Income Prediction Project Presentation

Leveraging Data to Predict Income Thresholds

Stakeholders and Problem Statement

- The goal of this project is to predict whether an individual's income is above or below a certain threshold.

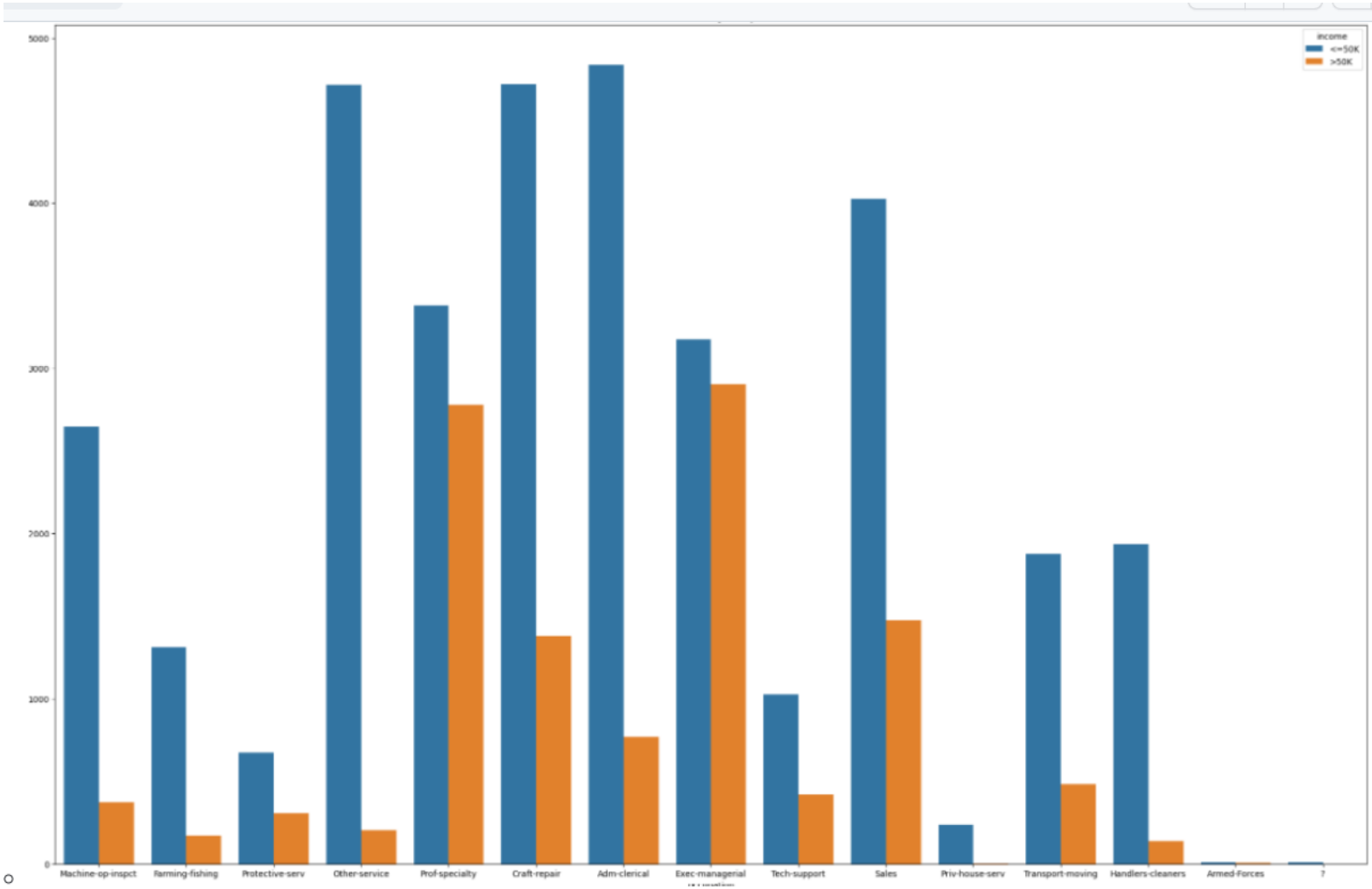
- The stakeholders include organizations or individuals who are interested in understanding the factors that influence an individual's income and want to use this prediction for decision-making.

Data Overview

- The dataset includes various features such as education level, age, gender, occupation, and more, which are believed to influence an individual's income.

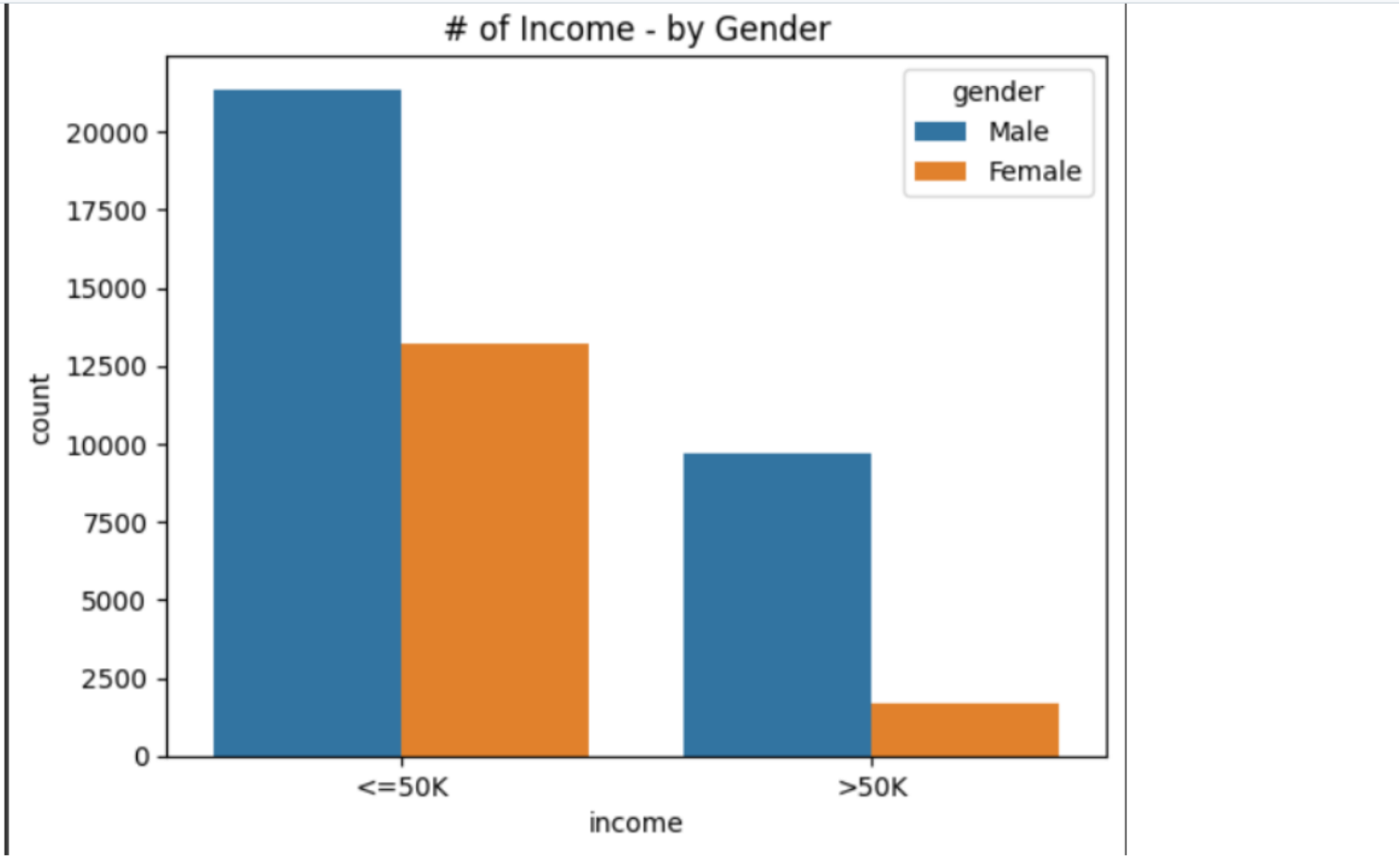
- The target variable is whether the income is above or below a threshold.

Key Finding 1: Education Matters



- The distribution of educational-num is shifted to the right for individuals earning >50K, indicating a higher educational level.

Key Finding 2: Age and Income



Visualizing the income distribution showed that a significant portion of individuals have incomes below the threshold, which could impact the model's

Model Performance

- **DecisionTreeClassifier**
- Training Data:
  - Accuracy: 1.00
- Test Data:
  - Accuracy: 0.81
- **LogisticRegression**
- Training Data:
  - Accuracy: 0.85
- Test Data:
  - Accuracy: 0.85
- **Model with PCA (Logistic Regression)**
- Training Data:
  - Accuracy: 0.83
- Test Data:
  - Accuracy: 0.84

Model Evaluation

- Based on the accuracy scores, the DecisionTreeClassifier achieves the highest accuracy on the training data, but it has the lowest accuracy on the test data, suggesting potential overfitting.

- The PCA with DecisionTreeClassifier also shows signs of overfitting, as it performs perfectly on the training data but significantly worse on the test data.

- Among the remaining models, both the GridSearchCV model with DecisionTreeClassifier and the LogisticRegression model have similar accuracy scores on both the training and test data.

- These two models generalize well and have a balance between precision and recall for both classes.

Which Model

Given that the primary business goal is likely to have a model that generalizes well to new, unseen data, the production model choice should be based on the balance between performance on the test data and avoiding overfitting. In this case, the LogisticRegression model would be a suitable choice for production.

Recommendations

- Feature Importance:
  - Provide stakeholders with insights into the most important features that influence an individual's income. This can help them target specific areas for interventions or improvements.

- Education Initiatives:
  - Based on the analysis, education level was found to be a significant predictor of income. Stakeholders could consider investing in education initiatives to help individuals improve their education and potentially increase their income.

- Targeting Age Range:
  - Individuals in the age range where the probability of earning >50K is higher could be targeted for higher-income financial products or services.

Conclusion

- This project aimed to predict an individual's income using various features from the Adult Income Dataset.
- The analysis provided insights into the relationships between features and income, and the predictive models helped in achieving accurate income predictions.
- The recommendations provided can guide stakeholders in making informed decisions to address income-related challenges.

Q&A