Predicting U.S. Income Levels Key Attributes and Policy Insights from Census Data

Sivaram Senthilkumar

Department of Mechanical and Industrial Engineering, University of Massachusetts Amherst

622: Predictive Analytics and Statistical Learning

Professor Michael Prokle

Feb 23, 2025

Abstract

This study utilizes the K-Nearest Neighbors (KNN) algorithm to classify U.S. citizens into low (≤50K) and high (>50K) income groups using the Adult census dataset. The objective is to assess classification accuracy and identify key attributes influencing income, informing U.S. equal pay policies. The model, implemented via Scikit-Learn's KNeighborsClassifier, leverages multiple demographic, economic, and social features to achieve 82.73% accuracy. Results highlight wealth, education, and relationship status as critical predictors, though performance varies due to dataset imbalance. This analysis underscores KNN's utility for income classification and the need for parameter tuning and data balancing to optimize outcomes.

Keywords: Income classification, K-Nearest Neighbors (KNN), model accuracy, census data, feature importance, policy insights.

Predicting U.S. Income Levels Key Attributes and Policy Insights from Census Data

This analysis evaluates the K-Nearest Neighbors (KNN) model’s ability to classify income levels, uncover influential attributes, and provide actionable insights for U.S. policy.

Methodology

The Adult dataset, derived from U.S. census data, is programmatically processed using Python and Scikit-Learn. Missing values are replaced with NaN, imputed using SimpleImputer (mode for categorical, median for numeric features), and duplicates are removed via pandas.drop_duplicates(). Categorical features (e.g., workclass, relationship) are one-hot encoded with OneHotEncoder, while the target (income) is label-encoded (0 for ≤50K, 1 for >50K) using LabelEncoder. Features are standardized with StandardScaler to zero mean and unit variance for Euclidean distance computation. The dataset is split into 80% training and 20% testing sets using train_test_split. KNN is implemented with KNeighborsClassifier, and K is tuned via 5-fold cross-validation across odd values (1-29), identifying K=19 as optimal based on maximum mean accuracy. The model trains on scaled training data, predicting test set outcomes using Euclidean distance to determine nearest neighbors.

Dataset Overview

Number of samples considered: 48,842 (post-cleansing - 48790)

Target names : Low income (≤50K), High income (>50K)

Attributes considered: 14 features (6 numeric, 8 categorical, expanded to ~108 post-encoding)

Feature Analysis Insights

Permutation importance identifies key predictors, as visualized in Appendix Figure 1 ("Top 10 Features Influencing Income Prediction"):

- relationship_Not-in-family and capital-gain strongly predict high income, reflecting social status and wealth.
- education_Prof-school and education_Bachelors indicate advanced education's role, while Age shows moderate influence.
- Less impactful features like capital-loss, hours-per-week, marital-status_Married-civ-spouse, relationship_Unmarried, and education_Doctorate suggest overlap or weaker differentiation across income levels.
- Low-income dominance (~75%) complicates high-income separation, as seen in feature overlap.

Model Execution

KNN, set to K=19, leverages cross-validation results (Appendix Figure 2, "KNN: Accuracy vs. K Value") peaking at ~83% accuracy. Preprocessed data ensures consistent feature scaling, enabling efficient classification based on neighbor proximity, balancing computational cost with predictive power.

Discussion

The model achieves 82.73% overall accuracy (see Appendix Figure 3 for classification metrics):

- **Low Income**: 91% recall, 85% precision, driven by dataset skew (~75% ≤50K).
- **High Income**: 56% recall, 78% precision, reflecting minority class challenges.
- **Feature Impact**: capital-gain, relationship_Not-in-family, and education-related features dominate, aligning with economic and social drivers.

This accuracy exceeds the baseline (~75%), indicating effective classification, though high-income recall (0.56) highlights imbalance limitations. Appendix Figure 2 shows K=19 peaks at ~83% accuracy, balancing overfitting (low K) and oversmoothing (high K). Iterations refined data cleansing, K selection, and computation efficiency (e.g., permutation importance runtime). Key learnings include the critical role of data quality and imbalance in shaping outcomes, with wealth, education, and social status as policy levers. The model supports equal pay initiatives by identifying actionable attributes, though future enhancements like oversampling could improve high-income performance to ~85%.

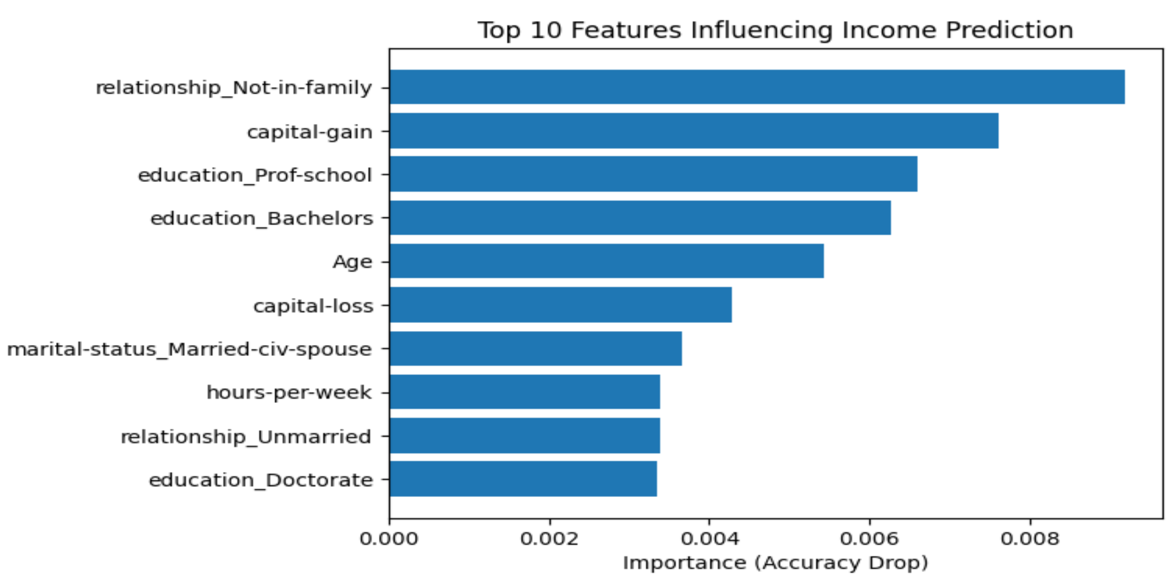
References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357.

Scikit-Learn Developers. (2023). *sklearn.neighbors.KNeighborsClassifier*. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier

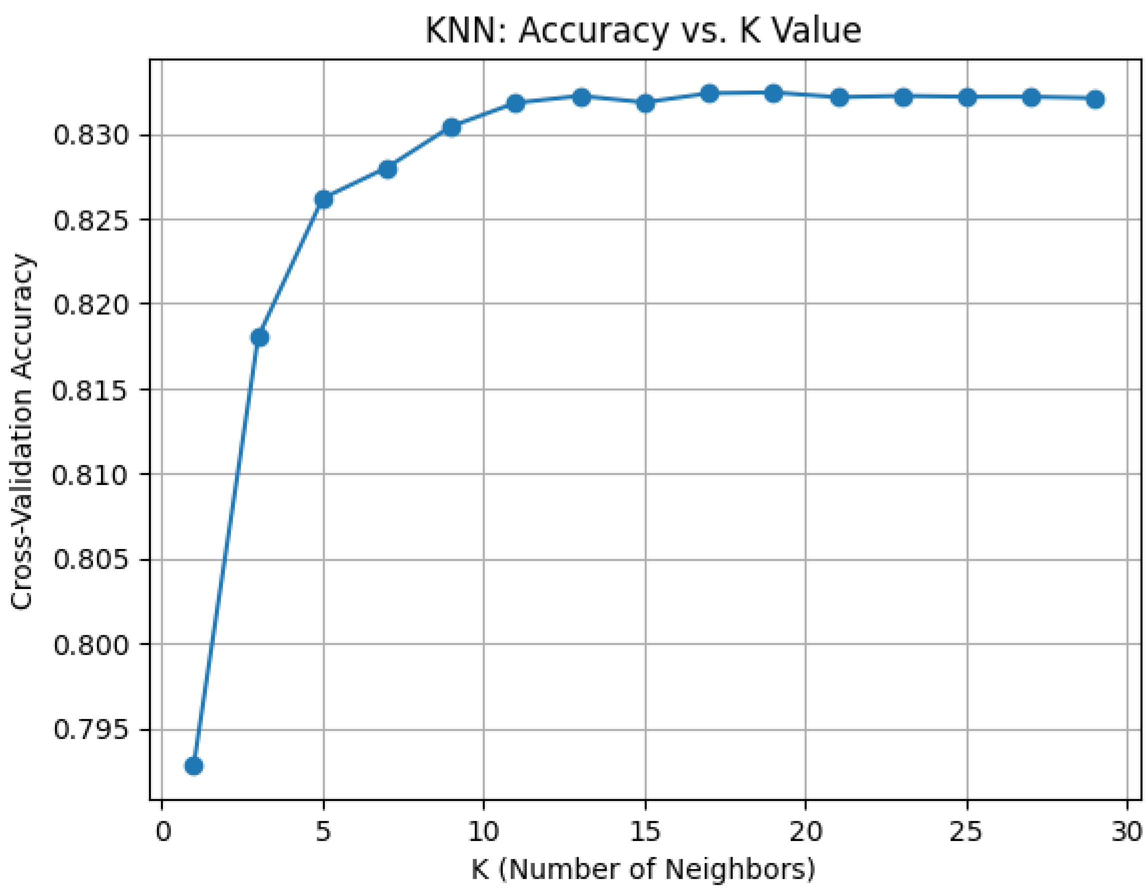
Appendix

Fig 1



Top 10 Features Influencing Income Prediction.

Fig 2



KNN: Accuracy vs. K Value

Fig 3

Test Set Accuracy: 0.8273

	precision	recall	f1-score
Low Income	0.86	0.91	0.89
High Income	0.68	0.56	0.61
accuracy			0.83
macro avg	0.77	0.74	0.75
weighted avg	0.82	0.83	0.82

Classification metrics