Overview

The goal of this project was to predict whether individuals make over $50,000/year using 1996 US census data. For this binary classification problem, I created a logistic regression model to maintain the interpretability of its predictions. The logistic regression model achieved an accuracy of 85.7% on unseen data. Additionally, I created an initial random forest model which achieved a higher accuracy of 86.27% but lacks interpretability. The random forest model can however be tuned to improve predictive power.

Data Exploration

The dataset comprises 48,842 individuals with 14 variables concerning education level, race, occupation, etc. The variables have several interesting relationships with the target. One of these relationships is shown in Figure 1. There was an approximately equal number of individuals in the data whose reported occupation was “Craft-repair” or “Exec-managerial,” however, the proportion of these individuals who make over $50,000/year is appreciably higher for Exec-managerial individuals, see the underlined columns.

A graph with different colored bars

Description automatically generated

Figure 1: Distribution of those who make over $50,000/year (blue) or not (red), by occupation.

Data exploration drove the choice to implement a logistic regression model. It preserves insights often lost in more predictive machine learning models.

Data pre-processing

Before modeling, I separated the data into training, validation, and test sets to assess model performance on new data. I then investigated any separation issues in the training data. I found several variable values that had separation concerns with low observation counts. I solved this problem by combining categories and strategically binning variables based on their relationship to the target. After I addressed the separation concerns, all variables were categorical and ready to be added to a logistic regression model.

Modeling

I selected variables to include in the model through backward selection, where non-useful variables were dropped. This process created a model with 12 of the 14 variables. I found two pairs of variables were closely correlated, resulting in erroneous model predictions: marital status and relationship and work class and occupation. I removed the least predictive variable from each pair, marital status and work class, and was left with a final model with 10 variables, see Table 1. It is important to note that information such as race and country of origin were included in this model and the random forest model. This can easily be altered if there are any ethical concerns.

Table 1: Variables included in the logistic regression model and their description.

|  |  |
| --- | --- |
| Variable | Description |
| Age | An individual’s age |
| Education num | An individual’s current education level |
| Capital gain | Post-social insurance income, in the form of capital gains |
| Capital loss | Post-social insurance income, in the form of capital losses |
| Hours week | The number of hours per week an individual worked |
| Country name | An individual’s native country |
| Occupation | An individual’s occupation |
| Race | An individual’s race |
| Relationship | An individual’s relationship status (Husband, Wife, Unmarried, etc.) |
| Sex | An individual’s sex (Male or Female) |

Results

The logistic regression model yielded an accuracy of 85.7% on the test dataset, but the true power of logistic regression is in its interpretable coefficients. For example, an individual 30 years old or older has 279% higher odds of making over $50,000/year than someone 29 years old or younger. A similar interpretation can be applied to all variables in the model. Since predictive power is also important, I created a machine learning model, random forest, to see if the model accuracy could be improved. Indeed, the random forest model yielded an accuracy of 86.27% on the test dataset.

Conclusion

I created a logistic regression model and an initial random forest model to predict whether an individual earns over $50,000/year based on census data. The random forest model slightly outperforms the logistic regression model's accuracy and can be further tuned to improve predictive power. The logistic regression model has the added benefit of having interpretable variables that could aid decision-making targeting specific population segments.