Riyadh | 16.07.2025

CENSUS INCOME PREDICTION

From Data to Deployment

DATAIKU TECHNICAL ASSESMENT

Mahmoud Emara
Mahmoud.emarah@cis.asu.edu.eg

AGENDA

1 PROJECT FOUNDATION

- Business Problem
- Project Overview & Methodology
- Data Exploration & Key Findings
- Feature Analysis & Relationships

2 DEVELOPMENT & MODELING

- Feature Engineering
- Model Development Strategy
- Model Performance Comparison

3 EVALUATION & DEPLOYMENT

- Bias Analysis & Fairness Assessment
- Production Deployment
- Interactive Web Application Demo
- Monitoring Strategy

4 INSIGHTS & FUTURE WORK

- Key Challenges
- Lessons Learned & Best Practices
- Next Steps & Future Enhancements
- Conclusion & Q&A Session

Riyadh | 16.07.2025

PROJECT FOUNDATION

BUSINESS PROBLEM & OPPORTUNITY

THE CHALLENGE:

- Income classification is critical for financial institutions, government agencies, and marketing companies
- Traditional methods rely on self-reported data which is often inaccurate or incomplete
- Need for automated, accurate prediction of income levels based on demographic and employment data

KEY BUSINESS QUESTION

- Can we build a reliable model to predict whether an individual earns more than $50K annually based on census data?

BUSINESS IMPACT:

- Targeted marketing: very important for customer segmentation solutions
- Policy planning: Better resource allocation for government programs
- Research: Enhanced demographic analysis for economic studies

Riyadh | 16.07.2025

PROJECT FOUNDATION

PROJECT OVERVIEW & OBJECTIVES

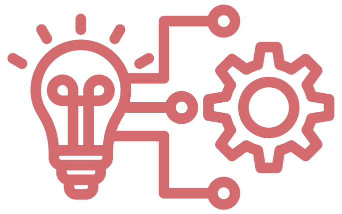
5-PHASE METHODOLOGY:

Our approach follows a structured 5-phase methodology to ensure thorough analysis, robust model development, and production-ready deployment.

1- EXPLORATORY DATA ANALYSIS

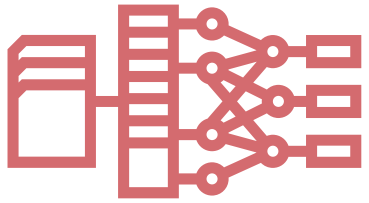
- System Setup
- Data Ingesion
- Exploratory Data Analysis

2- FEATURE ENGINEERING



- Data Cleaning
- Handel Missing Values
- Create New Features

3- MODEL DEVELOPMENT

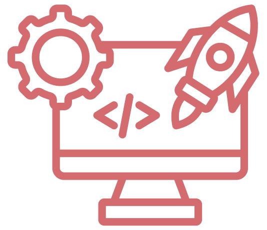


- Selecting Models
- Training Models
- Tuning Hyperparameters

4- MODEL EVALUATION

- Assessing Performance
- Validating results
- Check Bias

5- PRODUCTION READINESS



- Model Deployment
- GitHub readiness
- Web Application

PROJECT FOUNDATION

PROJECT OVERVIEW & OBJECTIVES

SUCCESS METRICS

PRIMARY

ROC-AUC > 90%, Recall > 85% for high-income class

BUSINESS

Enhancment on marketing applications

SECONDARY

Model interpretability score, bias metrics < 5%

Riyadh | 16.07.2025

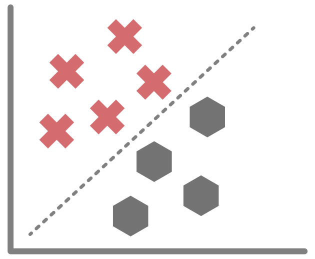
PROJECT FOUNDATION

DATA EXPLORATION & INSIGHTS

US CENSUS INCOME

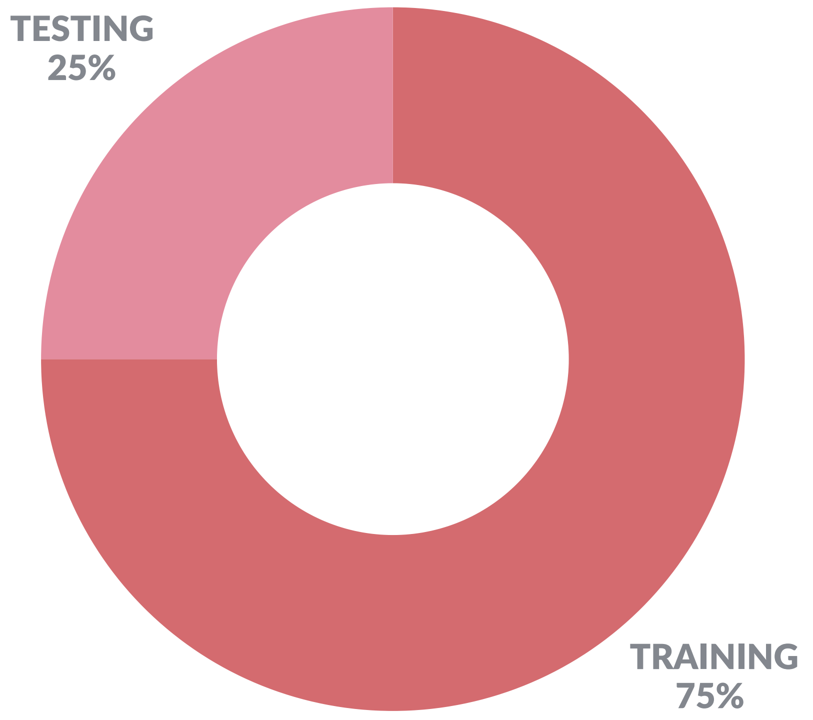
400 K

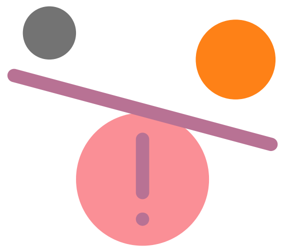
199,523	99,762	41
Training Samples	Testing Samples	Features



Target: Binary income classification (<50K$ or >=50K$)

DATA SPLIT



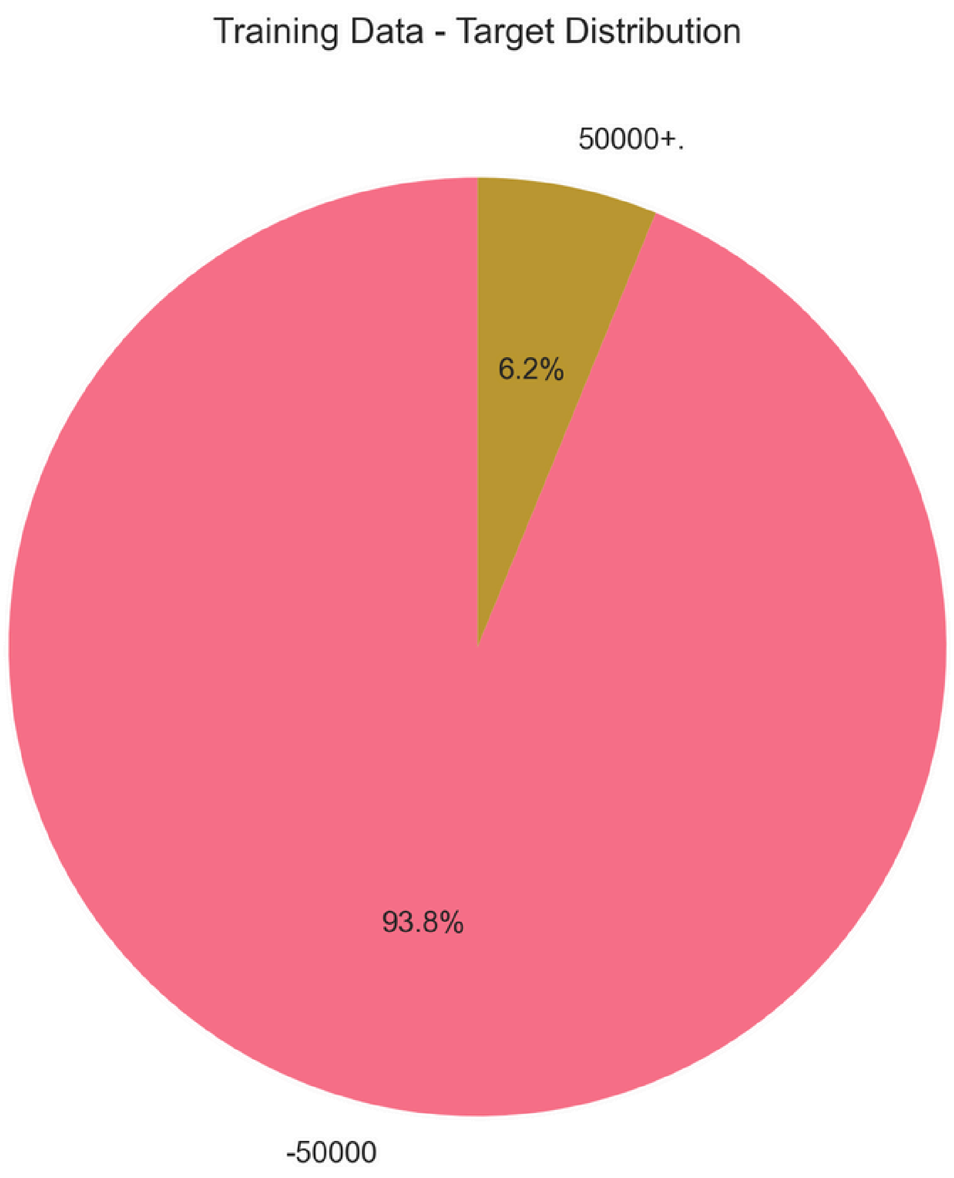


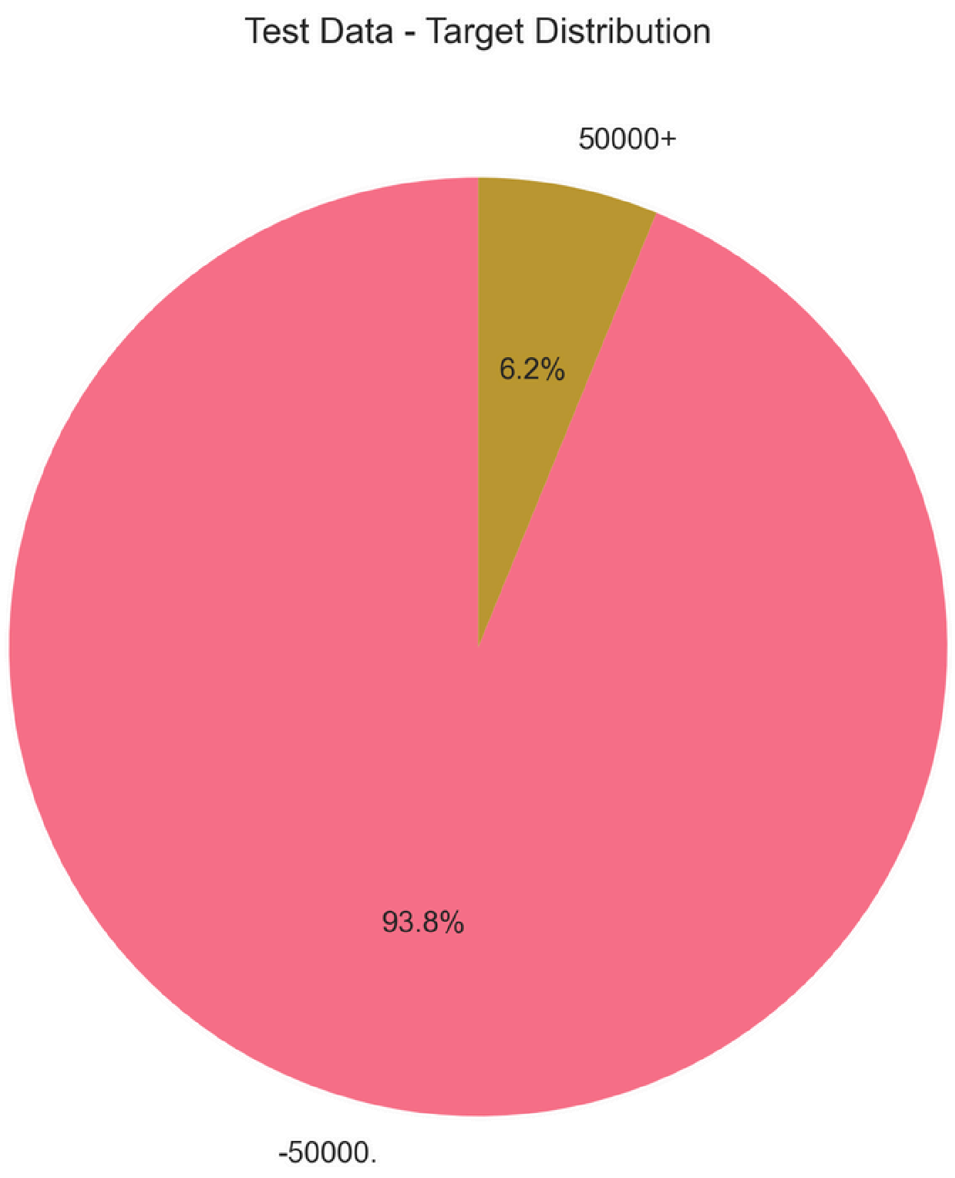
Class imbalance: 93.8% low-income vs. 6.2% high-income

PROJECT FOUNDATION

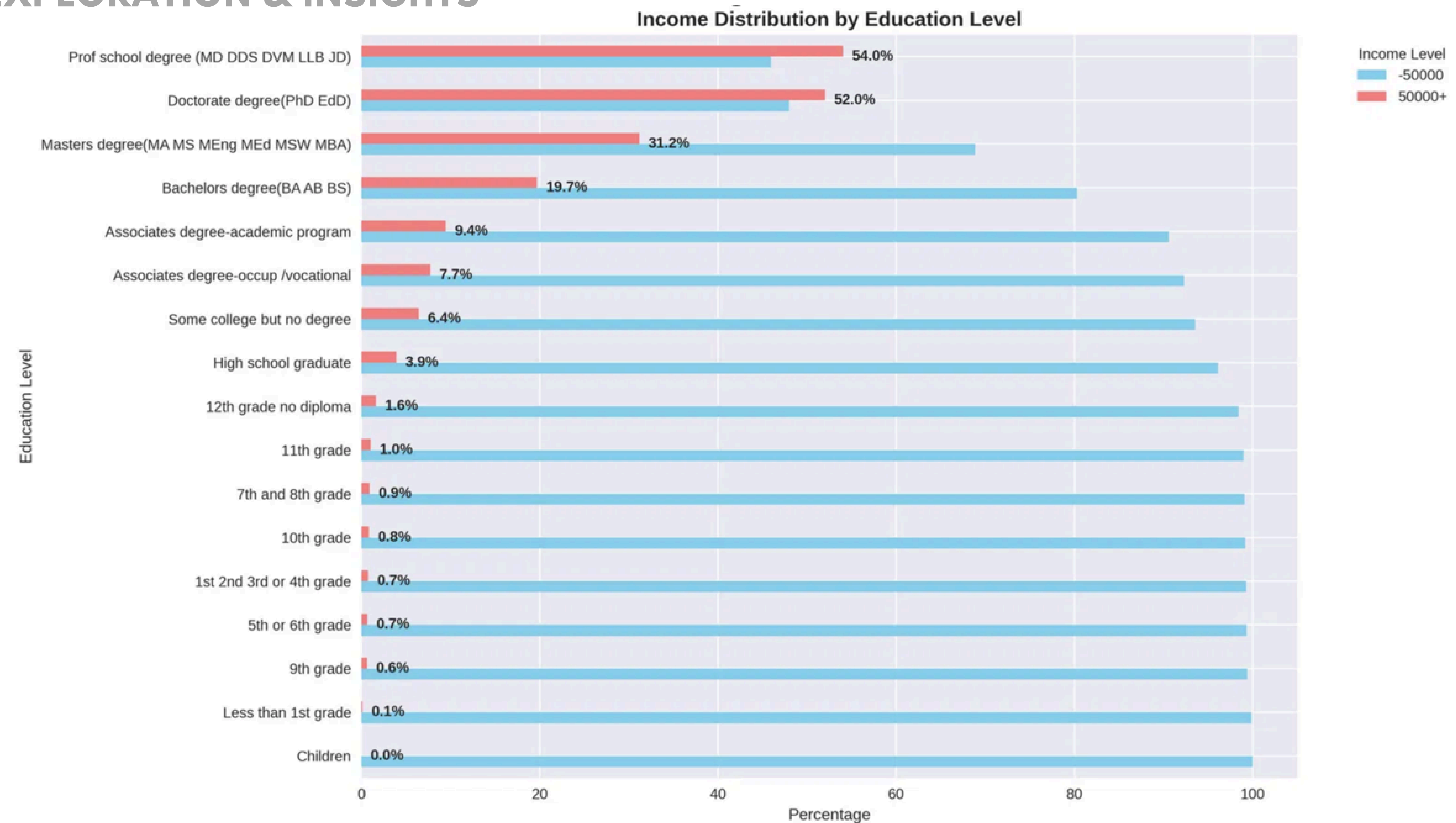
DATA EXPLORATION & INSIGHTS

TARGET DISTRIBUTION



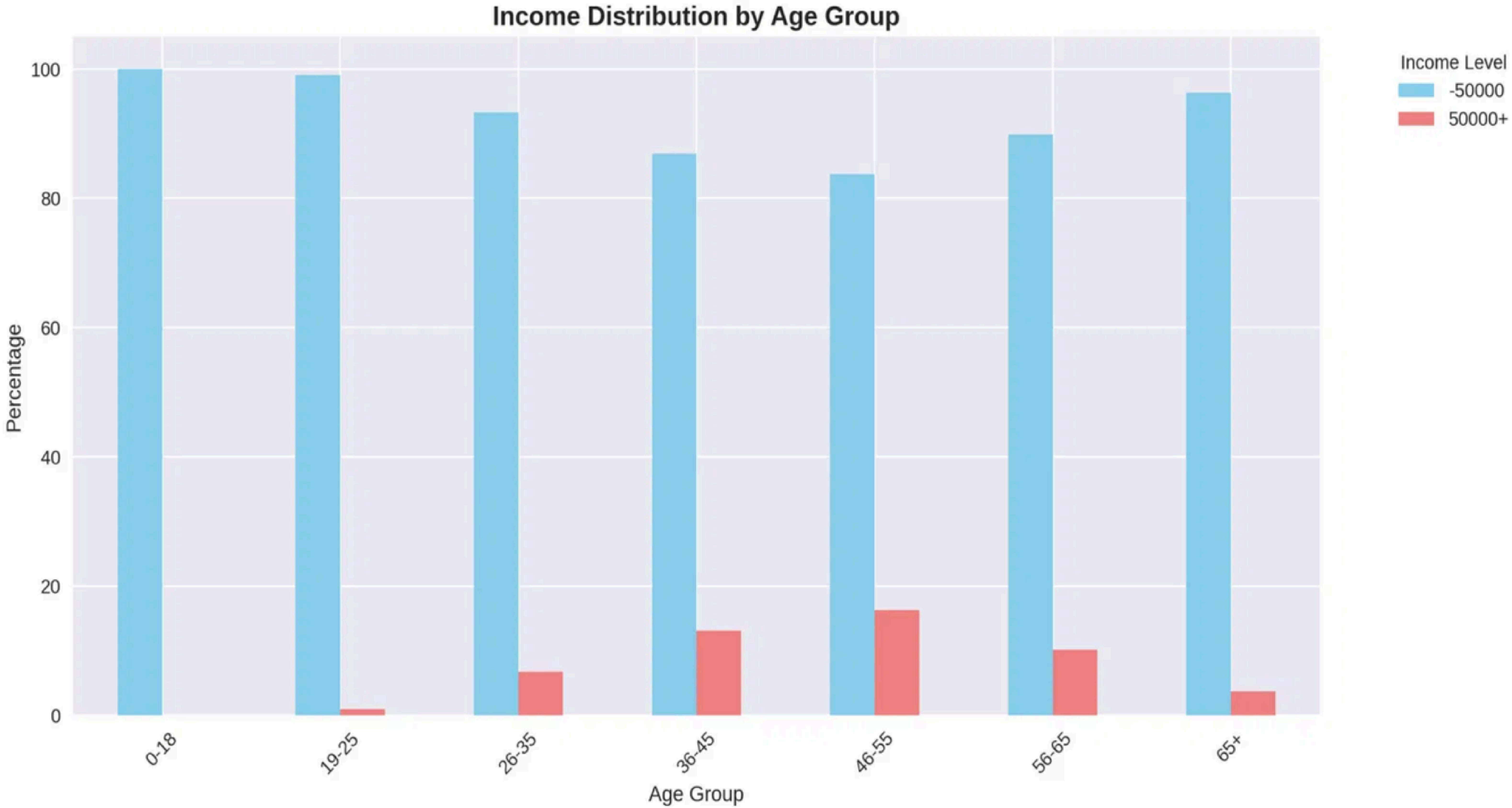


PROJECT FOUNDATION
DATA EXPLORATION & INSIGHTS



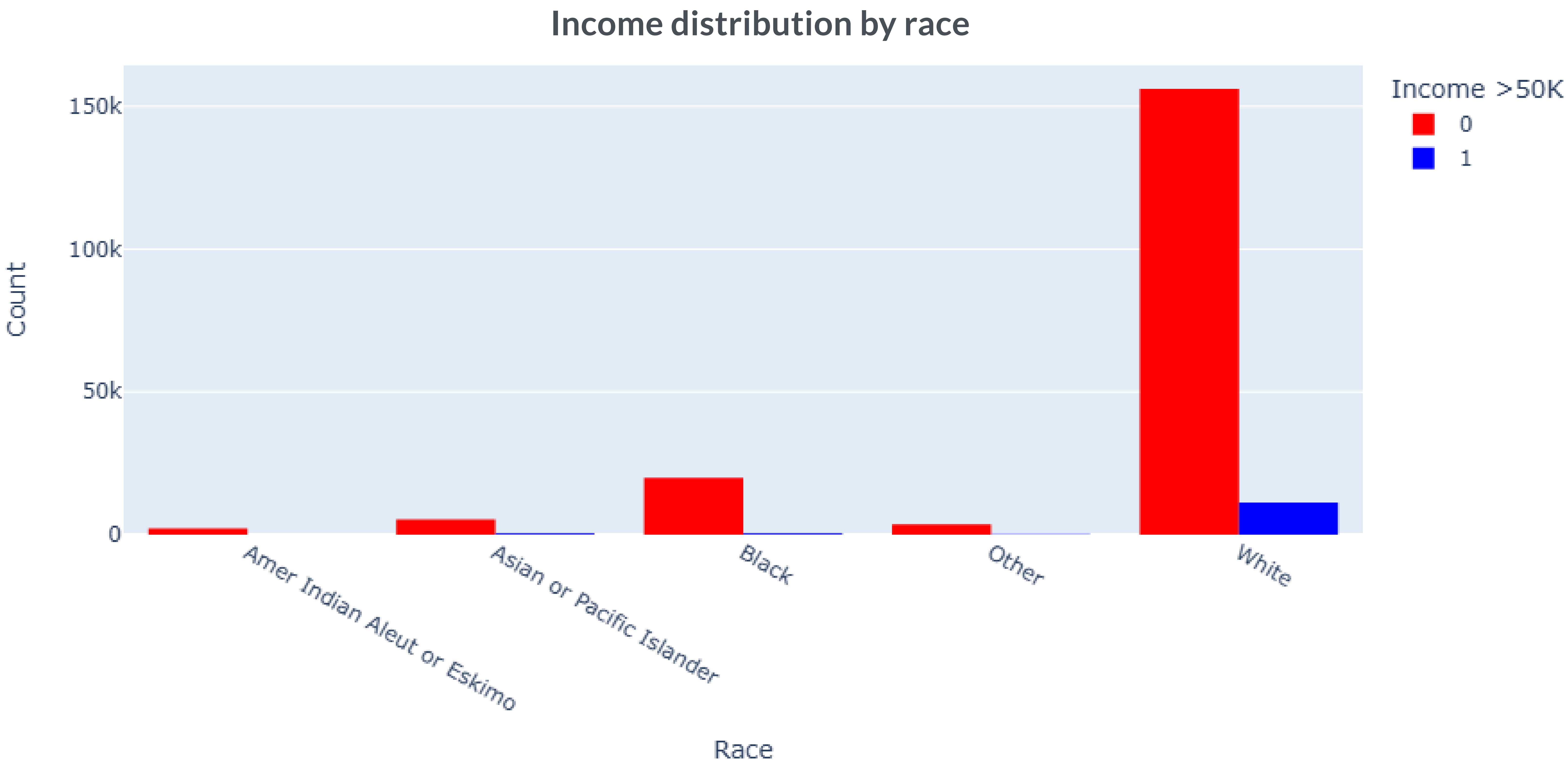
PROJECT FOUNDATION

DATA EXPLORATION & INSIGHTS

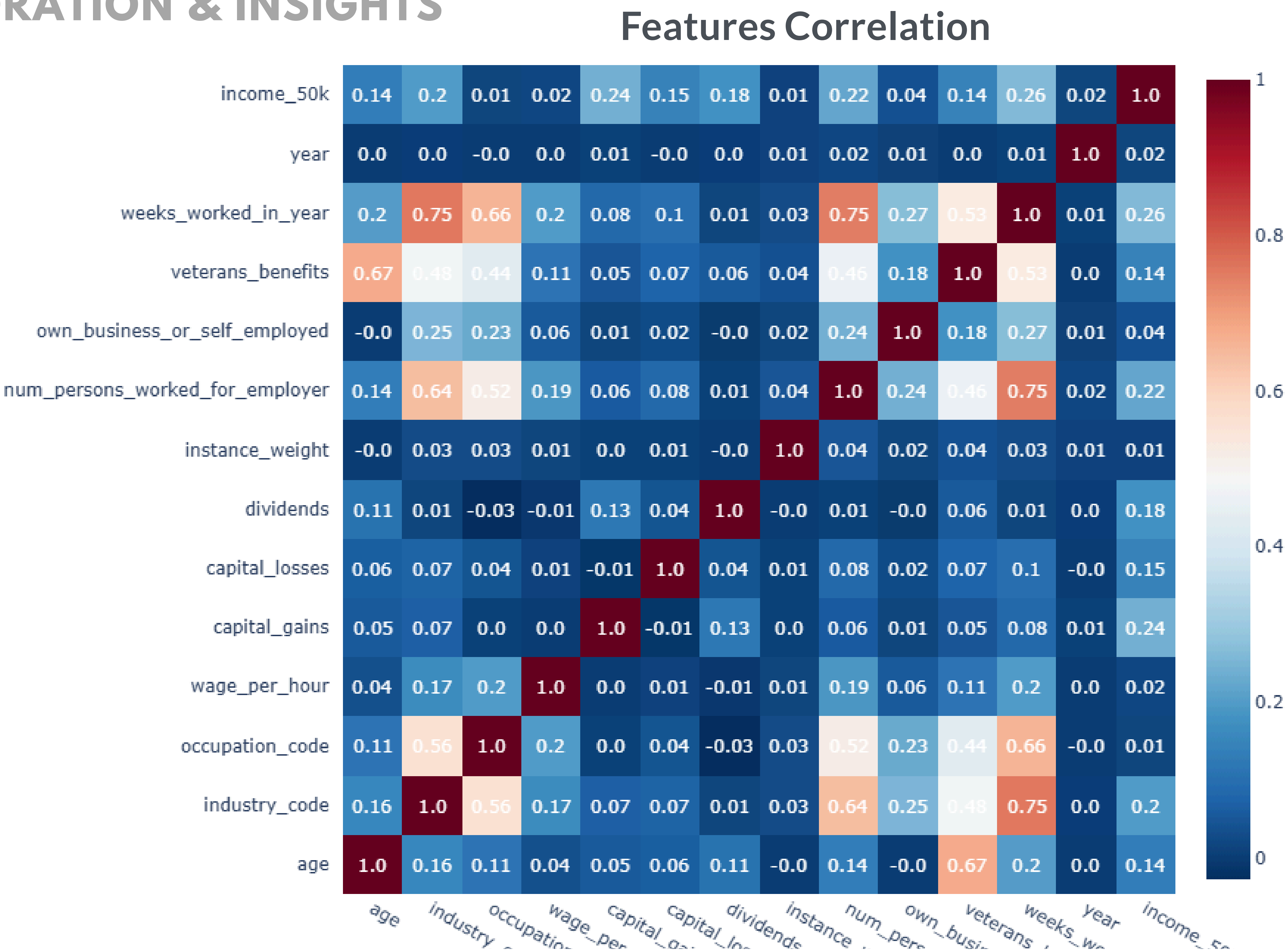


PROJECT FOUNDATION

DATA EXPLORATION & INSIGHTS

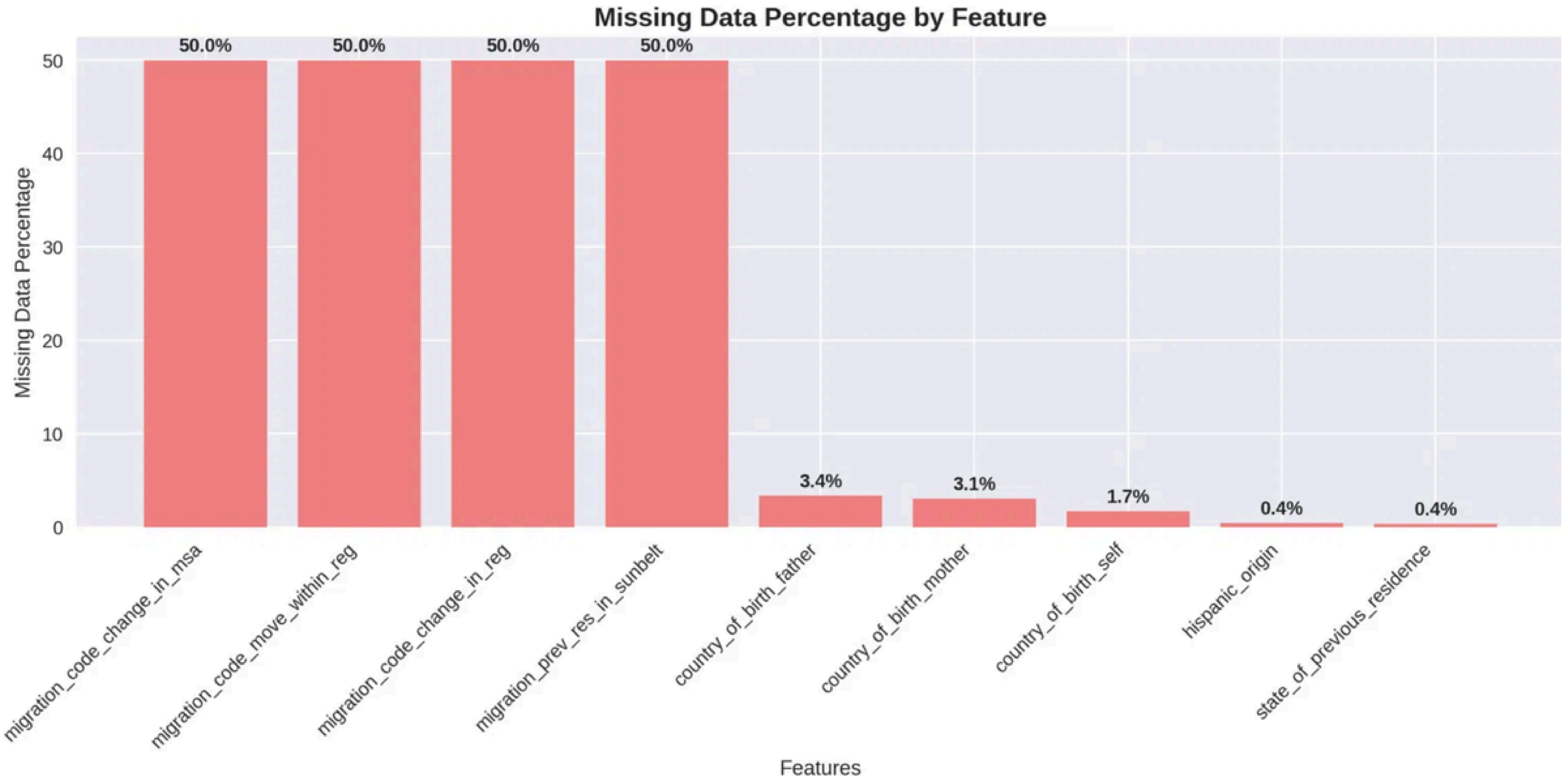


PROJECT FOUNDATION
DATA EXPLORATION & INSIGHTS



PROJECT FOUNDATION

DATA EXPLORATION & INSIGHTS



Riyadh | 16.07.2025

PROJECT FOUNDATION

DATA EXPLORATION & INSIGHTS

KEY INSIGHTS

Severe Class Imbalance

93.8% of individuals earn ≤$50K vs. only 6.2% earning >$50K, requiring special handling during modeling.

Missing Data Patterns

Migration-related features show ~50% missing values, while other features are relatively complete.

Education is Critical

Professional degree holders have a 54.7% high-income rate, compared to only 3.2% for those with less than high school education.

Age Matters Significantly

High earners average 46.3 years old vs. 33.7 years for low earners, showing a clear age-income relationship.

Work Class Impact

Self-employed incorporated workers show 34.7% high-income rate, while private sector workers show only 6.2%.

Key Correlations

Strong positive correlations between education, occupation, and income level. Age and hours-per-week also show moderate positive correlation with income.

No Data Leakage

Careful analysis confirmed no data leakage between features and target variable. All correlations represent genuine predictive relationships.

Riyadh | 16.07.2025

PROJECT FOUNDATION

DATA CHALLENGES & SOLUTIONS

Several significant data quality challenges required strategic solutions to ensure model reliability.

High Cardinality in Categorical Features

Several categorical features had high cardinality (many unique values), making one-hot encoding impractical and risking overfitting.

Solution: Rare Category Grouping + Target Encoding

Grouped rare categories (frequency <1%) into an "Other" category and applied target encoding for high-cardinality features to create meaningful numerical representations.

Duplicate Records

Identified 53,878 duplicate records (27% of training data), which could lead to data leakage between training and validation sets.

Solution: Duplicate Detection & Removal

Implemented robust duplicate detection and kept only the first occurrence of each record, reducing training data from 199,523 to 152,807 samples.

Overfitting Risk with 188 Features

Feature engineering expanded the feature space from 40 to 188 dimensions, increasing the risk of overfitting.

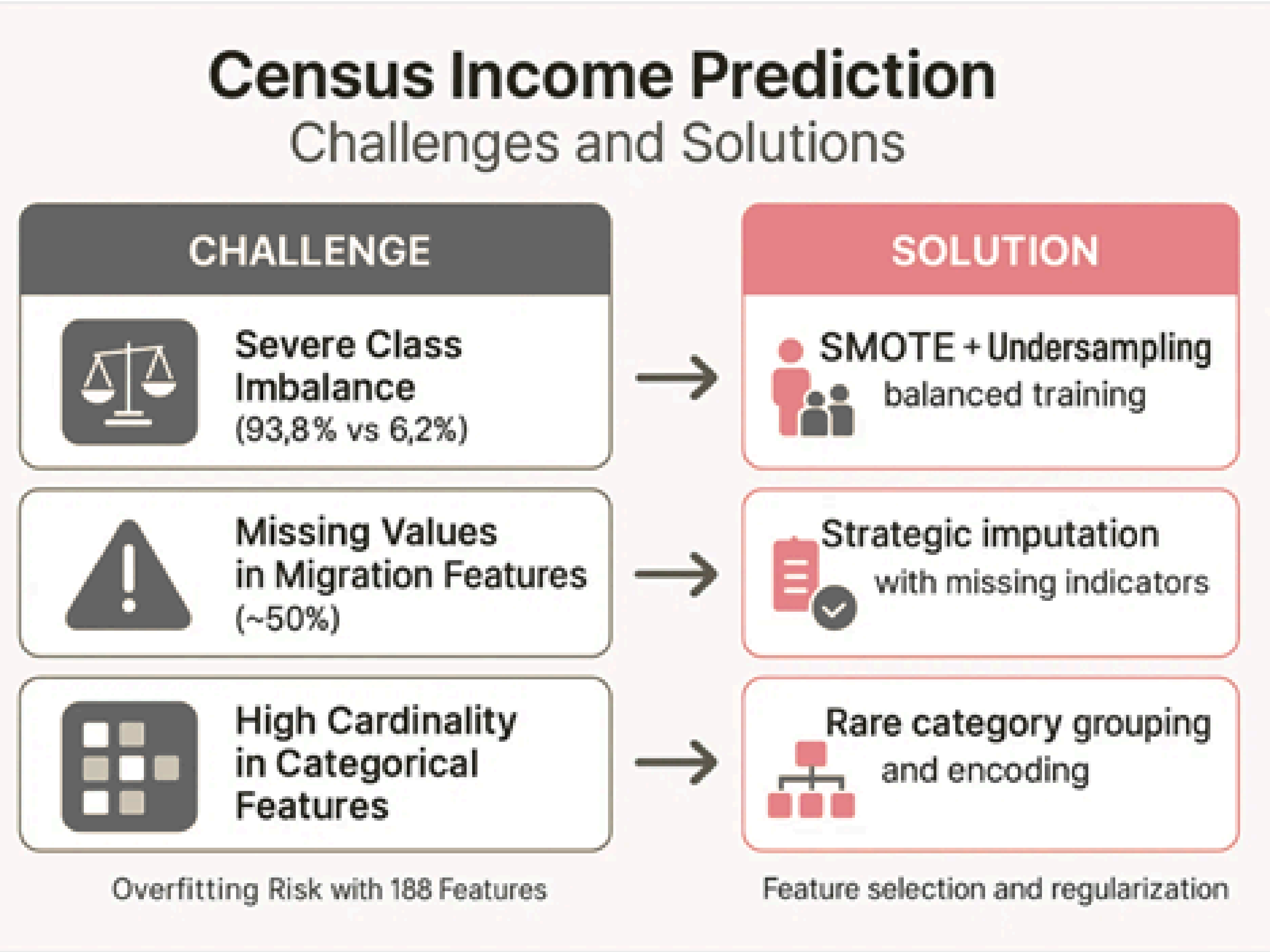
Solution: Feature Selection + Regularization

Applied feature importance analysis to select the top 50 most predictive features and implemented regularization techniques (L1/L2) to prevent overfitting.

PROJECT FOUNDATION

KEY DATA QUALITY ISSUES

Several significant data quality challenges required strategic solutions to ensure model reliability.



Severe Class Imbalance (93.8% vs 6.2%)

The extreme imbalance between income classes risked creating models biased toward the majority class.

Solution: SMOTE + Undersampling

Combined Synthetic Minority Over-sampling Technique (SMOTE) with undersampling to create a perfectly balanced training dataset (1:1 ratio) while preserving data characteristics.

Missing Values in Migration Features (~50%)

Migration-related features showed significant missing data, potentially limiting their usefulness.

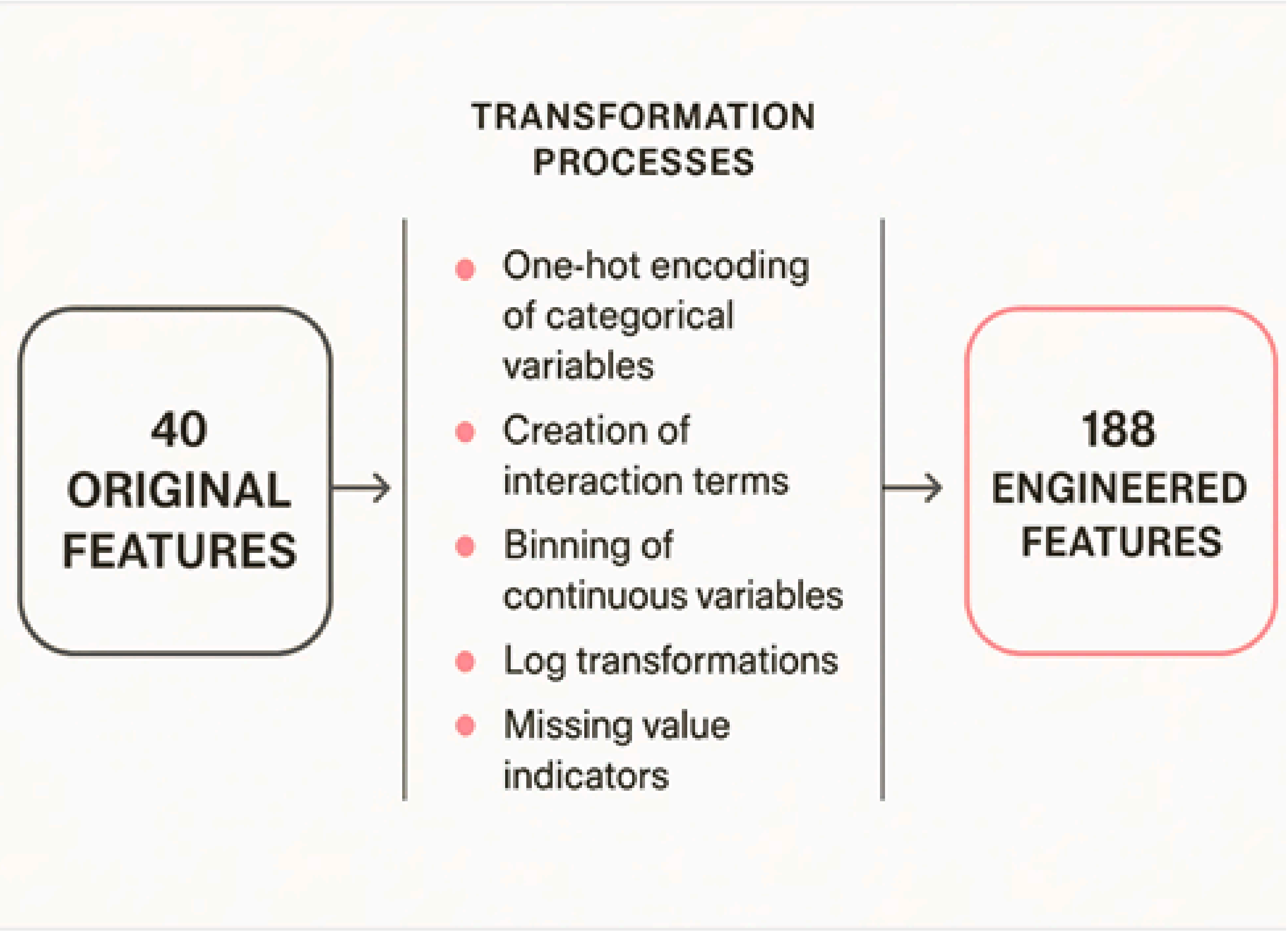
Solution: Strategic Imputation + Missing Indicators

Implemented KNN imputation for moderate missing values and created binary missing indicators to capture missingness patterns as potential signals.

PROJECT FOUNDATION

FEATURE TRANSFORMATION PROCESS

We expanded the feature space from 40 original features to 188 engineered features through a strong transformation process.



Age-Based Features

age_group (binned into 5-year intervals)

is_senior (age >= 65)

is_young_adult (age < 25)

age_squared (to capture non-linear effects)

Work-Based Features

is_self_employed (binary indicator)

is_government_worker (binary indicator)

work_intensity (hours_per_week / 40)

is_full_year_worker (weeks_worked >= 50)

PROJECT FOUNDATION

FEATURE TRANSFORMATION PROCESS

We expanded the feature space from 40 original features to 188 engineered features through a strong transformation process.

Education-Based Features

- education_level (ordinal encoding)
- has_college_degree (binary indicator)
- has_advanced_degree (Masters/PhD/Prof)
- education_years (estimated years of education)

Financial Features

- has_capital_gains (binary indicator)
- has_capital_losses (binary indicator)
- log_capital_gains (log transformation)
- has_investment_income (binary indicator)

Interaction Features

- age_education_interaction
- married_with_children
- education_occupation_interaction
- age_work_class_interaction

Riyadh | 16.07.2025

PROJECT FOUNDATION

PREPROCESSING PIPELINE ARCHITECTURE

We built a robust, production-ready preprocessing pipeline using scikit-learn's Pipeline , ensuring consistent transformations between training and inference.

Duplicate Handling

Detect and remove duplicate records to prevent data leakage between train/validation sets

Data Cleaning

Handle missing values, outliers, and inconsistent data formats

Feature Engineering

Create new features, interactions, and transformations

Encoding & Scaling

Convert categorical features and scale numerical features

Class Balancing

Apply SMOTE and undersampling to create balanced training data

DEVELOPMENT & MODELING

MODEL DEVELOPMENT STRATEGY

MODELING APPROACH

We developed and evaluated multiple competing models with different algorithmic approaches to identify the best performer.

Logistic Regression

Linear model with L2 regularization serving as our baseline. Surprisingly strong performance with high interpretability.

ROC-AUC	Accuracy	Recall
94.11%	86.83%	88.69%

XGBoost

Gradient boosting implementation with advanced regularization. Top performer with excellent generalization.

ROC-AUC	Accuracy	Recall
99.30%	96.32%	95.43%

Random Forest

Ensemble of 100 decision trees with bootstrap sampling. Excellent performance with moderate interpretability.

ROC-AUC	Accuracy	Recall
99.01%	95.16%	95.67%

Neural Network

3-layer MLP with dropout regularization. Good performance but more complex to interpret and deploy.

ROC-AUC	Accuracy	Recall
96.76%	91.06%	92.88%

DEVELOPMENT & MODELING

MODEL DEVELOPMENT STRATEGY

DEVELOPMENT BEST PRACTICES

Cross-Validation Strategy

Implemented 5-fold stratified cross-validation to ensure robust performance evaluation across different data subsets. Standard deviations across folds were consistently low (±0.1-0.4%), indicating stable model performance.

Hyperparameter Tuning

Performed systematic grid search and random search for each model type, optimizing for ROC-AUC. Key parameters tuned included regularization strength, tree depth, learning rate, and ensemble size.

Handling Class Imbalance

Trained models on perfectly balanced data (1:1 ratio) using SMOTE + undersampling, but evaluated on original imbalanced distribution to ensure real-world performance. This approach prevented majority class bias while maintaining realistic evaluation.

Model Interpretability

Prioritized interpretable models and techniques, including feature importance analysis, partial dependence plots, and SHAP values to explain model decisions. This ensures transparency and builds trust with stakeholders.

EVALUATION & DEPLOYMENT

COMPREHENSIVE METRICS COMPARISON

All models were evaluated on multiple metrics to ensure a balanced assessment of performance.

Model	ROC-AUC	Accuracy	Precision	Recall	F1-Score
XGBoost	**99.30%**	**96.32%**	**97.16%**	95.43%	**96.29%**
LightGBM	99.30%	96.25%	97.07%	95.37%	96.21%
Random Forest	99.01%	95.16%	94.70%	**95.67%**	95.19%
Neural Network	96.76%	91.06%	89.63%	92.88%	91.22%
Logistic Regression	94.11%	86.83%	85.52%	88.69%	87.07%

EVALUATION & DEPLOYMENT

BIAS ANALYSIS & FAIRNESS

We conducted a comprehensive fairness analysis across demographic groups to ensure the model performs equitably.

Demographic Group	Equal Opportunity	Demographic Parity	Accuracy Parity
Female	0.96	0.87	0.97
Male	0.95	1.00	0.96
College Degree	0.97	1.02	0.98
No College Degree	0.94	0.85	0.95
Age ≥ 40	0.96	1.05	0.97
Age < 40	0.93	0.82	0.94
Native-Born	0.95	0.98	0.96
Immigrant	0.94	0.89	0.95

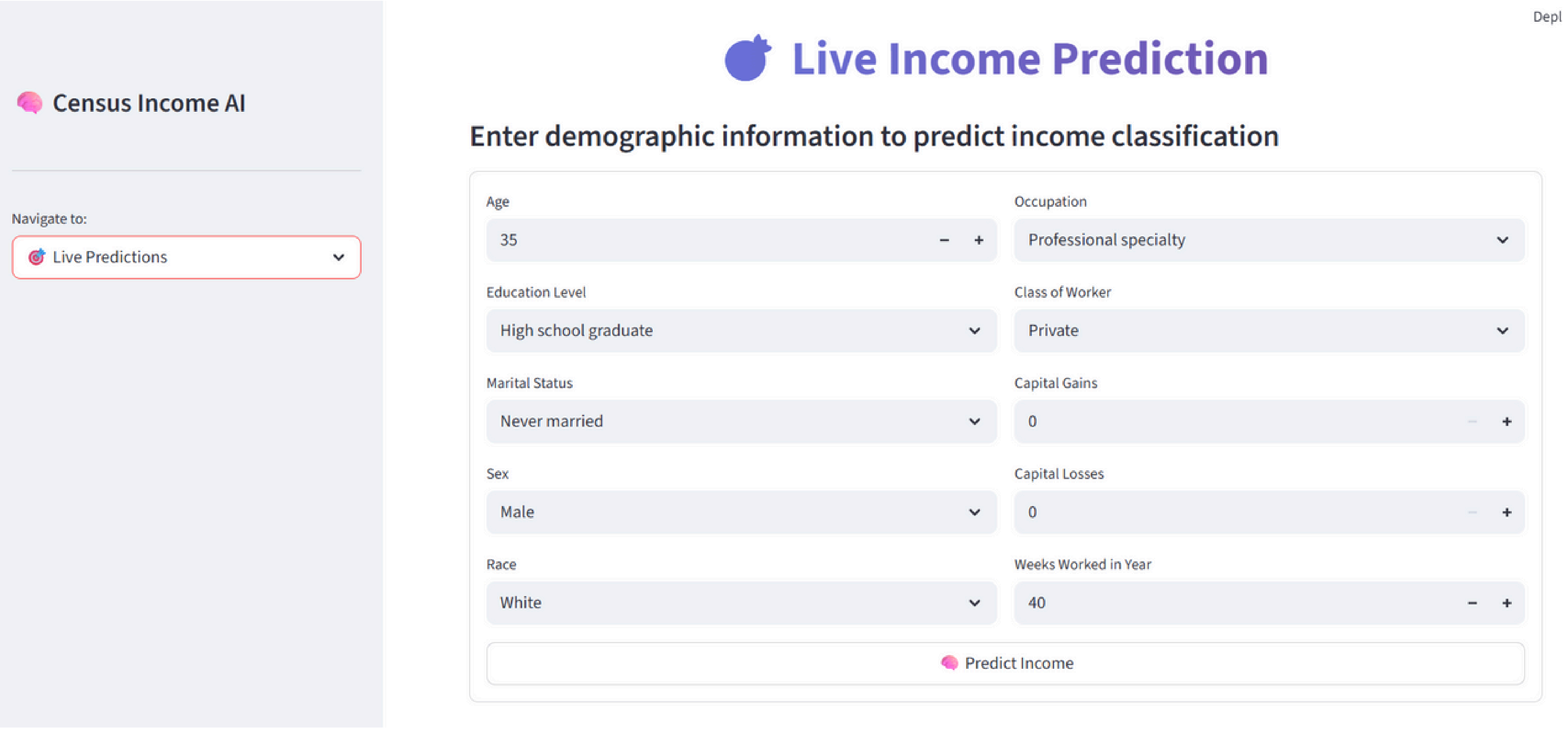
Riyadh | 16.07.2025

EVALUATION & DEPLOYMENT

DEPLOYMENT FEATURES

We designed a robust architecture for deploying the income prediction model in production.

STREAMLIT WEB INTERFACE



API ENDPOINTS

POST /api/v1/predict

Single prediction endpoint for real-time inference

```
# Request
{
  "age" : 42,
  "workclass" : "Private" ,
  "education" : "Bachelors" ,
  "occupation" : "Exec-managerial" ,
  "hours_per_week" : 45
  // Additional features...
}

# Response
{
  "prediction" : ">50K" ,
  "probability" : 0.92,
  "model_version" : "v1.2.3" ,
  "feature_importance" : {
    "education" : 0.35,
    "age" : 0.22,
    // Top features...
  }
```

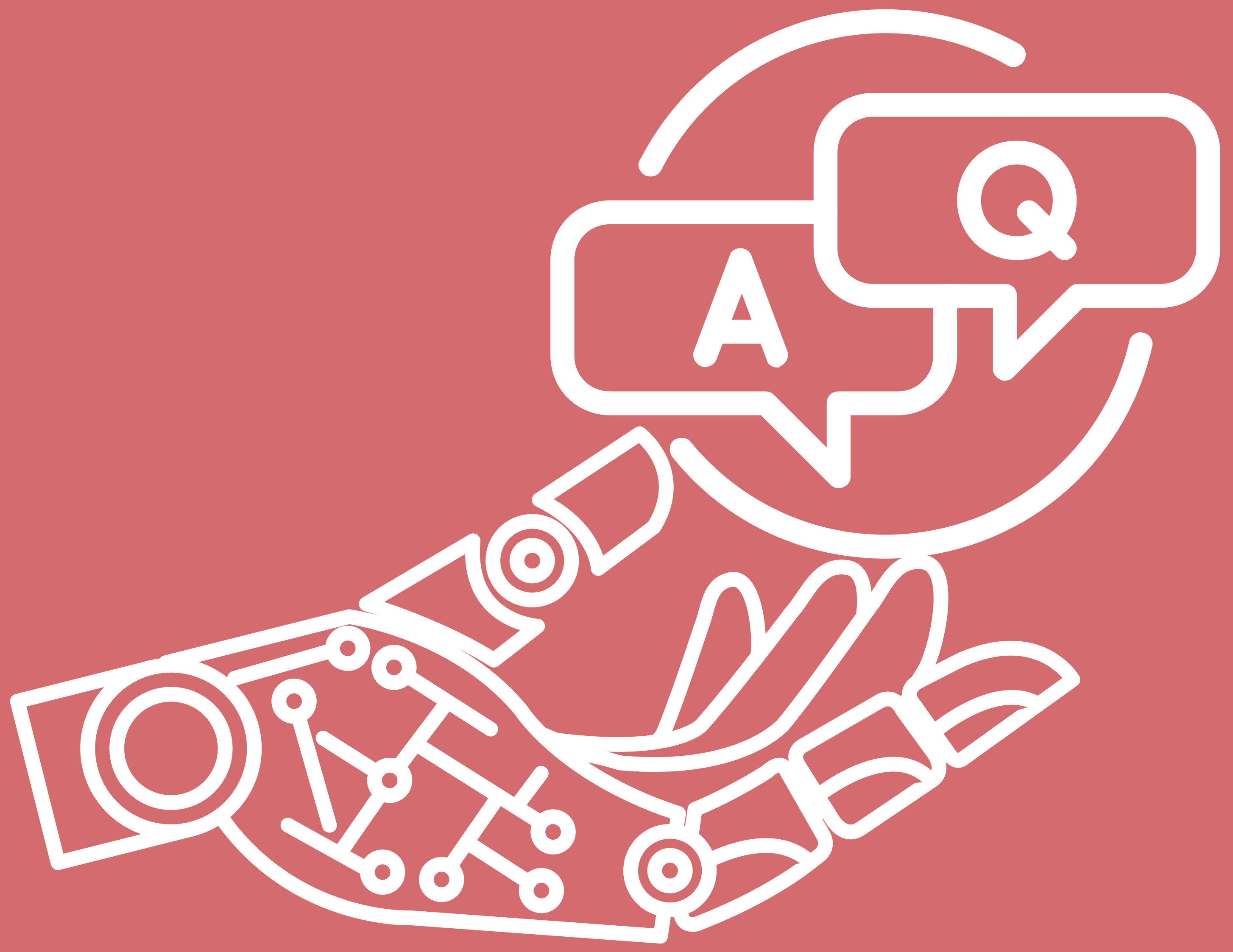
FUTURE RECOMMENDATIONS

EXTERNAL DATA INTEGRATION

APPLY MLOPS TECHNQUES

EXPLORE MORE ADVANCED MODEL ARCHITECTURES

THANK YOU


A
Q

DISCLAIMER

Some of the content was created with the help of GenAI models, but all provided information is my work output

EXTREME GRADIENT BOOSTING (XGBOOST)

- Model ensamble technique
- Univariate model
- Multivariate implementations

- Direct Multioutput
- Model 1: Given X, predict torque1
- Model 2: Given X, predict torque2
- ..
- Model 6: Given X, predict torque6

- Chained Multi-output;
- Model 1: Given X, predict torque1.
- Model 2: Given X and torque1, predict torque2.
- ..
- Model 6: Given X, predicted torque1, predicted torque2,. ., and predicted torque5, predict torque6.