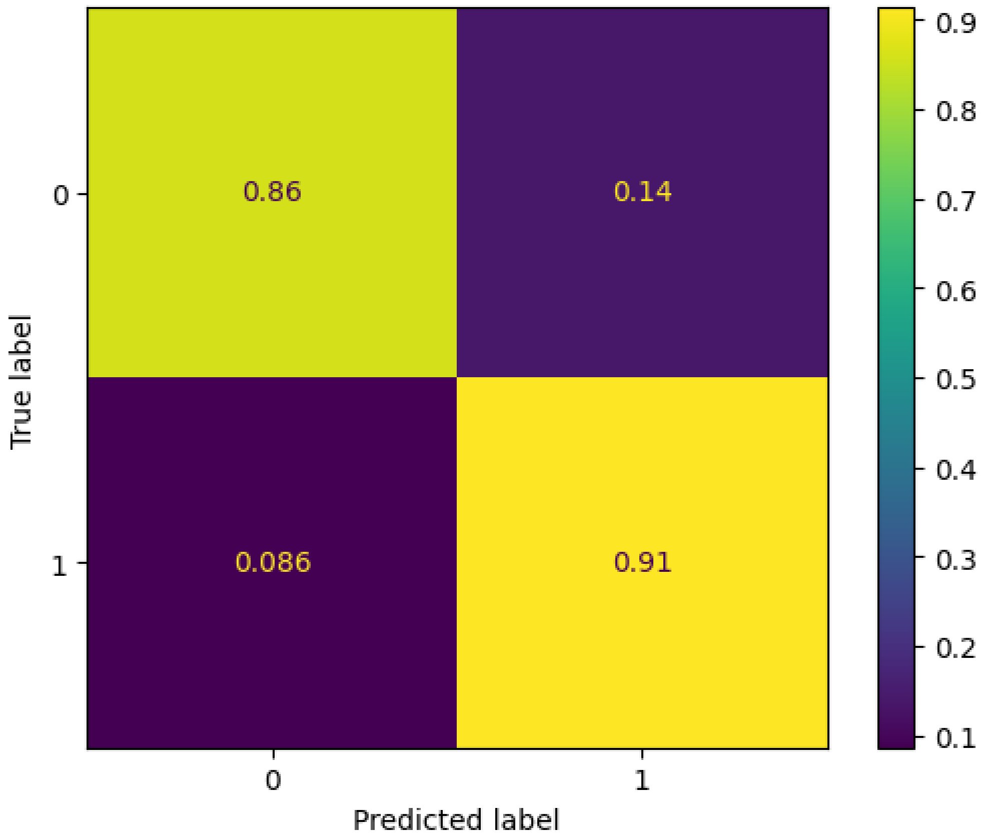
Feature Engineering (Skipped due to time constraints)

Modeling

Evaluation

Deployment

Modelling Efforts - Validation



Data Split

Learn.csv: 75% train, 25% validation

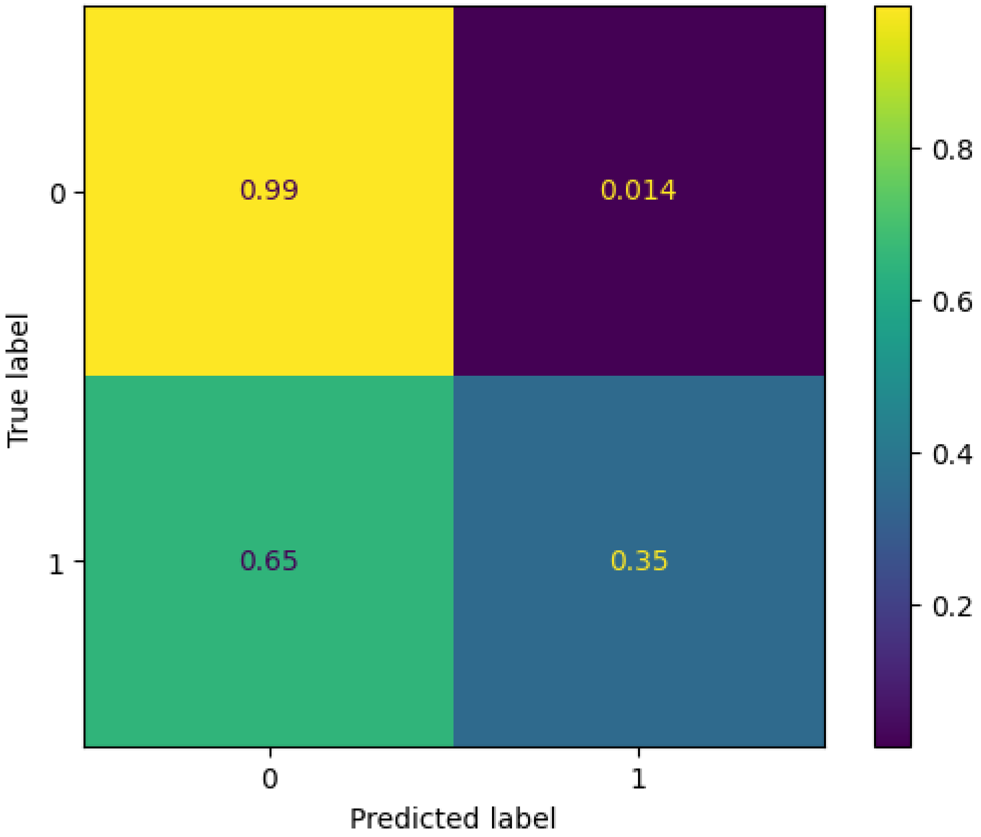
Models Considered

- Logistic Regression (as baseline): ROC-AUC 0.94
- LightGBM: ROC-AUC 0.955
- NaiveBayes (not implemented due to time constraints)

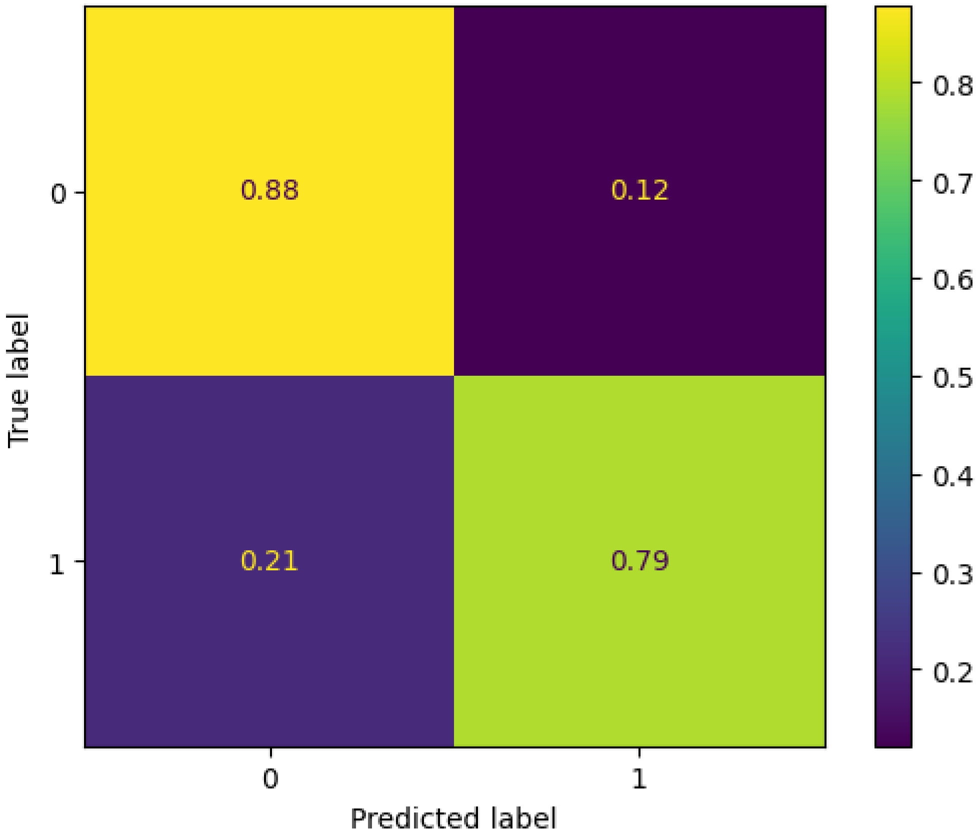
Optimal Threshold from validation: 0.059

F1: 0.47

Modelling Efforts – Challenges



Using Threshold from Validation Set
F1: 0.45



Using Threshold from Tuned on Test Set
F1: 0.43

Thank You!