Data Mining on Census Income Dataset

Xiaomu Dong, Jincheng Li, Bo Wang
Texas A&M University

Abstract

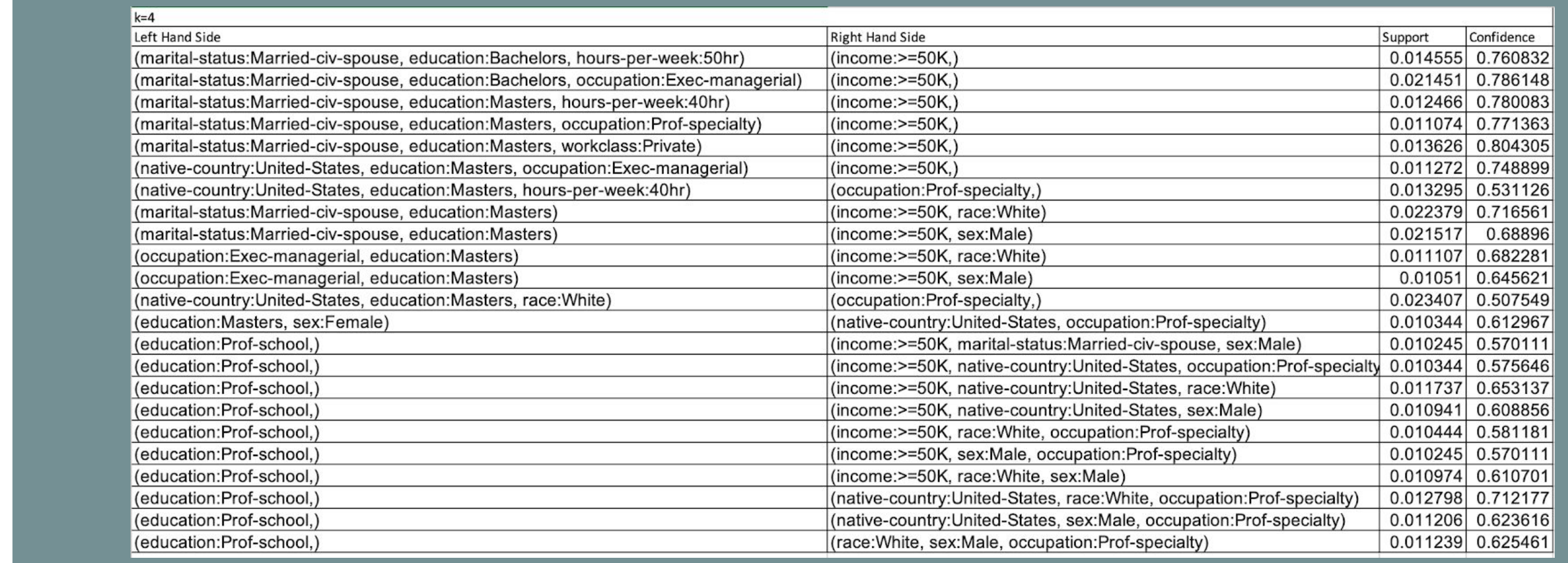
This project uses some advanced data mining techniques to find the important features that affect people's income in US. Many different features such as age, sex, race, education levels are analyzed in this project to find if there are any correlations between these features and the high income. We used Apriori Algorithms to find the association rules between different features and we find some interesting results such as people who are graduated with a higher education level are more likely to be engaged in professional work. Moreover, we used some machine learning algorithms such as logistic regression, random forest to build some binary prediction models which we can use to predict if a person have a high income. Using these models, we can get some important features in these models, which indicates which could affect if a person has a high income. We also applied some mathematical approach to analyze the correlations between different features and high income label.

Methodology

Census income data has always been an important role which measures the economic well-being of the nation. The census tells us a lot about the income distribution among different people group and helps us to determine how to distribute some resources such as schools and assistance. To get a insight from the dataset, we use some data mining techniques here such as Apriori algorithms to find the interesting association rules, we also built some machine learning classification models which can be used to predict if a person has a higher income. We also applied some other techniques to find the important features which may contribute to a higher income. Our results also indicates the income inequality between different sex, race and other factors.

Results

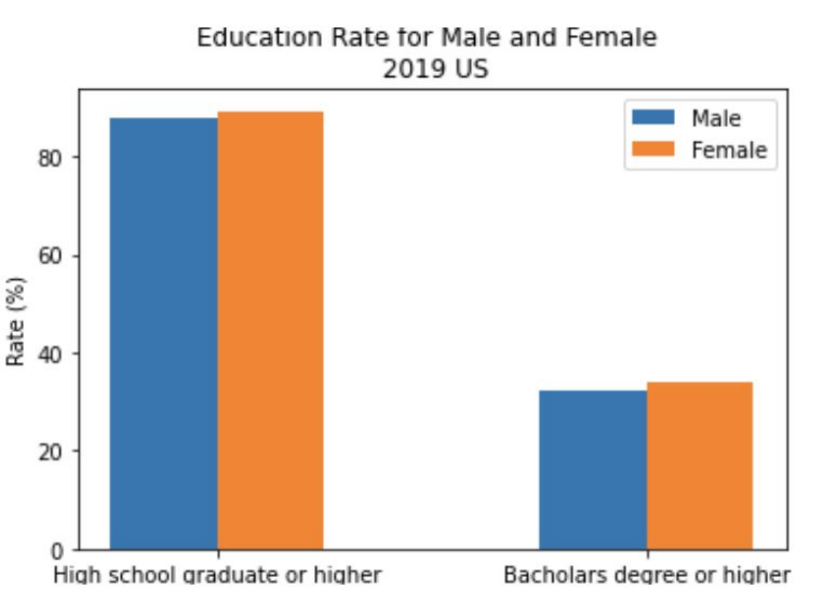
Interesting association rules found in census data:

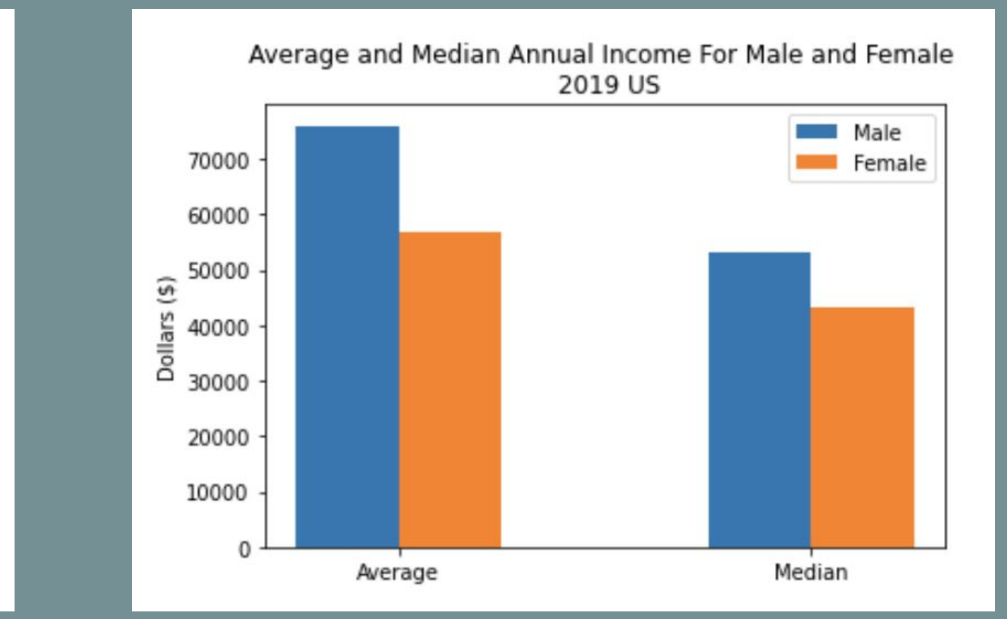


k=4			
Left Hand Side	Right Hand Side	Support	Confidence
(marital-status:Married-civ-spouse, education:Bachelors, hours-per-week:50hr)	(income:>=50K,)	0.014555	0.760832
(marital-status:Married-civ-spouse, education:Bachelors, occupation:Exec-managerial)	(income:>=50K,)	0.021451	0.786148
(marital-status:Married-civ-spouse, education:Masters, hours-per-week:40hr)	(income:>=50K,)	0.012466	0.780083
(marital-status:Married-civ-spouse, education:Masters, occupation:Prof-specialty)	(income:>=50K,)	0.011074	0.771363
(marital-status:Married-civ-spouse, education:Masters, workclass:Private)	(income:>=50K,)	0.013626	0.804305
(native-country:United-States, education:Masters, occupation:Exec-managerial)	(income:>=50K,)	0.011272	0.748899
(native-country:United-States, education:Masters, hours-per-week:40hr)	(occupation:Prof-specialty,)	0.013295	0.531126
(marital-status:Married-civ-spouse, education:Masters)	(income:>=50K, race:White)	0.022379	0.716561
(marital-status:Married-civ-spouse, education:Masters)	(income:>=50K, sex:Male)	0.021517	0.68896
(occupation:Exec-managerial, education:Masters)	(income:>=50K, race:White)	0.011107	0.682281
(occupation:Exec-managerial, education:Masters)	(income:>=50K, sex:Male)	0.01051	0.645621
(native-country:United-States, education:Masters, race:White)	(occupation:Prof-specialty,)	0.023407	0.507549
(education:Masters, sex:Female)	(native-country:United-States, occupation:Prof-specialty)	0.010344	0.612967
(education:Prof-school,)	(income:>=50K, marital-status:Married-civ-spouse, sex:Male)	0.010245	0.570111
(education:Prof-school,)	(income:>=50K, native-country:United-States, occupation:Prof-specialty	0.010344	0.575646
(education:Prof-school,)	(income:>=50K, native-country:United-States, race:White)	0.011737	0.653137
(education:Prof-school,)	(income:>=50K, native-country:United-States, sex:Male)	0.010941	0.608856
(education:Prof-school,)	(income:>=50K, race:White, occupation:Prof-specialty)	0.010444	0.581181
(education:Prof-school,)	(income:>=50K, sex:Male, occupation:Prof-specialty)	0.010245	0.570111
(education:Prof-school,)	(income:>=50K, race:White, sex:Male)	0.010974	0.610701
(education:Prof-school,)	(native-country:United-States, race:White, occupation:Prof-specialty)	0.012798	0.712177
(education:Prof-school,)	(native-country:United-States, sex:Male, occupation:Prof-specialty)	0.011206	0.623616
(education:Prof-school,)	(race:White, sex:Male, occupation:Prof-specialty)	0.011239	0.625461

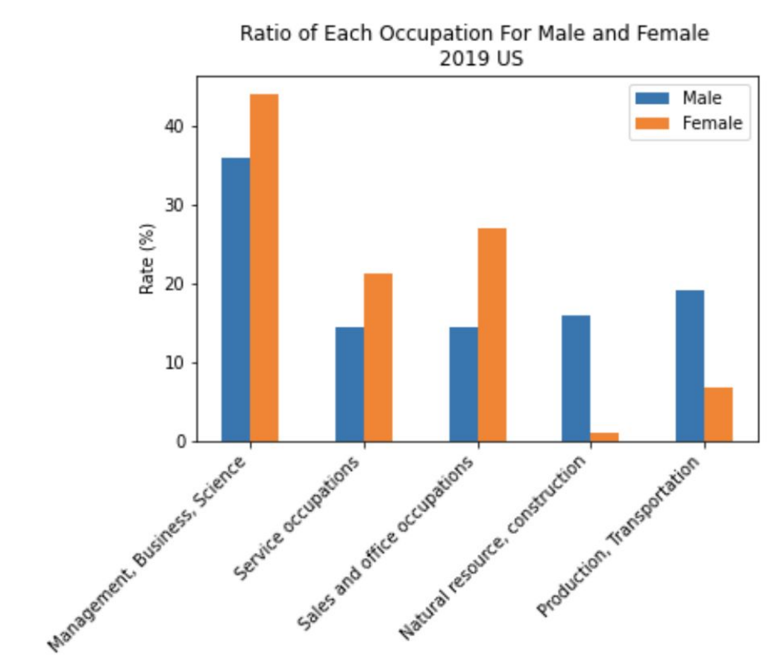
One thing to notice is that for a random person who went to professional school, the possibility of that person to be a man is much higher than 0.5. This means that there exists an obvious inequality in education chance for male and female.

Take a look at the education rate data shown in the figure below (LHS), the number of people who is high school graduate or higher is generally same for male and female. Combined with what we found from the association rules, we can get that more women are graduated from high school or college, but more men got high level education (master, phd). When it comes to annual income, the equality between male and female also exists. From the figure on the right, we can see that both the average and median annual income of male are obviously higher than that of female.

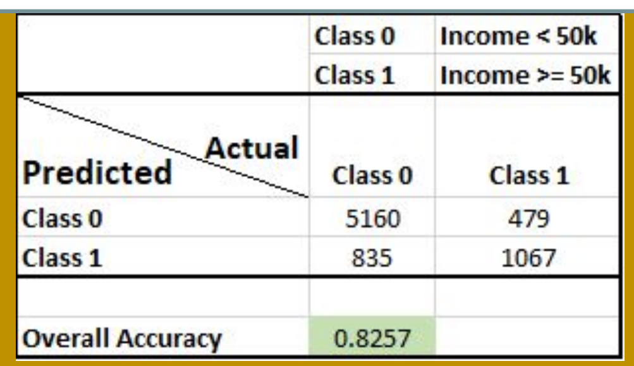




To analyze further in the occupation data, in occupations usually with higher salary like management, business, and science, the ratio of female employees is significantly higher than the ratio of male. However, from what we see from the previous figure, average income of female is much lower than that of male. Higher percentage of women have better jobs but this results in lower average annual income. Inferred from this, we can say that women are still unequally treated in 2019.

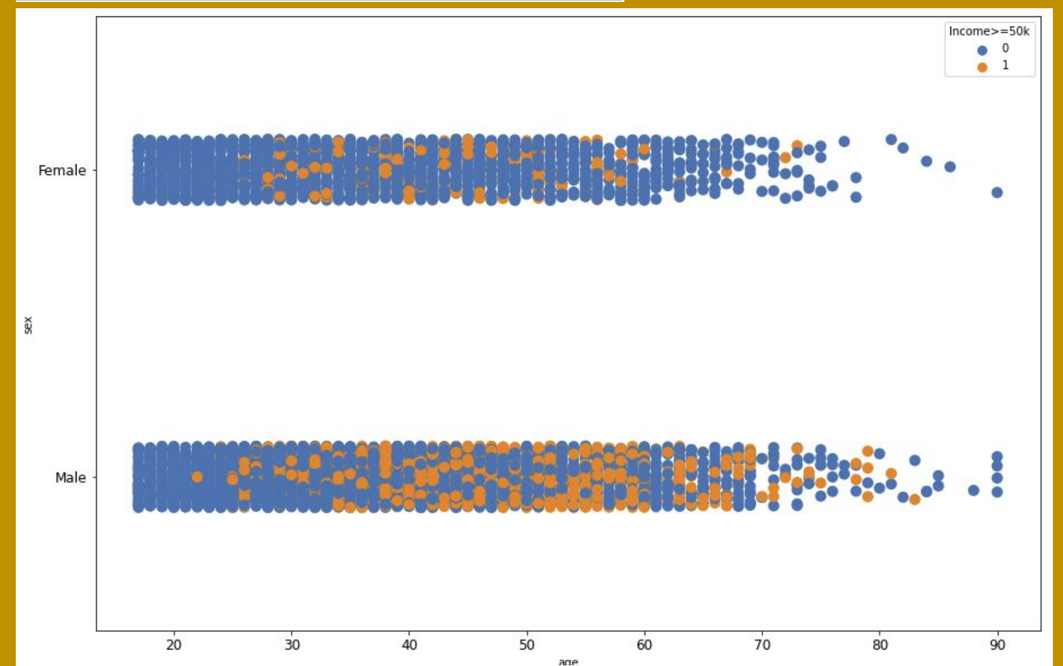


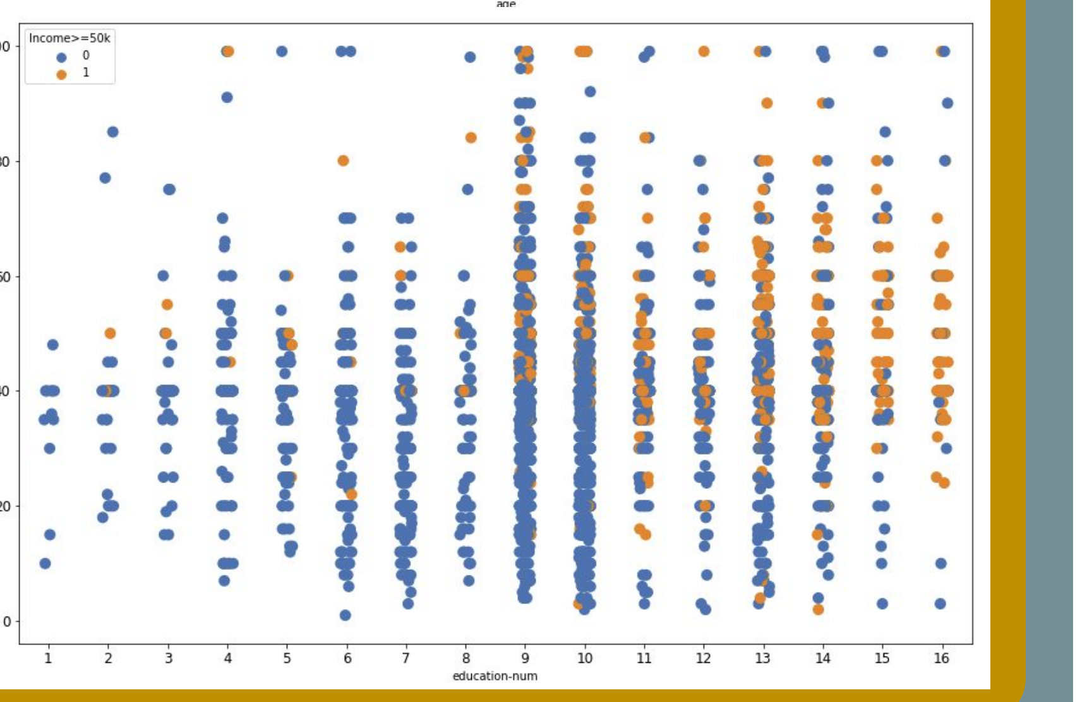
The overall logistic regression model has an accuracy of 0.8257 based on 75% of the data trained.



	Class 0	Income < 50k
	Class 1	Income >= 50k
Actual Predicted	Class 0	Class 1
Class 0	5160	479
Class 1	835	1067
Overall Accuracy	0.8257	

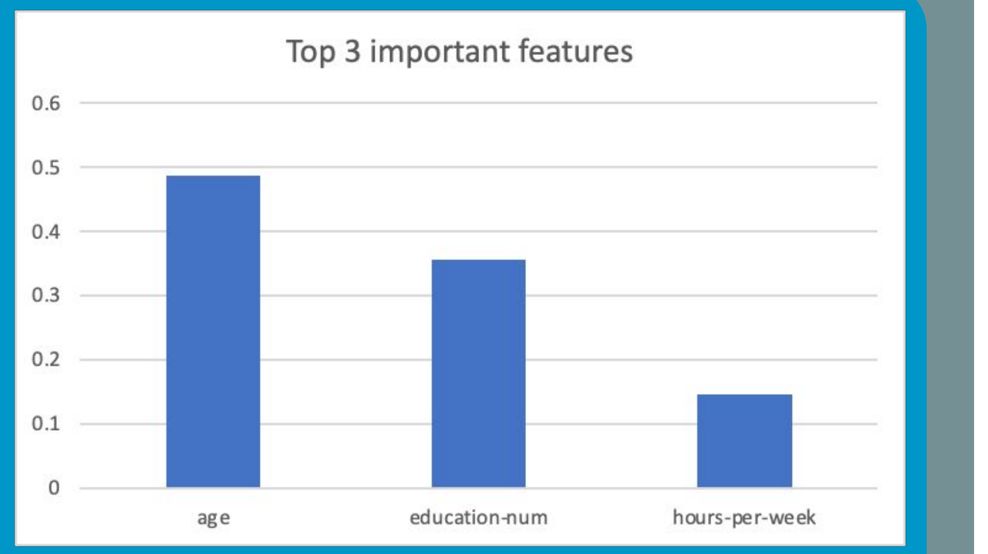
Since the logistic regression is trained on dozens of features, it is not possible to display all the features and labels in a 2D plot. However, from some features combination, we can observe that there is a distinct trend/ cluster that separates low income from high income. For example, gender and age plot indicates a clear gender inequality that more male had income over 50k than female. The hours-per-week and education year showed that most high income people are the ones who have over 10+ years of education since middle school, i.e. Bachelor's degree or higher. This indicates that education is one of the most important factor that contributes to a high income.

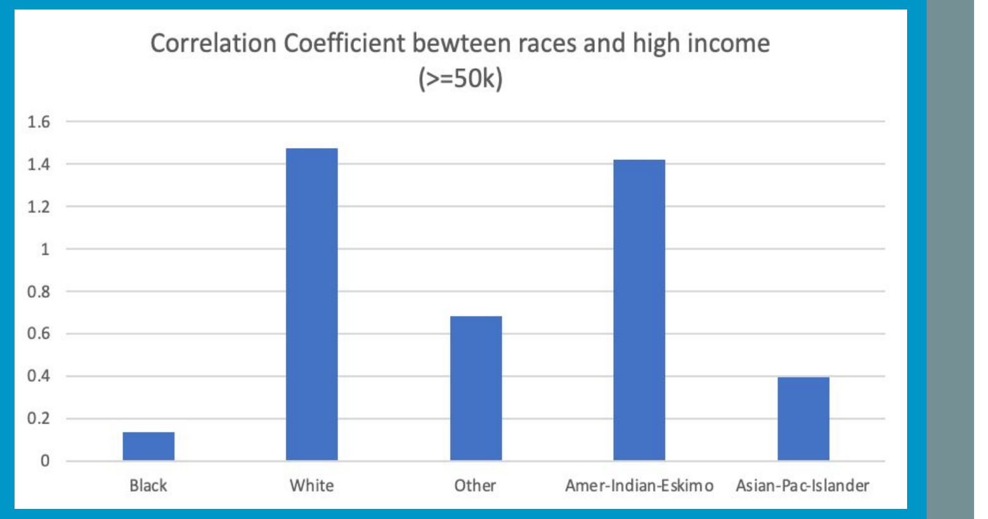


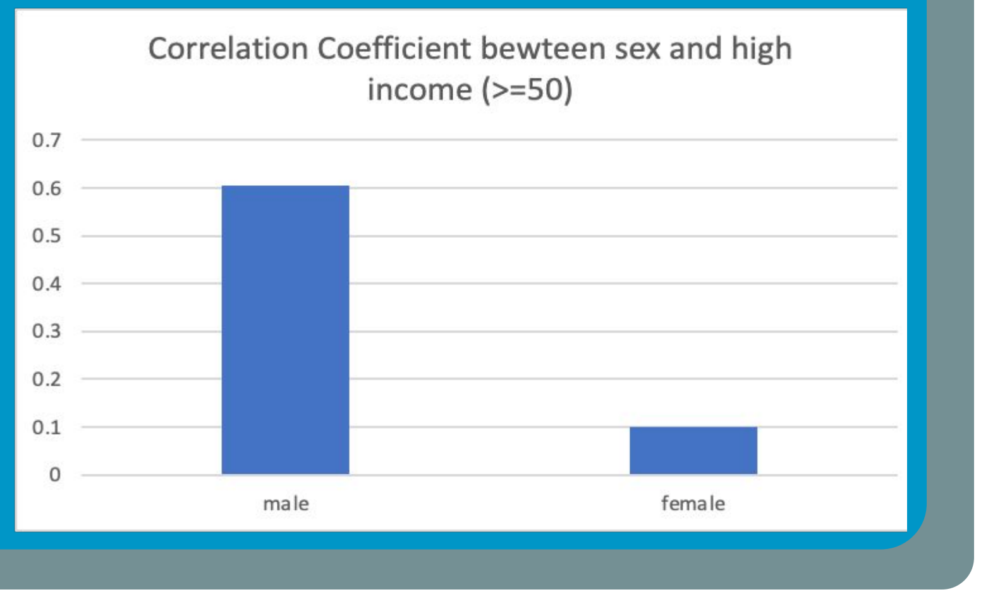


We used different techniques to find the important features that may contribute to a high income. Firstly, we built a random forest binary classifier which achieved a prediction accuracy score of 0.84. In this model, we can get the all the feature importance score, in which we pick out top 3 important features that may lead to a high income. These top 3 important features are age, education levels and working hours per week, which is shown on the right.

We also applied the the chi-square test to find the correlation coefficient between some noticeable features such as race and the high income. The result is shown on the right. From which we can find that some races have a high correlation with the high income. We also analyzed the correlation between sex and high income, we find that males are more likely to be related to high income.







Conclusions

- There is still inequality in multiple aspects. Gender inequality exists in the chance of getting higher level education (master, phd), but for high school and college, gender inequality is not obvious. Also, the gender inequality exists in income. Even when female get jobs that are commonly thought to be highly paid, they may be paid less compared with male who take the same kind of jobs. Racial inequality also exists with regards to income. In our data, white people still have the highest income.

- Usually, higher level education is related to higher income, but there is no clear evidence in our data that shows people with a doctorate degree are the most possible to get high income (>=50K).

- The most important features that decide if a person has high income are age, education and working time (hours per week).

References

https://data.census.gov/cedsci/table?q=&t=Education%3AIncome%20and%20Poverty&tid=ACSSPP1Y2019.S0201

https://archive.ics.uci.edu/ml/datasets/Census+Income

https://scikit-learn.org/stable/modules/classes.html

https://pytorch.org/docs/stable/index.html