Predicting Loan Applicants' Income

Introduction and Problem Identification

- Problem: predicting applicant income from other factors to decide if they’re eligible for a loan
- Income is known to affect likelihood of loan repayment
- Increasing number loans repaid to bank will increase profitability

Data Wrangling

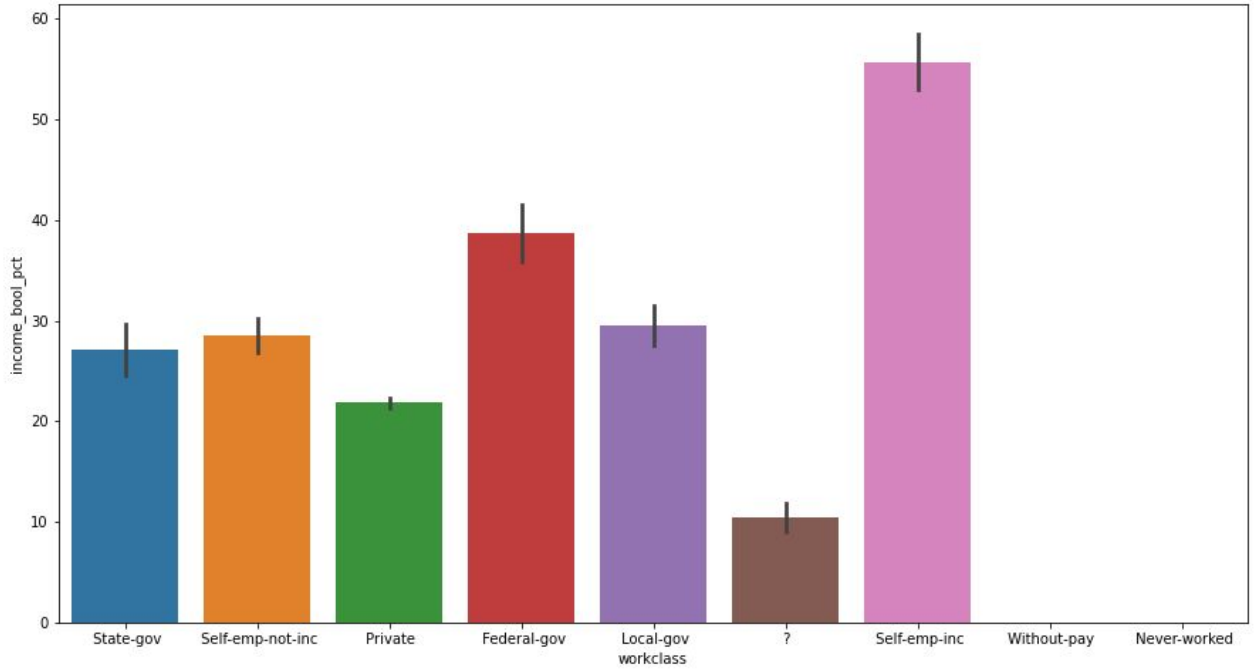
- The chosen census dataset was loaded into a pandas dataframe and jupyter notebook
- No null values were found
- Columns were summarized and inspected

Exploratory Data Analysis (EDA)

- Barplots of ‘income’ vs each explanatory variable were created
- Histograms of each explanatory variable were generated
- A pairplot was made showing relationships between variables

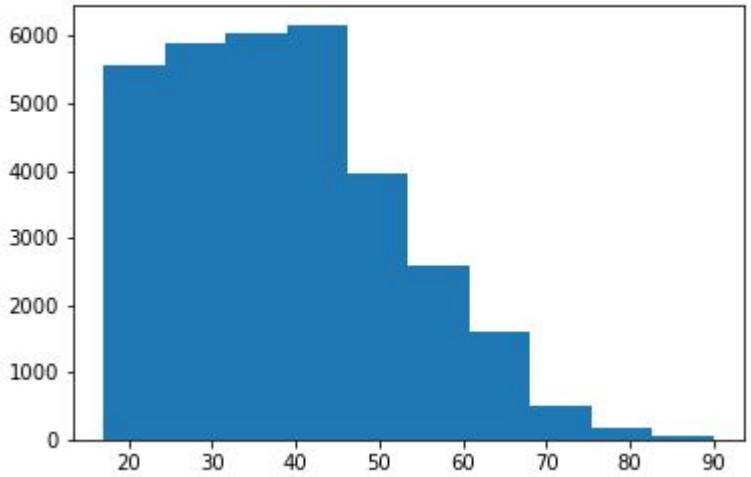
One EDA Barplot for Illustration

- Workclass on x-axis and income on y-axis



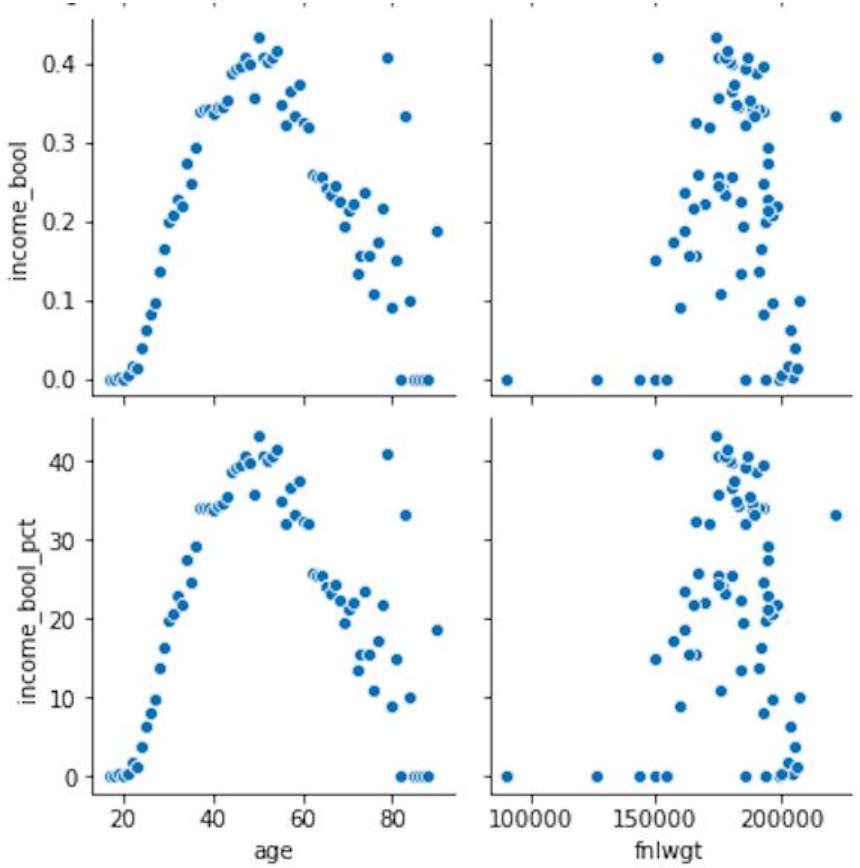
One EDA Histogram for Illustration

- Age (in years) on x-axis and frequency (# of individuals) on y-axis



EDA Pairplot

- Example shows 4 plots out of total of 64 in pairplot
- Variable names can be seen on x and y axes.



Data Preprocessing

- X was created using all columns of pandas dataframe except ‘income’
- Dummy variables were created for categorical variables in X
- y was created using only ‘income’ column
- Train_test_split function was used to split the data into X_train, X_test, y_train and y_test, test size as 0.2
- Finally, a scaler was fit based on X_train and then applied to X_train and X_test.

Modeling

- Four models were created
- Each model's hyperparameters were roughly tuned using GridSearchCV
- In rough order of increasing complexity the models are:
  - Logistic regression
  - Decision tree classifier
  - Random forest classifier
  - Gradient boosting classifier

Model Metrics Table

Model	Accuracy, rounded	Precision (<=50K)	Precision (>50K)	Recall (<=50K)	Recall (>50K)	F1-score (<=50K)	F1-score (>50K)
Logistic Regression	0.805	0.81	0.73	0.97	0.26	0.88	0.39
Decision Tree	0.856	0.88	0.76	0.95	0.56	0.91	0.64
Random Forest	0.859	0.87	0.79	0.96	0.54	0.91	0.64
Gradient Boosting	0.865	0.88	0.78	0.95	0.59	0.91	0.67

Modeling continued

- Logistic Regression performed markedly worse across all 7 metrics in the table than other models
- Other 3 models performed very comparably to each other across all metrics
- Thus, in terms of performance, it’s a toss up
- Decision Tree may be best choice due to Occam’s Razor
- Decision Tree-simpler, easier understand, and less computationally expensive

Conclusion

- Bottom line: a model that performs better than a human (or preexisting methods) at predicting loan repayment could help increase the bank's profitability
- It can improve profitability by preventing the bank from making loans that won't be repaid
- It can also improve profitability by causing the bank to make loans (that it otherwise wouldn't've made) that will be repaid

Recommendations

- If model performs better than preexisting methods, use it to decide whether an applicant is granted a loan or not
- Added bonus that using a model may decrease bank's liability or unconscious discrimination
- Could turn model into online tool that customers can use in advance to see if they will likely be granted a loan
- Could sell this model to other banks/businesses

Future Steps

- Try tuning each model over larger range of hyperparameters
- Perform more cross validation of models with dataset
- Find a larger (but similar) dataset with which to train the models on
- Somehow check veracity of individual’s responses in census dataset
- Try more different models beyond the four tried