US Adult Income

•••

Grannel Pinto
December, 2019

Introduction

The US Adult Census Dataset was retrieved by Barry Becker from the 1994 US Census Database. There are a total of 15 columns in this dataset, 14 of these variables will contribute whether that individual makes an income of ">50K" or "<=50K" in a given year.

United States® Census Bureau

Objective

The objective is to predict the "Income Bracket" by using binary classification on the dataset and evaluate which model has the best accuracy.

Tasks Performed

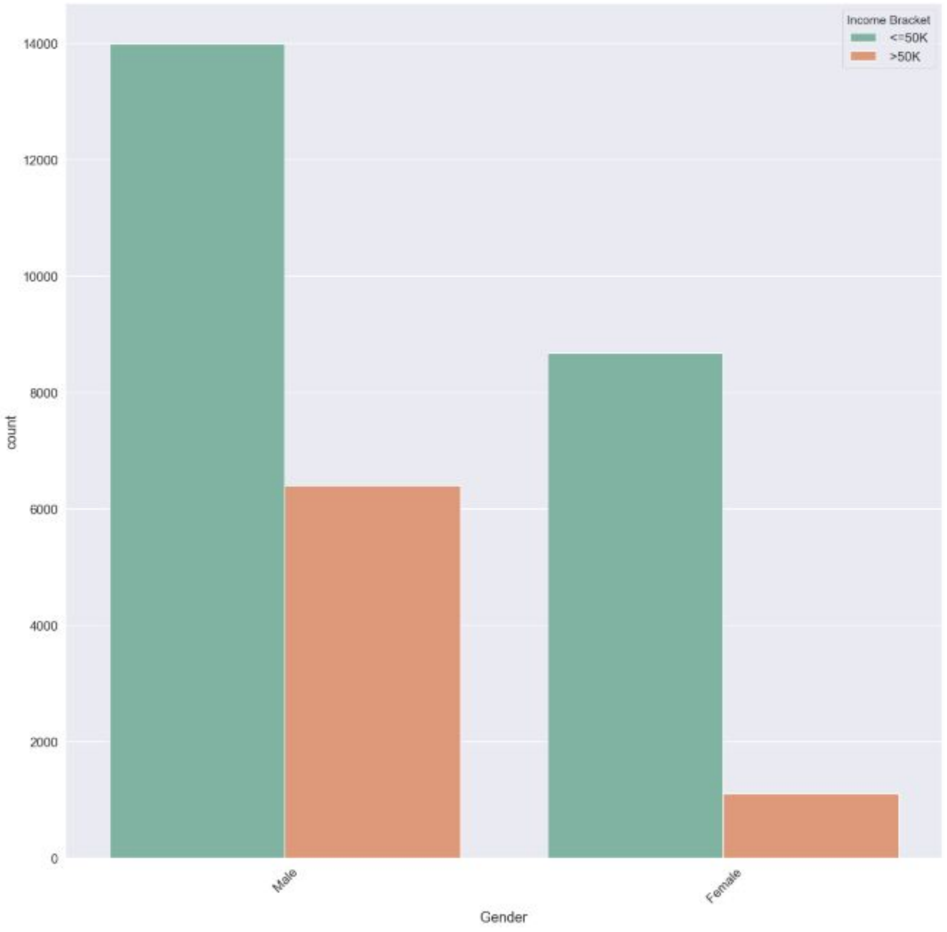
Data analysis & Data Cleaning

- Conceptualize the data
- Replace ‘ ? ’ with ‘nan’ and then dropped from the dataset

Machine Learning & Model Tuning

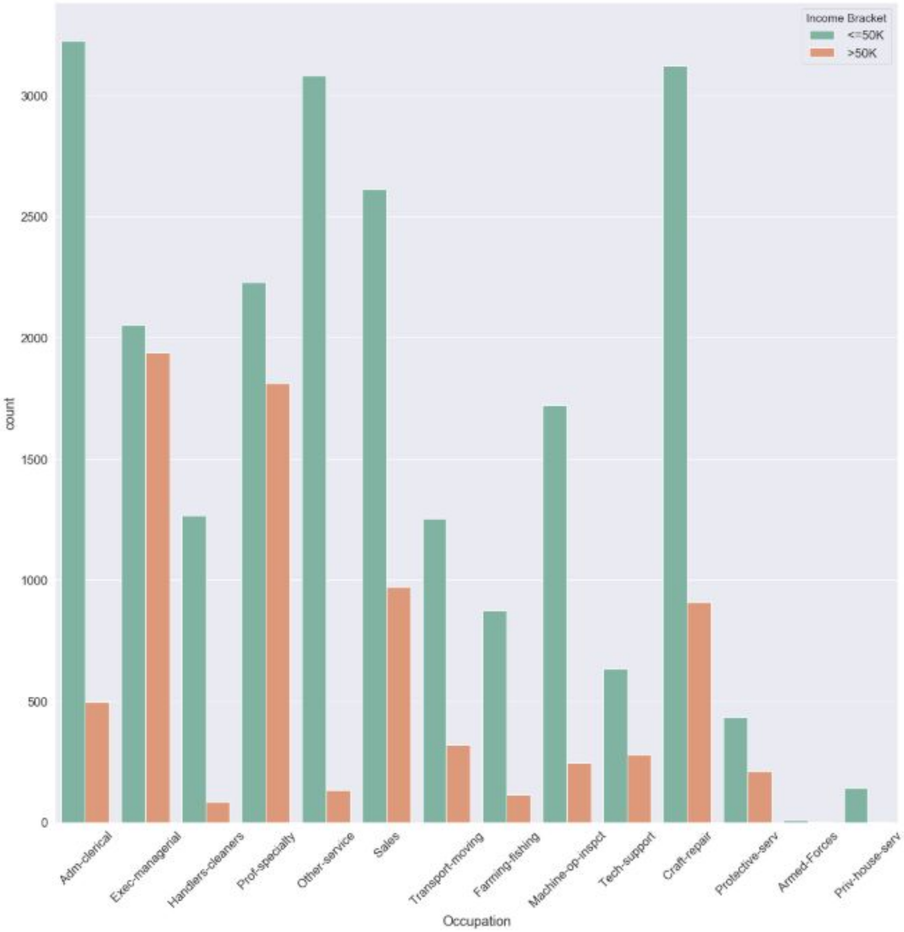
- Decision Tree Model, Logistic Regression Model, Random Forest Classifier Model and SVC Model will be used to determine the training accuracy.
- GridSearchCV will be used to get a higher accuracy.

Data Analysis



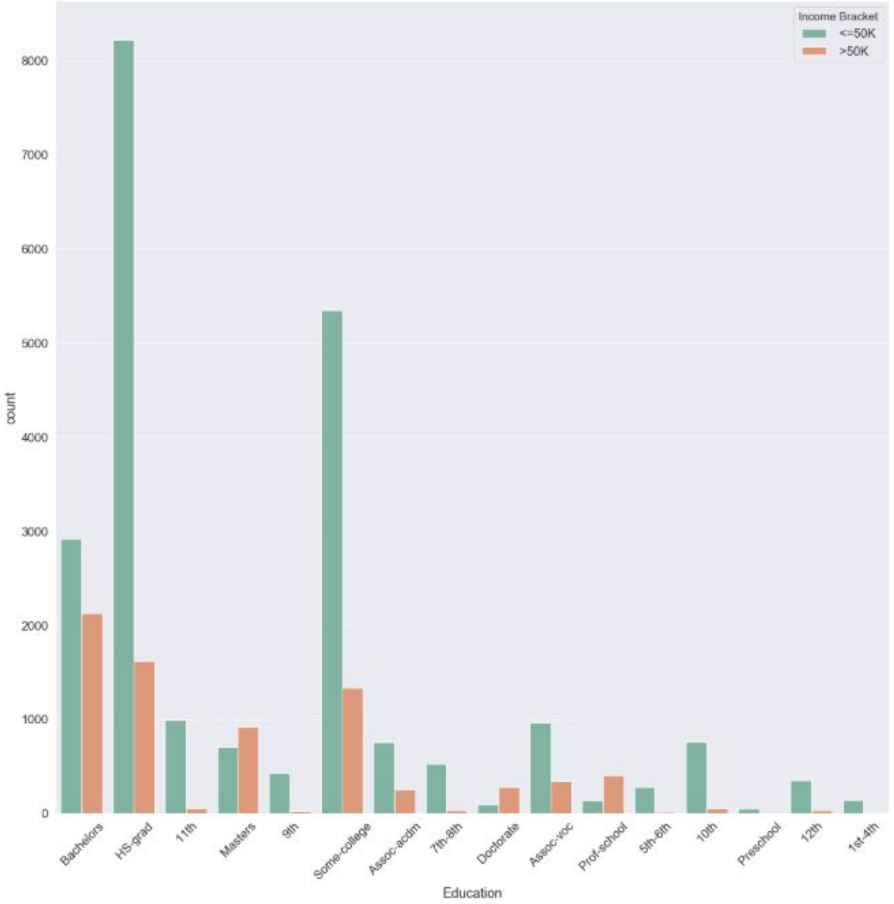
Incomes of <=50K and >50K by Occupation

Handlers-cleaners have the greatest difference of income. Exec-managerial has the least difference as there are far more people in this occupation making >50K.



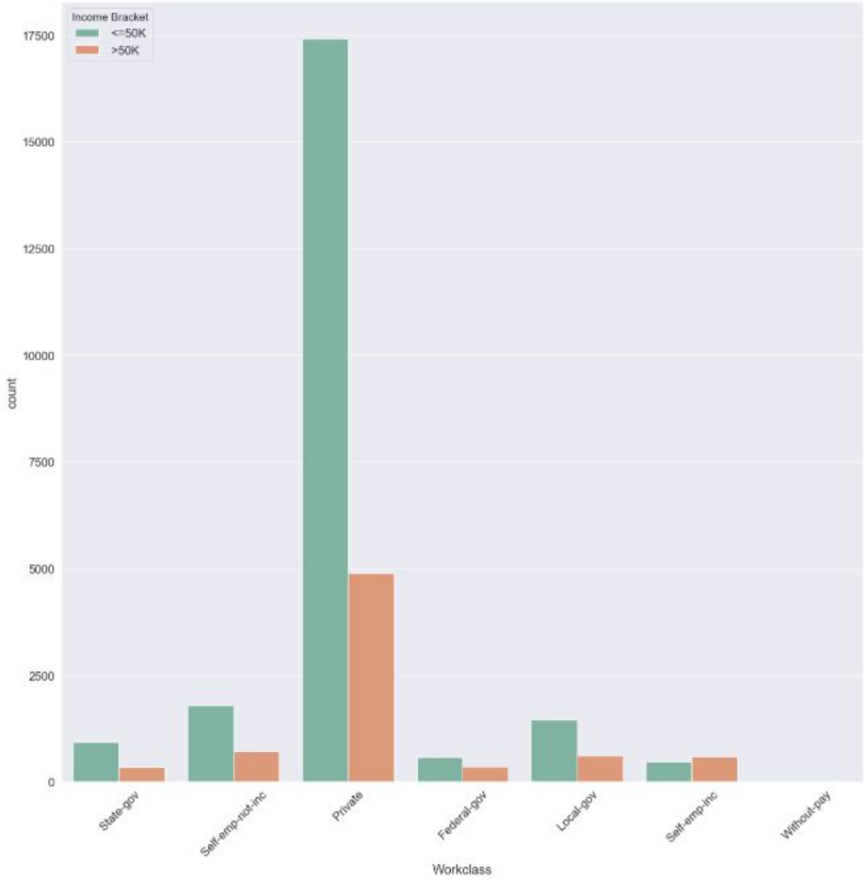
Incomes of <=50K and >50K by Education

Most individuals from this dataset have an education of highschool or more.



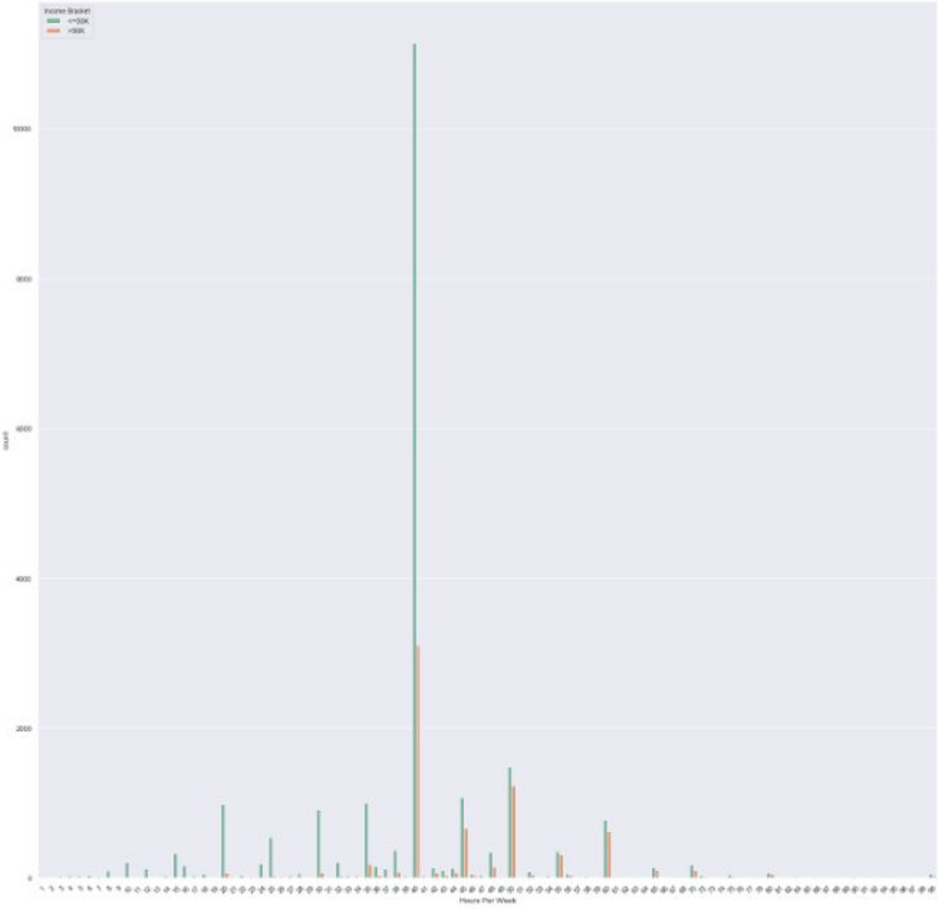
Incomes of <=50K and >50K by Work Class

Self employed workers have more individuals making over 50K. In all the other working classes, there is a huge gap between <=50K and >50K, most people in these sectors make <=50K.



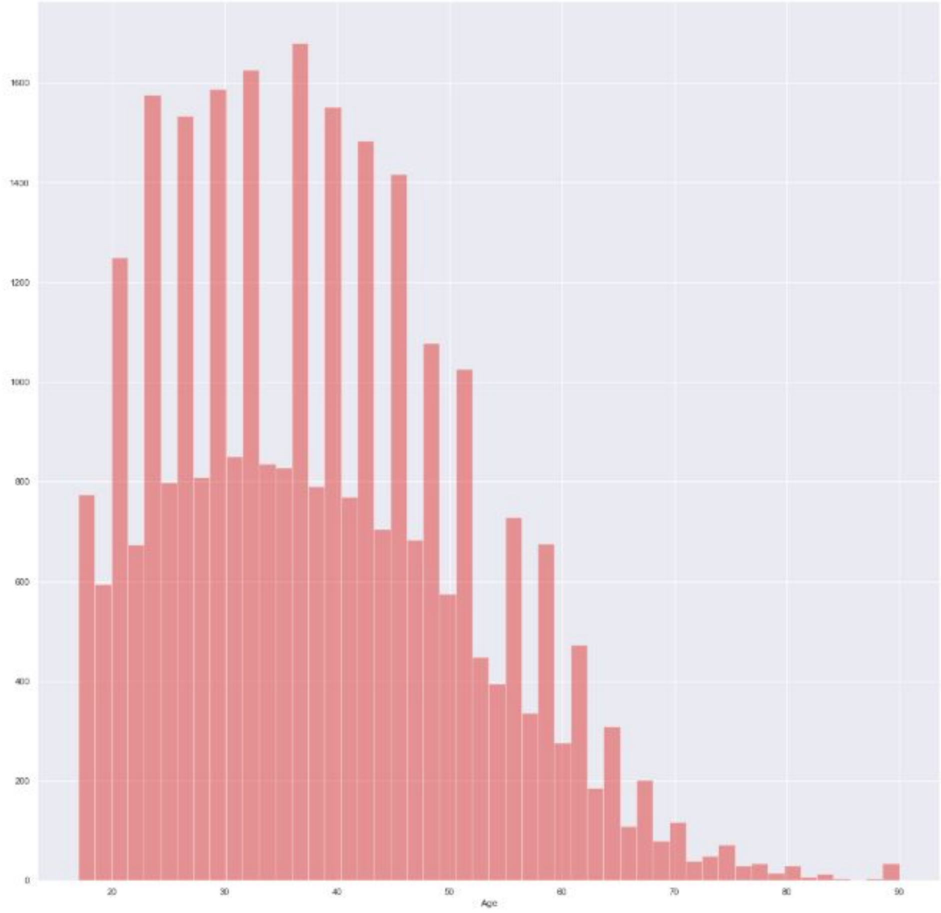
Incomes of <=50K and >50K by Hours Per Week

Most individuals are working 40 hours per week. A large amount of individuals who make >50K seem to work 40 hours or more per week.



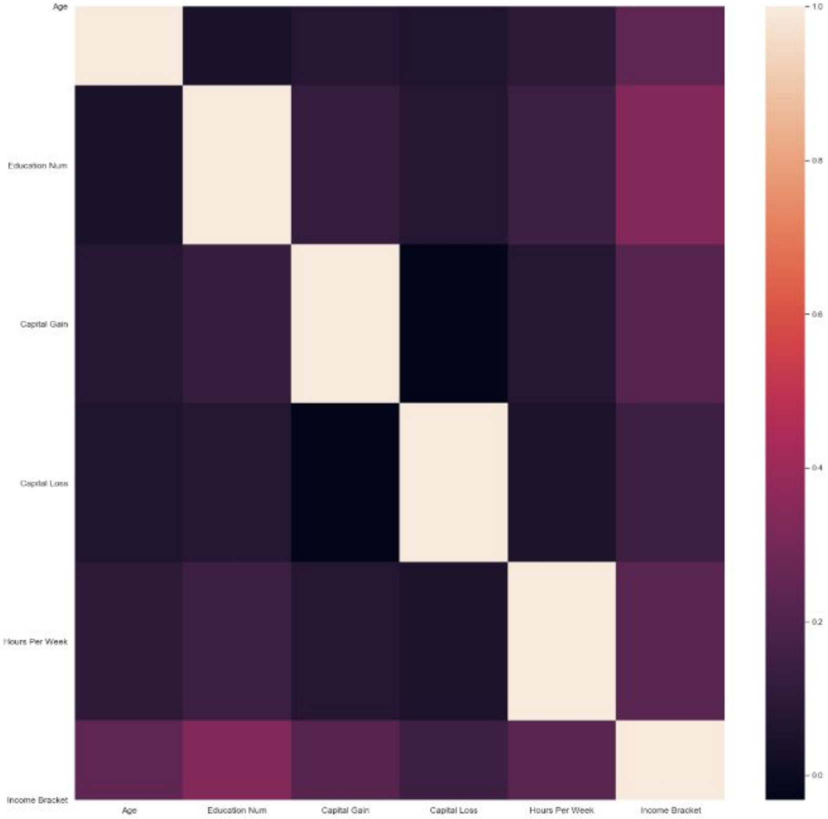
Histogram by Age

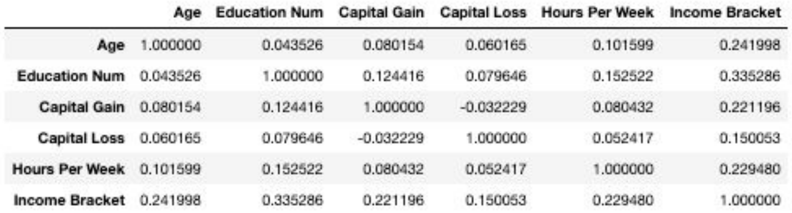
There is a wide age gap in this dataset, from 17 years old to 90 years old. The average age of an individual from the dataset is about 38 years of age.



Correlation Matrix

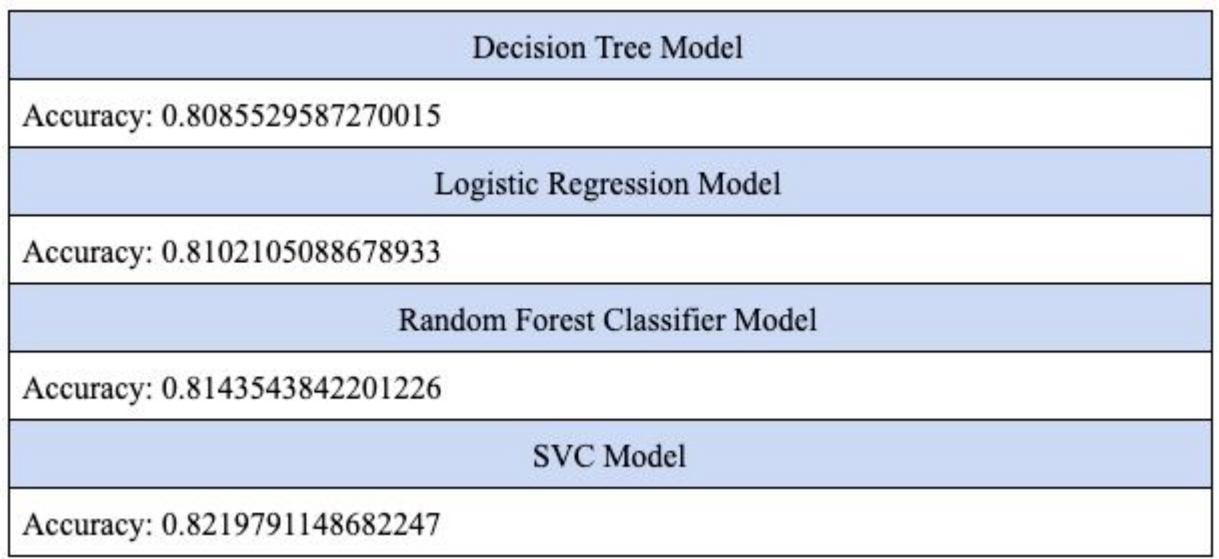
The correlation matrix is geared towards the Income Bracket and its correlation towards the other numerical variables. Education Num has the highest correlation while fnlwgt has the lowest correlation.





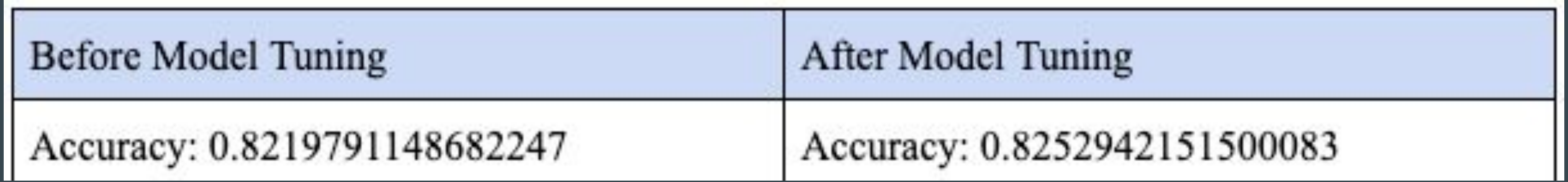
	Age	Education Num	Capital Gain	Capital Loss	Hours Per Week	Income Bracket
Age	1.000000	0.043526	0.080154	0.060165	0.101599	0.241998
Education Num	0.043526	1.000000	0.124416	0.079646	0.152522	0.335286
Capital Gain	0.080154	0.124416	1.000000	-0.032229	0.080432	0.221196
Capital Loss	0.060165	0.079646	-0.032229	1.000000	0.052417	0.150053
Hours Per Week	0.101599	0.152522	0.080432	0.052417	1.000000	0.229480
Income Bracket	0.241998	0.335286	0.221196	0.150053	0.229480	1.000000

Machine Learning



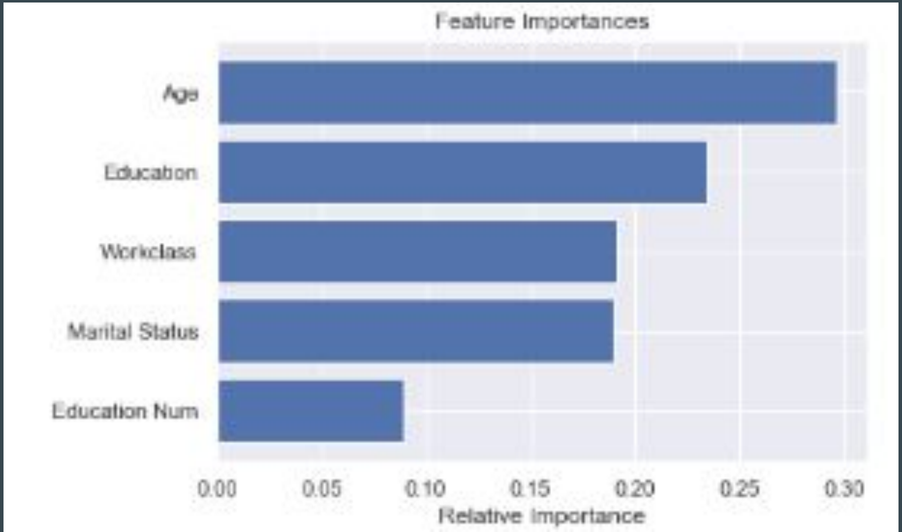
Decision Tree Model
Accuracy: 0.8085529587270015
Logistic Regression Model
Accuracy: 0.8102105088678933
Random Forest Classifier Model
Accuracy: 0.8143543842201226
SVC Model
Accuracy: 0.8219791148682247

Tuning SVC Model



Before Model Tuning	After Model Tuning
Accuracy: 0.8219791148682247	Accuracy: 0.8252942151500083

Feature Importance



Conclusion

- From the feature importances, Age had the most predictive power on the dataset.
- Out of the following classification models used, the SVC Model gave the best accuracy.
- After tuning this model by using GridSearchCV, I received an accuracy of 0.825.

Recommendations & Future Improvements

- Using more models will help to determine the model that gives the best accuracy.
- I also recommend using imblearn for future work, this will help over-sampling data such as the income bracket which has far more individual making <=50K than in comparison to >50K.