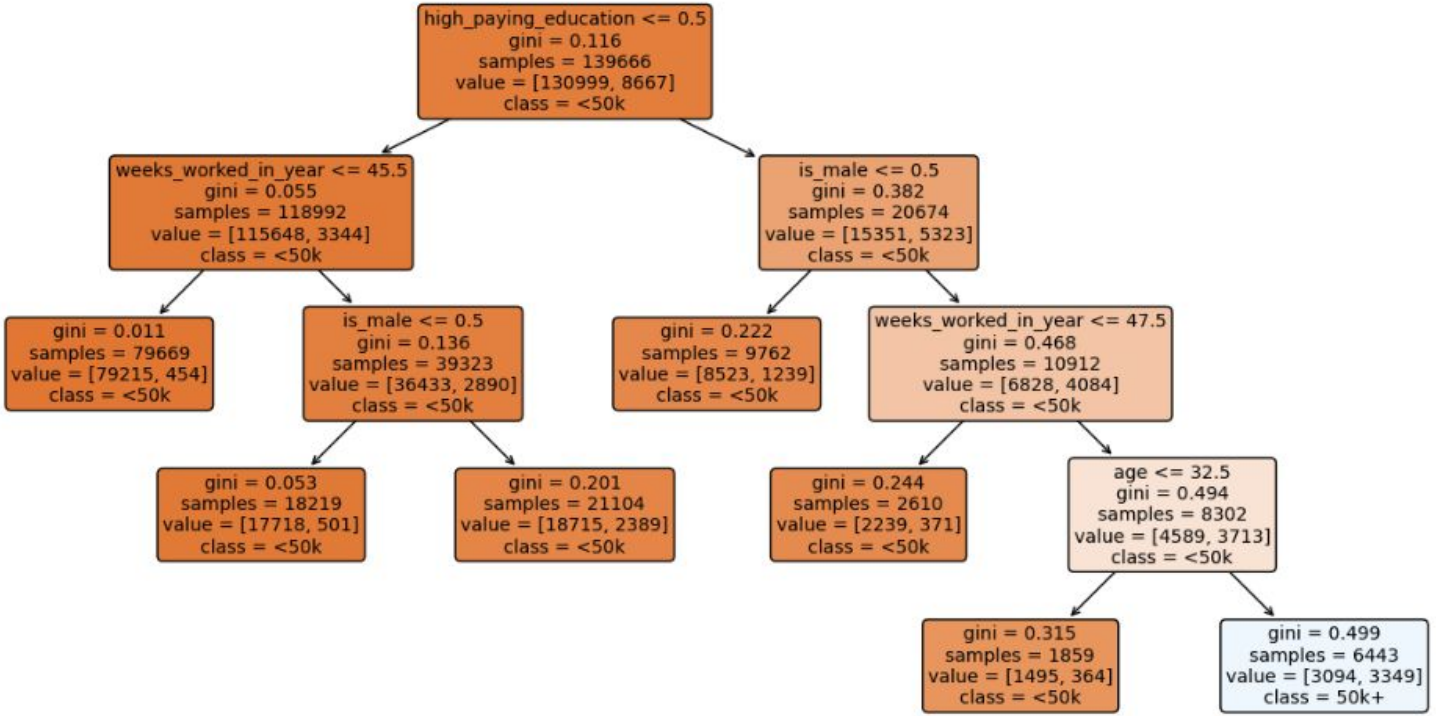
Census Income Classification

Jeff Mullahey
1/4/2023

Final Model Selected - Decision Tree



Target Variable:

Income $50,000+

Does the individual have an annual income of at least 50k per year?

Only 6% of the data did.

Predictor Variables:

1. Education

Having a bachelor's degree or higher was the strongest predictor of a $50k+ income

2. Gender

Male's had higher incomes than females on average

3. Weeks worked in year

Working less than 46 weeks a year led to a lower income

4. Age

Being over 32 years of age led to a higher income

Model Performance

Overall Accuracy:

94.0%

Confusion Matrix:

		Actual	
		<50k	50k+
Predicted	<50k	54,813	2,242
	50k+	1,329	1,473

Sensitivity: 40%

Meaning the model correctly identified 40% of the 50k+ group

Data Dictionary

Link to data dictionary

Link to provided metadata

Data Exploration Steps

---

1. Look at relationships between continuous inputs and categorical target
    a. Age seemed to have the only useful relationship
2. Look at relationships between categorical inputs and categorical target
    a. Does one class within a categorical variable have a higher proportion of 50k+?
    b. Education had the strongest relationship

Data Cleaning Steps

---

1. Add column names
2. Confirm input variables did not have missing values
3. Split data into training and testing sets

Competitor Model - Logistic Regression

Overall Accuracy:

94.2%

Confusion Matrix:

		Actual	
		<50k	50k+
Predicted	<50k	55,641	2,953
	50k+	501	762

Sensitivity: 21%

Meaning the model correctly identified 21% of the 50k+ group

Model Output:

Variable	Coefficients	P-Value
has_capital_gains	2.2	0
high_paying_own_business_self_employed	0.69	0
is_male	-0.97	0
high_paying_marital_stat	1.01	0
high_paying_education	1.9	0
weeks_worked_in_year	0.005	0
age	-0.09	0

Next Steps

---

1. More advanced classification models
2. New data
3. Tune parameters
4. Adjust classification threshold
5. Over/under sampling techniques
6. Use all classes of categorical variables for creating dummy variables (not just high vs low)