Adult Census Income

Edgar Galindo
Kevin Parton

Dataset

- Adult Census Income data set from Kaggle
- 32,561 samples 14 features.
- Binary label
- Predict whether income for a certain individual exceeds $50,000 a year based on census data.
- Data set contains features that are useful indicators of one's earning potential.
- Some of the more important features are race, education, occupation, and age.

Visualizations:

Converting categorical feature data

Education

15 => prof-school
14 => masters
13 => doctorate
12 => bachelors
11 => associate degree
10 => vocational school
9 => some college
1-8 => 1st to high school
0 => preschool

Gender

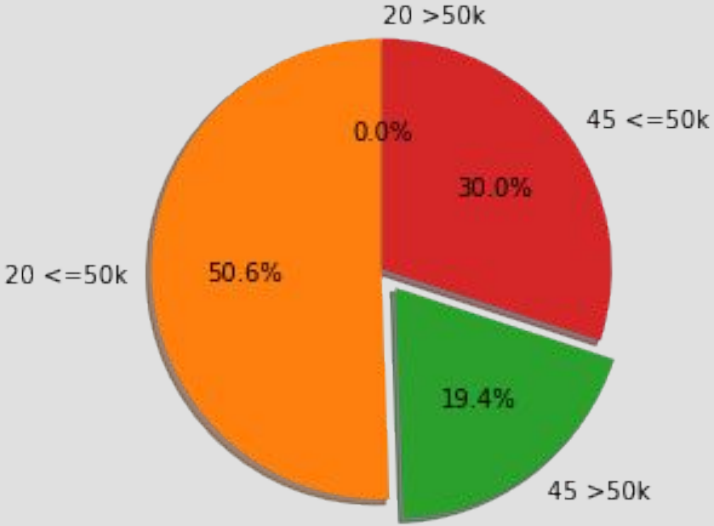
1 => male

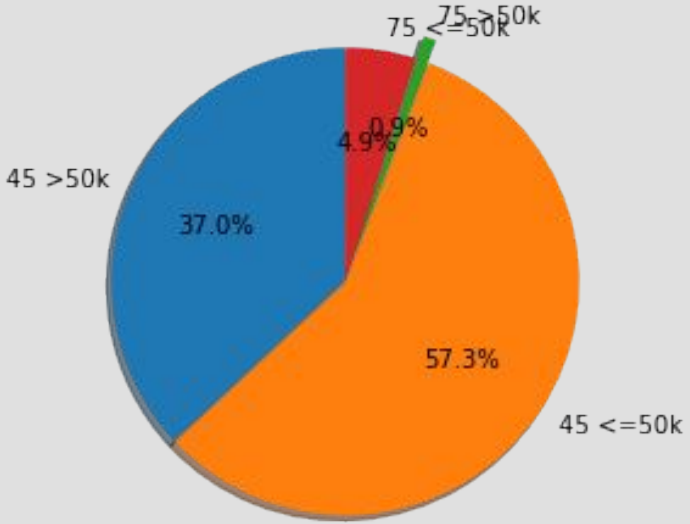
2 => female

Race

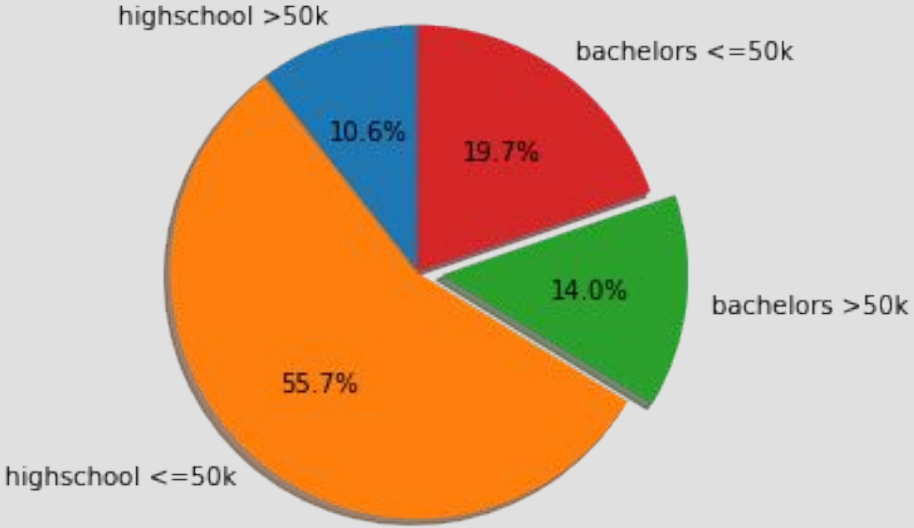
4 => white
3 => asian-pac-islander
2 => other
1 => black
0 => indian-eskimo

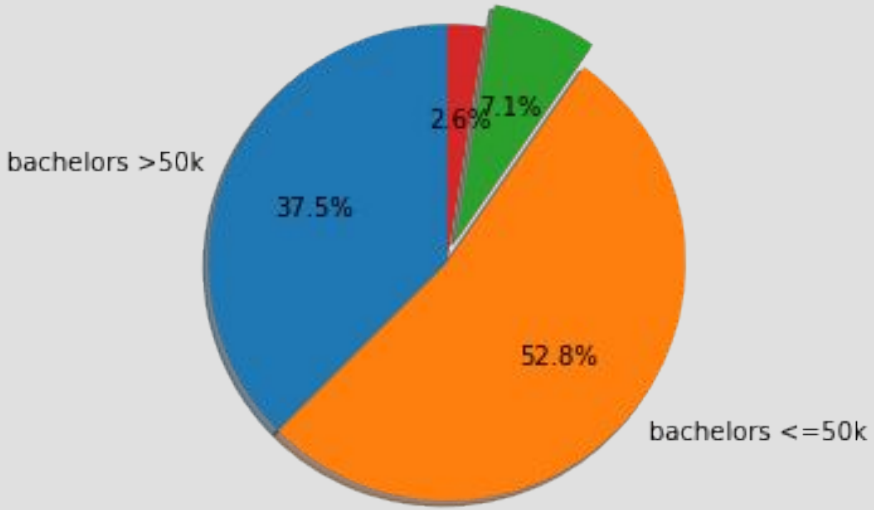
Age pie chart



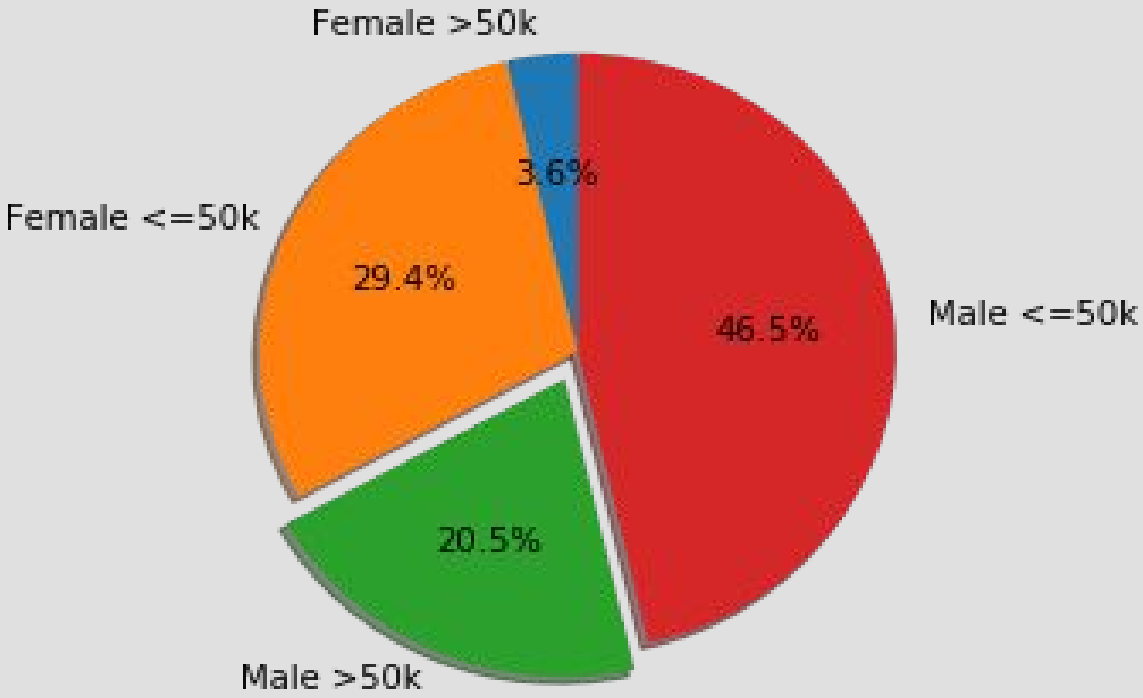


Education pie chart

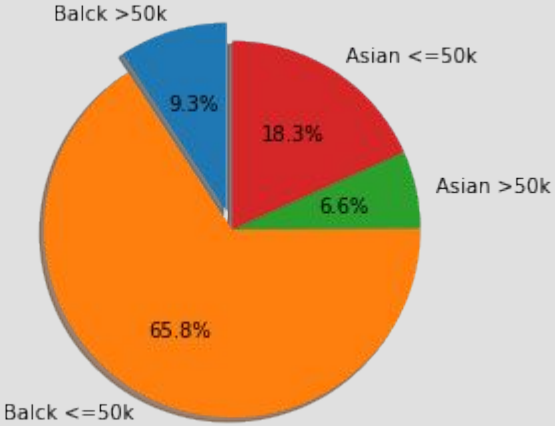


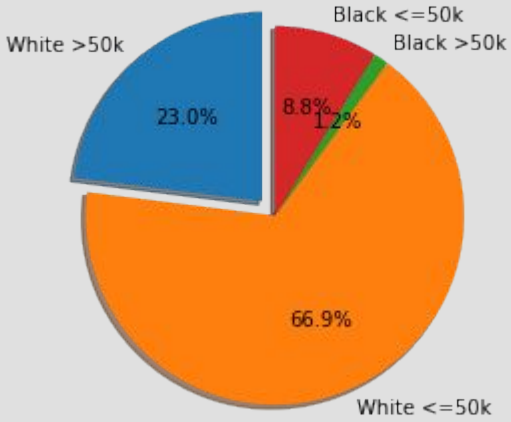


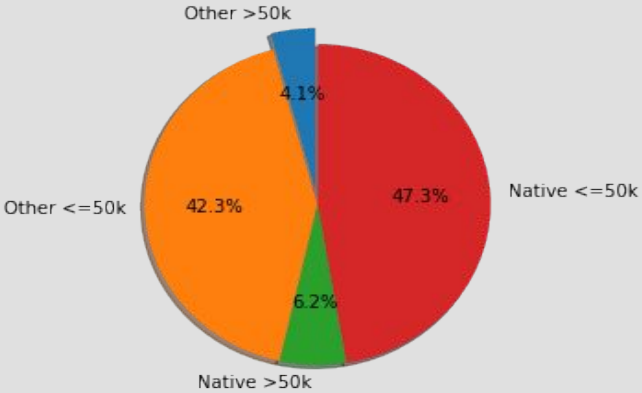
Gender pie chart



Race pie chart







Data Preprocessing

- Check data for missing values
  - Found ‘?’ in place of missing values
- Remove columns with more than 20% missing values
  - 27 rows removed
- Impute data
  - Replace with other.column_name
- One hot encoding
  - From 14 columns to 102
- Scale data
  - From -1 to 1

Results

	21% '3'	14% '2'	
• Logistic Regression:	83.12%	82.84%	82.89%
• Log Regression CV:	83.12%	82.71%	83.12%
• Decision Tree:	81.31%	81.00%	81.25%
• Decision Tree CV:	81.80%	81.33%	81.73%
• Bagging:	82.58%	82.39%	82.62%
• Bagging CV:	82.75%	82.37%	82.78%
• AdaBoost:	82.11%	81.83%	81.91%
• AdaBoost CV:	82.39%	81.82%	82.39%
• Random Forest:	82.23%	81.85%	81.77%
• Random Forest CV:	82.24%	81.85%	82.12%
• KNN: k=20	82.57%	82.57%	82.67%
• KNN CV:	80.20%	79.01%	79.83%
• ANN: 3 Neurons	82.90%	82.84%	83.09%
• SVM with PCA: 20F	82.03%	82.13%	81.85%
• **ANN Grid Search:**	**83.33%**		