Census Income Classification Project

Adil Gursel Karacor, PhD.
Senior Data Scientist Candidate at DataIku

Project Overview

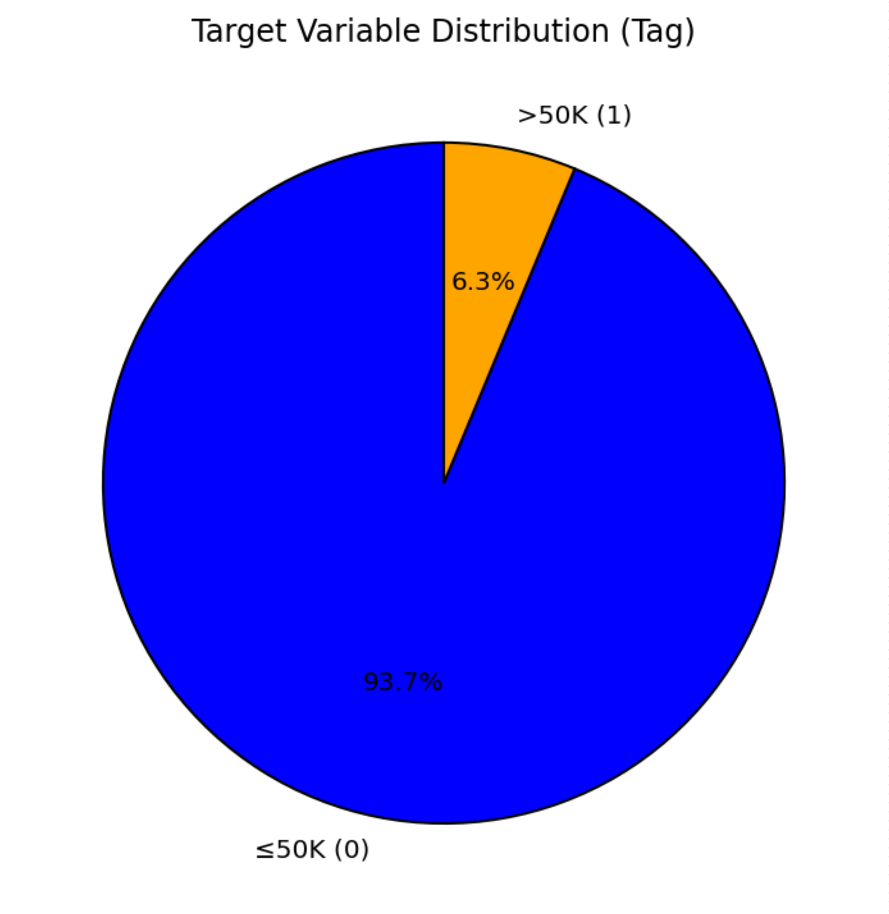
- Business Question
  - *“Which demographic and employment factors (US) best predict if an individual’s annual income exceeds $50K?”*
- Data Source
  - U.S. Census sample (~300K records)
  - Provided train/test CSVs with 40+ features.
- Project Objective
  - Develop a robust classification model to predict whether income > $50K
  - Identify which features drive income disparities.

Presentation Outline

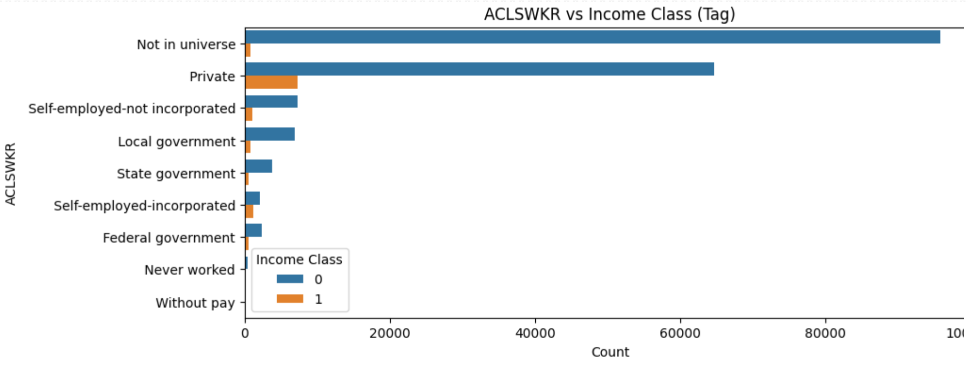
1. Exploratory Data Analysis (EDA)
    a. Check distribution, missing data, outliers.
    b. Visualize features.
2. Data Cleaning & Feature Engineering
    a. Resolve duplicates & conflicts.
    b. Encode or bin key variables.
3. Modeling
    a. Tried multiple classifiers (CatBoost, Random Forest, Logistic Regression).
    b. Address imbalance (AUC, F1).
4. Results & Evaluation
    a. Compare performance on validation set.
    b. Identify best model + important features.
    c. Deep dive into EDA insights, modeling strategy, final results, and recommendations.

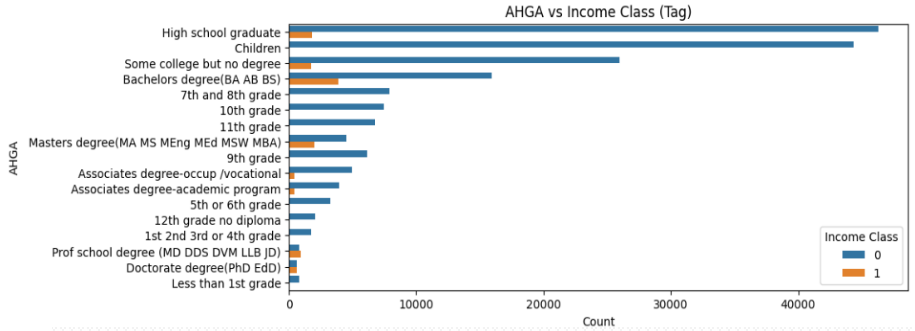
Data Overview & Key Distributions

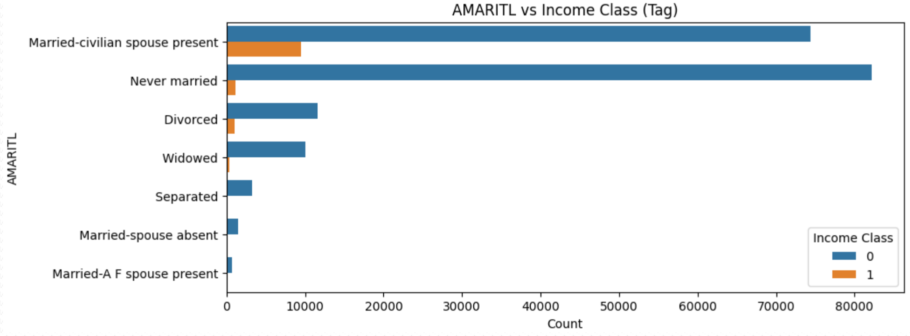
```
=== Loading Data ===
Initial training rows: 199523
Initial test rows: 99762

=== Duplicates Removed ===
Training duplicates removed: 3229
Test duplicates removed: 883

=== Conflict Rows Removed ===
Training conflicts removed: 379
Test conflicts removed: 119

=== Unexpected/Invalid Incomes ===
Training rows with invalid 'income': 0
Test rows with invalid 'income': 0

=== Final Summary ===
Initial training rows: 199523
Final training rows: 195915
Total removed from training: 3608  (Duplicates: 3229, Conflicts: 379, Invalid Income: 0)

Initial test rows: 99762
Final test rows: 98760
Total removed from test: 1002  (Duplicates: 883, Conflicts: 119, Invalid Income: 0)

Training 'Tag' distribution after cleaning:
Tag
0    183627
1     12288
Name: count, dtype: int64

Test 'Tag' distribution after cleaning:
Tag
0    92612
1     6148
Name: count, dtype: int64
```

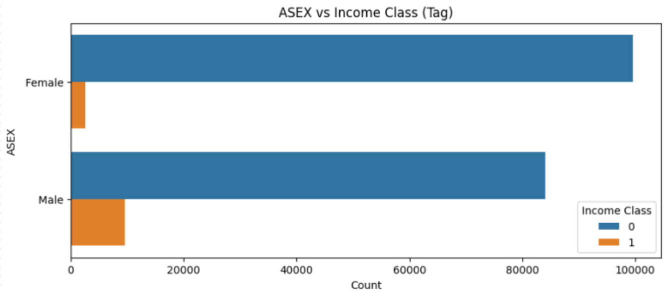


Categorical Variables Insights

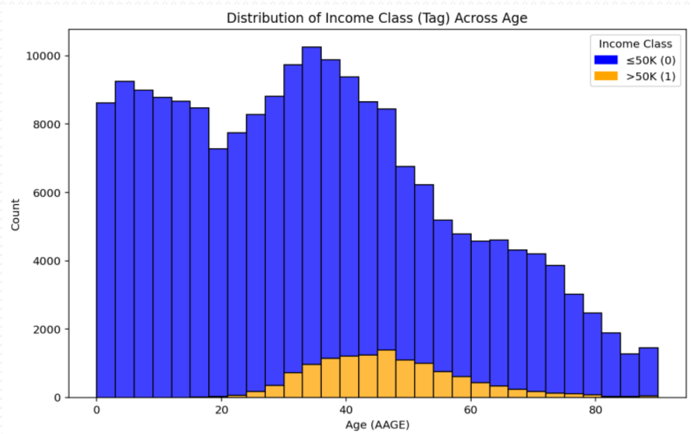


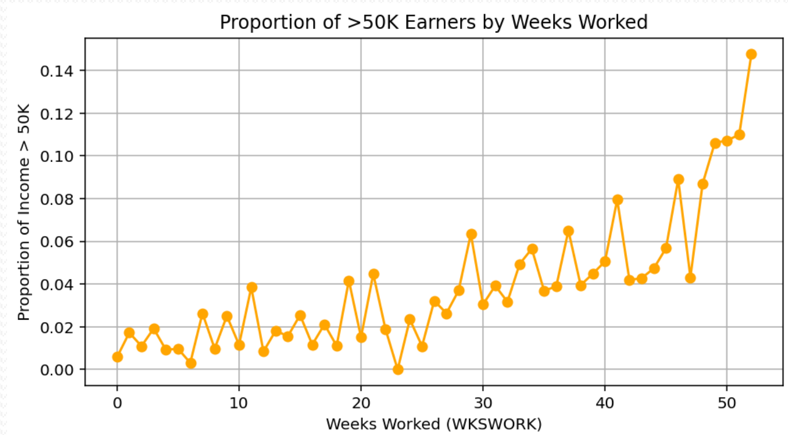


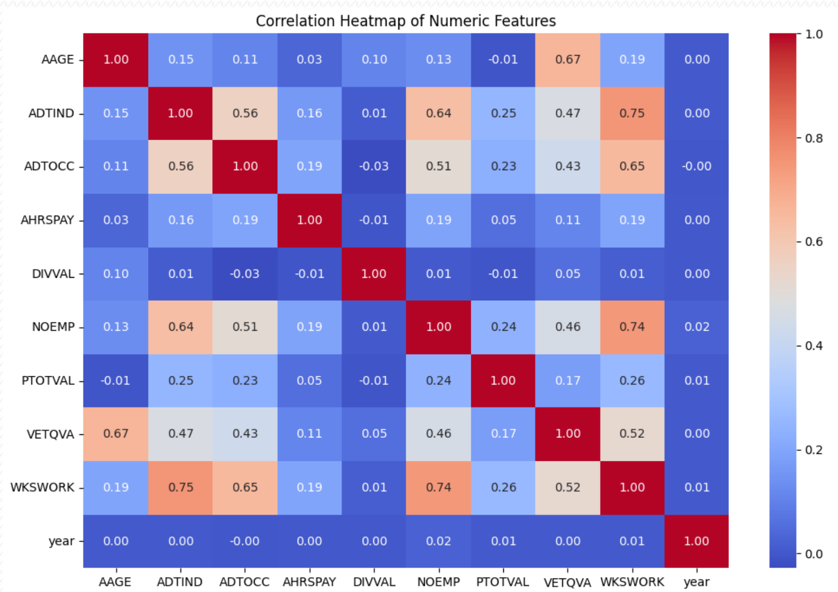




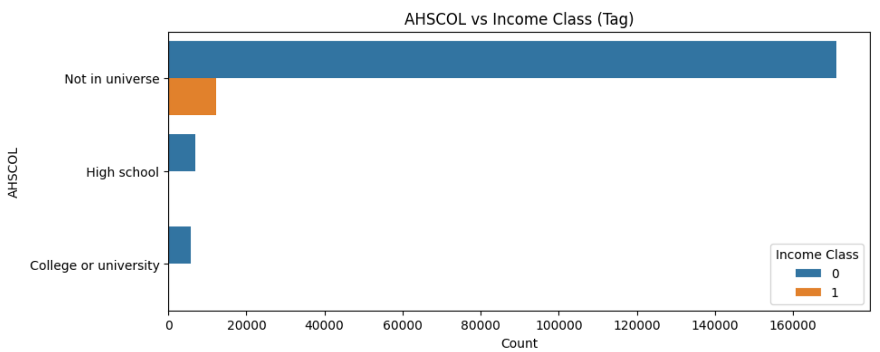
Numeric Variables & Correlations

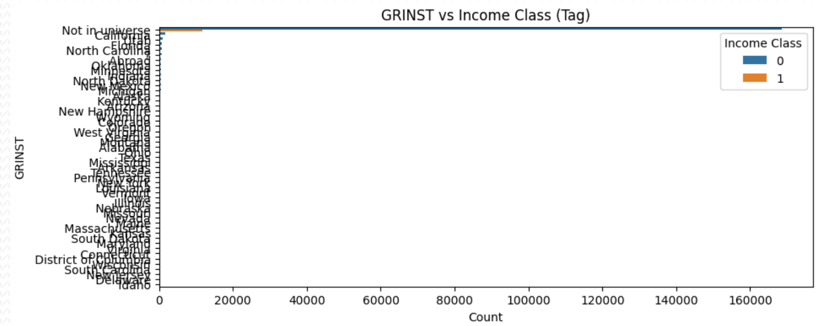


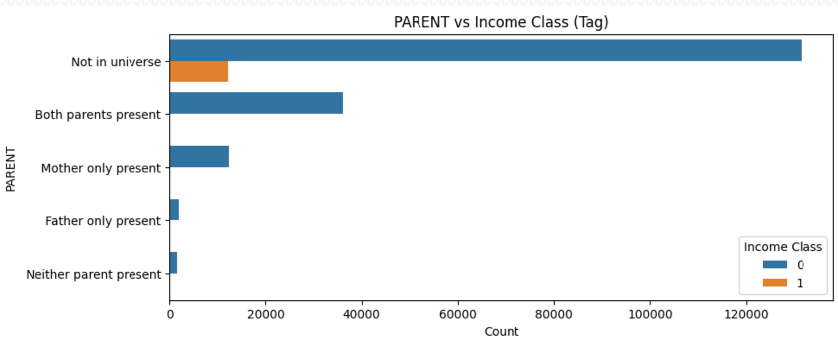


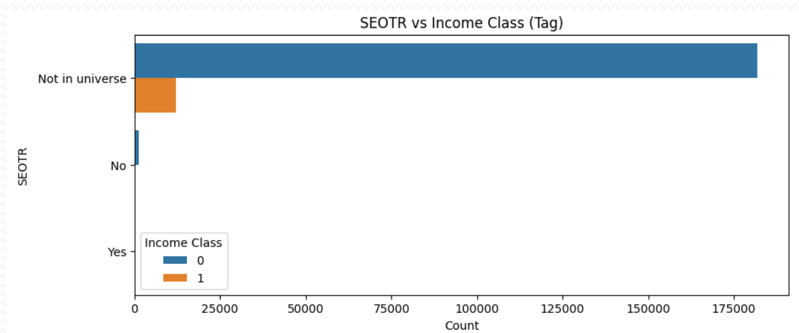


Feature Engineering – Dropped Features









Feature Engineering – Transformed Features

- **Reorganized HHDFMX (Household Relationship):**
  - Retained key categories: Householder, Nonfamily householder, Secondary individual, Spouse of householder
  - Grouped all others as 'Other'
  - **Why?** Reduces sparsity and improves model generalization

- **Created 'capgainloss' (Net Capital Gain/Loss):**
  - Computed as: **CAPGAIN - CAPLOSS**
  - Dropped original CAPGAIN & CAPLOSS to remove redundancy
  - **Why?** Captures financial impact in a single feature

Model Selection & Rationale

- **Logistic Regression**
  - Pro: Simple, interpretable coefficients, good baseline
  - Con: Assumes linear relationships; might underfit complex data.
- **Random Forest**
  - Pro: Robust to outliers, handles non-linearities well, less feature engineering
  - Con: Can be slower for large data if trees are deep, less interpretable than linear models.
- **Catboost**
  - Pro: Categorical feature handling, strong performance on tabular data, well-suited to imbalanced classification
  - Con: More complex to tune, can be memory-intensive.
- From simple linear to advanced tree-based approaches

Hyperparameter Tuning & Cross-Validation

- **Hyperparameters**
  - **Logistic Regression:** C (inverse regularization strength), solver
  - **Random Forest:** n_estimators, max_depth, min_samples_split
  - **CatBoost:** learning_rate, iterations, depth, l2_leaf_reg
- **Cross-Validation**
  - **Stratified K-Fold** (5-fold) → ensures balanced class splits.
  - **Metric: ROC AUC** due to imbalanced classes. Also tracked **F1** and **precision/recall**.
- **Tuning Method: Manual**

Model Performance Comparison

- Catboost

```
--- Final Model Performance on Validation Set ---
ROC AUC: 0.9581
Accuracy: 0.9588
Precision: 0.7695
Recall: 0.4834
F1 Score: 0.5938

Confusion Matrix:
[[91722   890]
 [ 3176  2972]]
```

- Random Forest

```
--- Final Model Performance on Validation Set ---
ROC AUC: 0.9399
Accuracy: 0.9455
Precision: 0.8692
Recall: 0.1470
F1 Score: 0.2515

Confusion Matrix:
[[92476   136]
 [ 5244   904]]
```

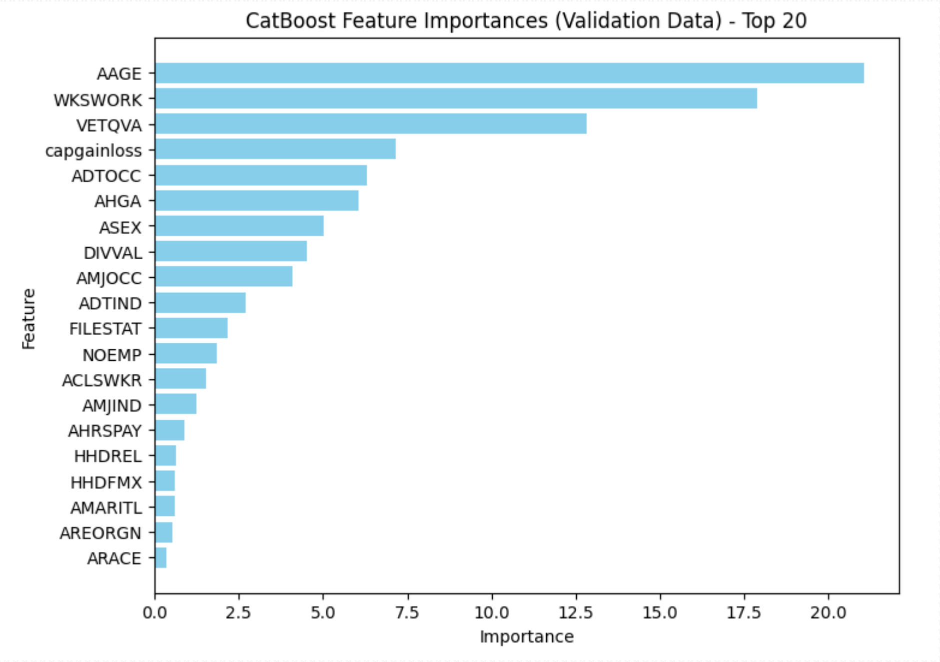
- Logistic Regression

```
--- Final Model Performance on Validation Set ---
ROC AUC: 0.9274
Accuracy: 0.9470
Precision: 0.6989
Recall: 0.2598
F1 Score: 0.3788

Confusion Matrix:
[[91924   688]
 [ 4551  1597]]
```

CatBoost outperformed on AUC, followed by Random Forest, then Logistic Regression.

Results Interpretation

- Top features
  - Age
  - Weeks Worked
  - Veteran's Benefits
  - Net Capital Gains/Losses (engineered)



Potential Policy & Business Usage

- Targeted Workforce Development
  - Use key features affecting income (e.g., age, education, occupation) to develop training programs that improve employability and earnings potential.
- Personalized Financial Advice
  - Utilize predictions to offer tailored financial counseling, particularly in managing capital gains, investment strategies, and veteran benefits.
- Equity and Inclusion Efforts
  - Pinpoint workforce segments that rarely exceed $50K income, enabling targeted interventions to address income inequality.

Potential Policy & Business Usage

- Efficient Resource Allocation
  - Optimize welfare programs or business strategies by identifying populations at risk of lower earnings, ensuring precise resource distribution.
- Human Resource Optimization
  - Businesses can leverage predictive insights to enhance employee retention, career development, and salary growth initiatives.

Next Steps & Possible Improvements

- Collect Additional Data
  - Explore more detailed employment history, educational specializations, or geographic factors.
  - Incorporate real-time economic indicators (e.g., recession/unemployment data) to enhance model.
- More Advanced Feature Engineering
  - Derivate more variables from the most important features.
  - Investigate interactions (e.g., age × occupation).
  - Refine binning or transformations for heavily skewed variables.

Next Steps & Possible Improvements

- Advanced Modeling Approaches
  - Try ensemble methods combining multiple algorithms.
- Fairness & Bias Analysis
  - Evaluate potential bias in predictions across demographic groups.
  - Mitigate issues with fairness-aware algorithms.

Conclusion & Key Takeaways

- Age (AAGE), Weeks Worked (WKSWORK), VETQVA, and Capital Gains/Losses (capgainloss) are the strongest predictors of income.
- Individuals around 50 who work full weeks and have positive capital gains have a significantly higher chance of earning >$50K.
- The class imbalance impacted modeling; AUC was a key metric to handle this.

Conclusion & Key Takeaways

- **CatBoost** performed best.
- Feature engineering (e.g., reorganized HHDFMX, capgainloss) improved the model's predictive power.
- Additional feature interactions and external economic data could further enhance predictions.

Conclusion & Key Takeaways

- **Main Recommendation:**
  - Implement the CatBoost model for income classification with potential refinements.
  - Use the model for **targeted policy-making, workforce planning, or financial decision-making**.
  - Consider periodic retraining to account for economic shifts and workforce changes.