Income Prediction Analysis: Insights and Recommendations

Maksym Solodarenko, PhD

January 26, 2025

Problem Statement

Understanding Income Predictors

- **Business Goal**: Identify characteristics associated with individuals earning more than $50,000 annually.
- **Key Challenges**:
    - Class imbalance in the dataset: Most individuals earn less than $50,000.
    - Need for interpretable and actionable insights for stakeholders.

Dataset Overview

Data Source and Composition

- **Dataset**: US Census Income Data
- **Training Set**: 200,000 rows
- **Testing Set**: 100,000 rows
- **Key Variables**:
  - Target: Income Class (<=50K or >50K)
  - Features: Age, Education, Marital Status, Capital Gains, Hours Worked, etc.
- **Challenge**: Highly imbalanced target variable (6% earning >50K).

Data Preprocessing Steps

Cleaning and Preparation Pipeline

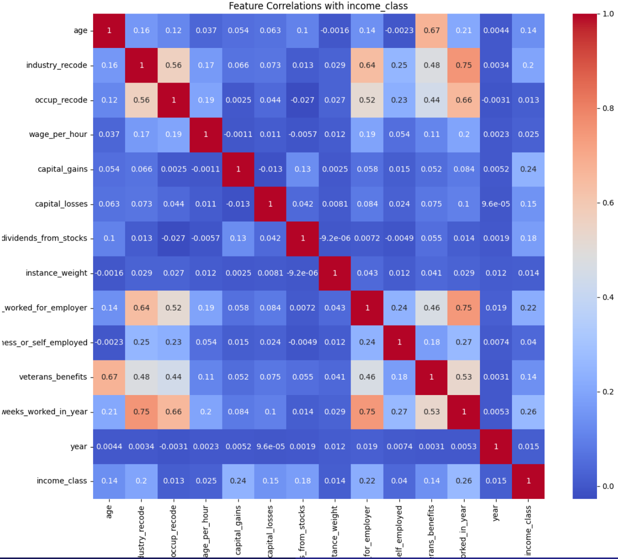
1. **Target Encoding**:
    - Converted income class into binary (1 for `>50K`, 0 for `<=50K`).
2. **Handling Skewed Features**:
    - Applied log-transformation to `capital_gains`, `capital_losses`, and `dividends_from_stocks`.
3. **Encoding Categorical Variables**:
    - Label encoded features like `marital_status` and `education`.
4. **Scaling Numeric Features**:
    - Standardized features like age and `hours_worked_per_year` using `StandardScaler`.
5. **Addressing Class Imbalance**:
    - Applied SMOTE (Synthetic Minority Oversampling Technique) to balance training data.

Exploratory Data Analysis

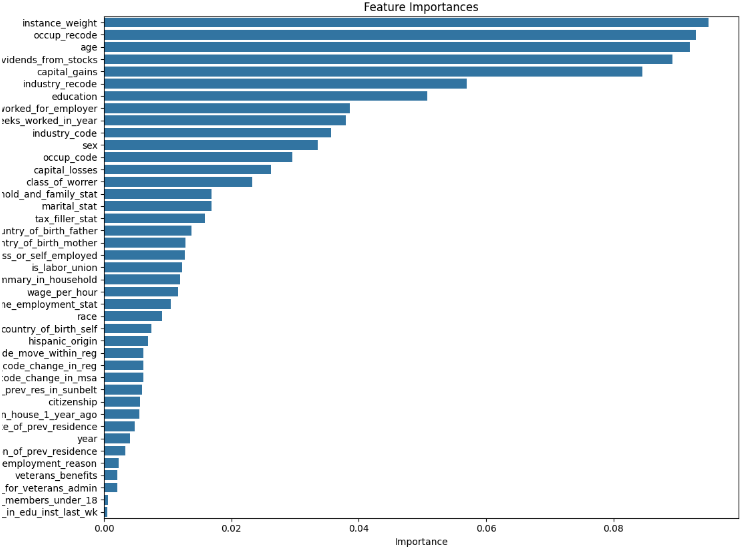
Key Insights

- **Feature Correlations with Income**:
  - `Number Of Weeks Worked In Year`: Strong positive correlation with income.
  - `Number Of People In Household Working`: A higher number of people working for an employer correlates with higher income.
  - `Capital Gains`: Significant predictor of earning `>50K`.
- **Class Distribution**:
  - Only $\approx 6\%$ of individuals in the training set earn `>50K`.
- **Feature Importance (Random Forest)**:
  - Top Features: Occupation, Age, Dividends From Stocks, Capital Gains, Industry, Education

Feature Correlations with Income



Feature Importance (Random Forest)



Models Evaluated

Algorithms and Performance

1. **Logistic Regression**
   - **Strengths**: Simplicity and interpretability.
   - **Accuracy**: 95%, but struggled with recall (28%) for >50K class.
2. **Random Forest**
   - **Strengths**: Non-linear relationships and feature importance.
   - **Accuracy**: 95%, improved recall (39%) and F1-score for >50K.
3. **XGBoost**
   - **Strengths**: Robust handling of imbalanced data.
   - **Accuracy**: 95%, slight decrease in recall (38%) for >50K.

Improvements through Hyperparameter Tuning

Random Forest with GridSearchCV

- **Optimized Parameters**:
  - `n_estimators`: 200
  - `max_depth`: 10
  - `min_samples_split`: 5
- **Performance**:
  - Accuracy: 90%
  - Recall for `>50K`: 76%
  - Highlight: Improved balance between precision and recall.

Ensemble Approach

Custom Soft Voting Ensemble

- **Method**:
  - Averaged probabilities from Logistic Regression, Random Forest, and XGBoost.
- **Performance**:
  - Accuracy: 82%
  - Recall for `>50K`: 92%
  - Precision for `>50K`: 24%
  - Tradeoff: Improved recall at the cost of precision.

Summary of Key Takeaways

- **Random Forest** was the best model overall, with a balanced tradeoff between precision and recall.
- **Feature Importance** highlighted age, education, and capital gains as key predictors.
- **Custom Ensemble** improved recall significantly but suffered from low precision.

Questions Discussion

Let's Collaborate!

- Questions about methodology?
- Suggestions for further improvements?