B565 Final Project

Predicting Income from Census Data

Karthik Vegi
kvegi@iu.edu

Melita Dsouza
dsouzam@iu.edu

1. Goal of the Project

The goal of the project is to create a classification model to identify the individuals with a potential of earning more than $50,000 USD. The response variable is a binary classifier that will identify whether a person will make $50k and the predictor variables are census information like age, marital status, education etc. We would like to identify the significance of the variables that were used as predictors. A couple of classification techniques were used to compare the accuracy of the classification. The focus of this project will be more on the interpretation and less on implementing a classification algorithm from scratch.

2. Statistics of the dataset used

The dataset used is from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Adult
No of Observations: 48842

No of variables: 14
Attribute type: categorical/numeric

Predictor variables: age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, weekly-hours, native-country, income

Response variable: Income

3. Tools Used

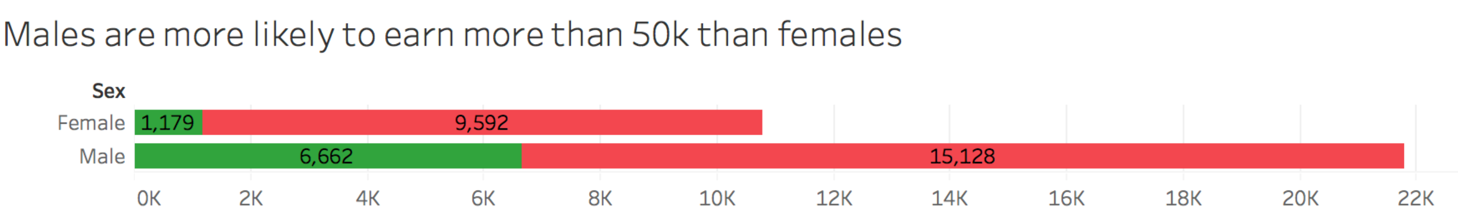
Analysis: R
Visualization: Tableau, R

4. Feature Selection

We will exclude the irrelevant variables that will not go into the construction of our classification model. The variables **fnlwgt** and **education-num** are assumed to be of no or less significance and are removed from the training and test datasets. Attributes like **education**, **workclass** and **relationship** are strong features that will affect the response variable.

5. Data Visualization

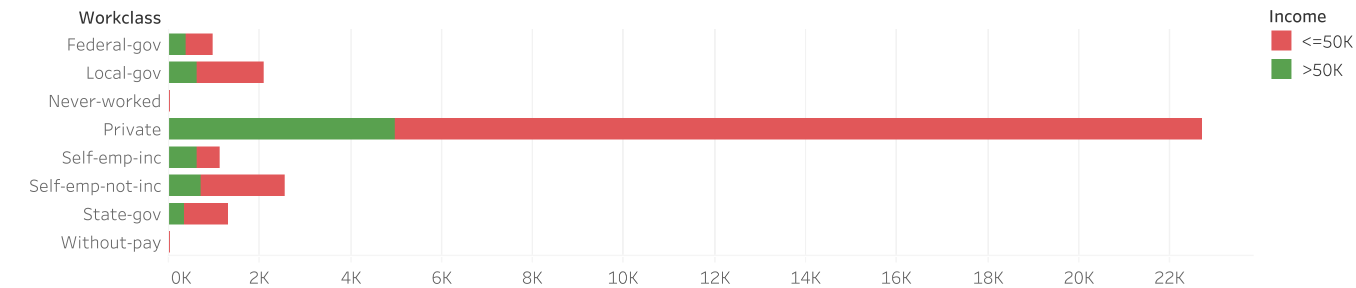
We visualized the data in tableau to understand the significance of some important attributes



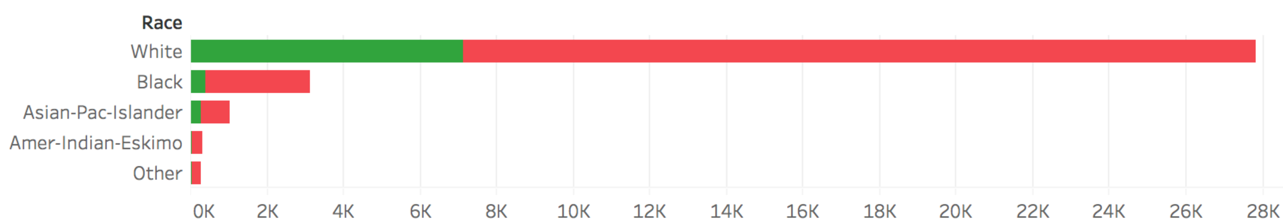
Divorced, Seperated and Widowed are more likely to earn less than 50k

	Marital-Status						
Income	Married-civ-spouse	Never-married	Divorced	Separated	Widowed	Married-spouse-absent	Married-AF-spouse
<=50K	8,284	10,192	3,980	959	908	384	13
>50K	6,692	491	463	66	85	34	10

Income Vs Work class



White Population tend to earn more than 50k



6. Choice of classification technique

- We chose to implement two classification algorithms so that we can compare the performance
- **Naïve bayes** is simple and works well even with a small training set
- **Decision Tree** is fast and scalable to compute, especially in this case where there are a lot of records

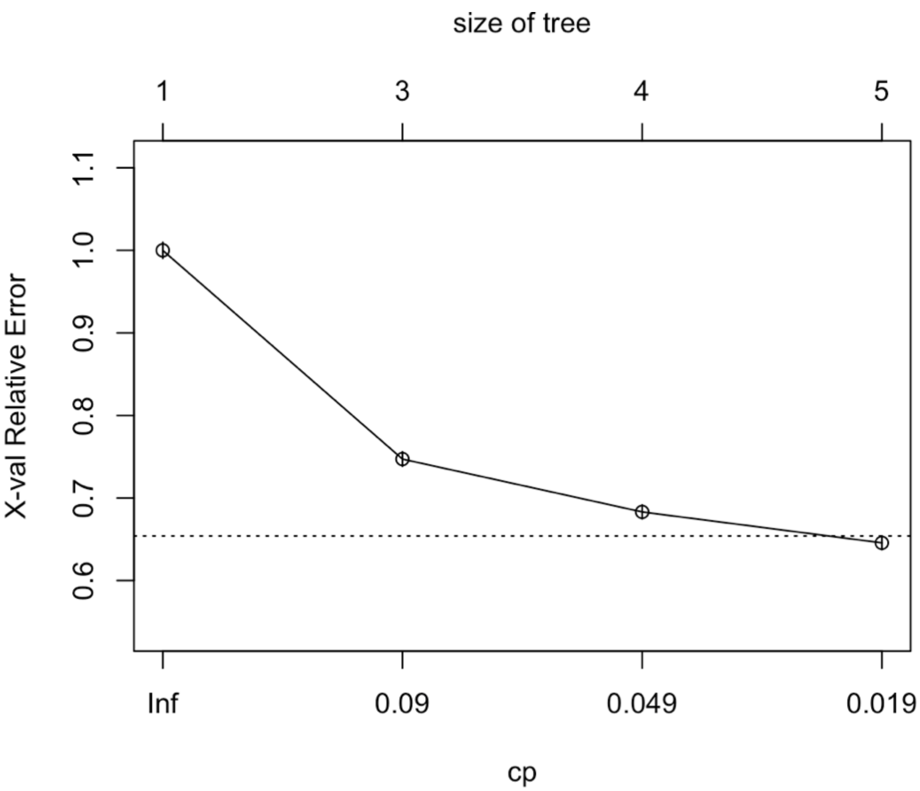
(i) Naïve Bayes classifier: Below are the test results for naïve bayes classifier

```
Naive Bayes: Confusion matrix for training set..
train.pred  <=50K  >50K
     <=50K  23273  4445
     >50K    1447  3396

Naive Bayes: Accuracy of classifier on the training set is..[1] 81.90473

Naive Bayes: Confusion matrix for test set..
naive.pred  <=50K  >50K
     <=50K  11846  2522
     >50K     589  1324

Naive Bayes: Accuracy of classifier on the test set is..[1] 80.89184
```

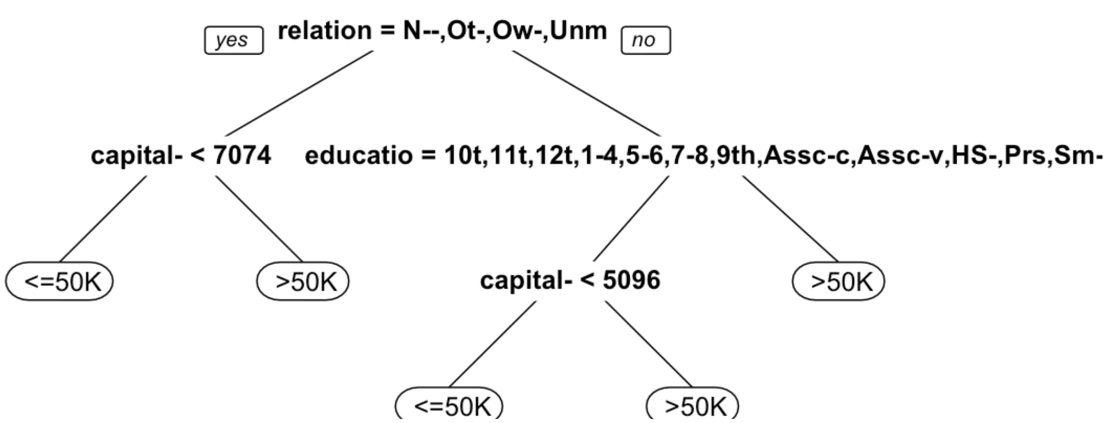
(ii) Decision Tree classifier:

```
Variables actually used in tree construction:
[1] capital-gain education    relationship

Root node error: 7841/32561 = 0.24081

n= 32561

        CP nsplit rel error  xerror      xstd
1 0.126387      0   1.00000 1.00000 0.0098399
2 0.064022      2   0.74723 0.74723 0.0088402
3 0.037495      3   0.68320 0.68320 0.0085321
4 0.010000      4   0.64571 0.64571 0.0083394
```

Error Rate Plot:



Decision Tree:

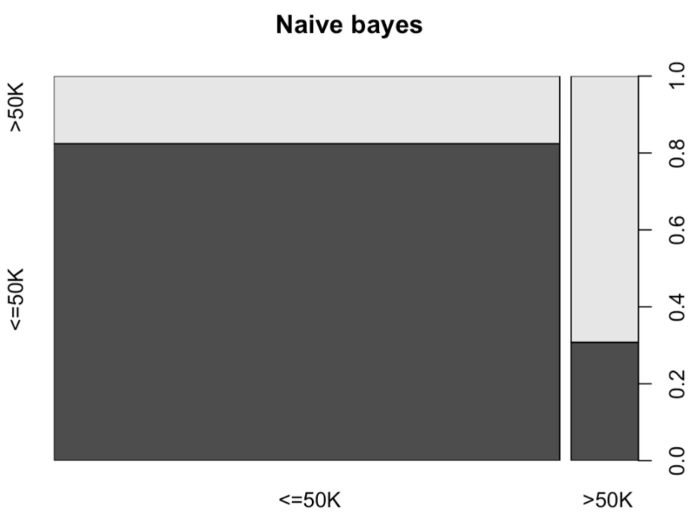


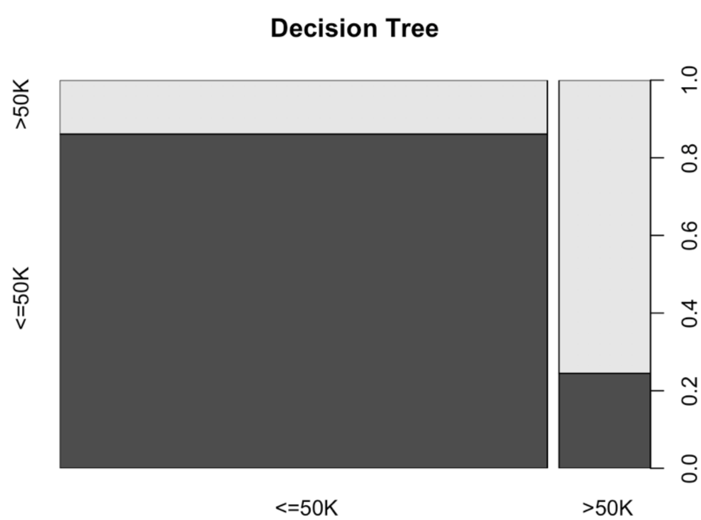
Decision Tree Statistics:

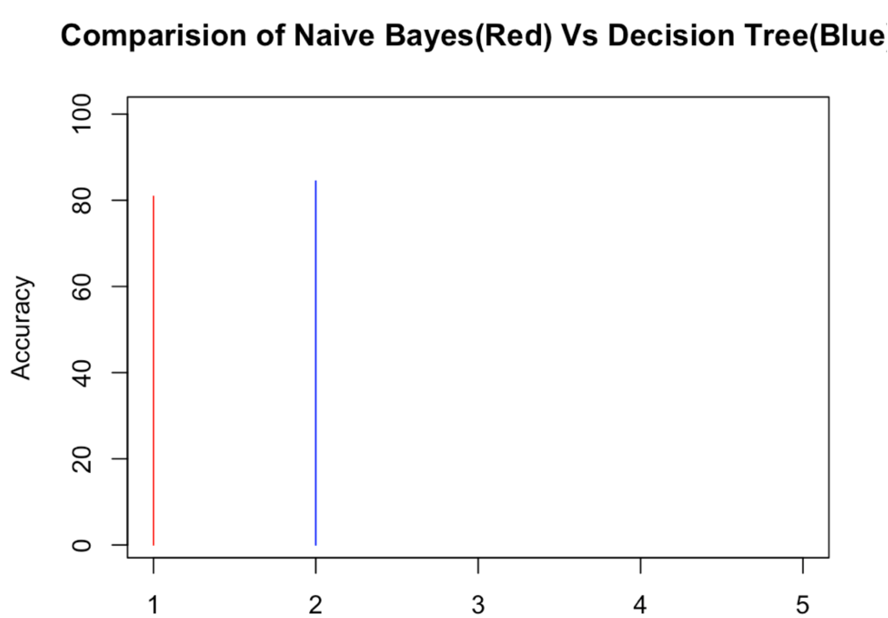
```
Decision Tree: Confusion matrix for training set..
train.pred  <=50K  >50K
     <=50K  23473  3816
     >50K    1247  4025

Decision Tree: Accuracy of classifier on the training set is..[1] 84.45072

Decision Tree: Confusion matrix for test set..
tree.pred  <=50K  >50K
    <=50K  11805  1901
    >50K     630  1945

Decision Tree: Accuracy of classifier on the test set is..[1] 84.45427
```

7. Evaluating Performance

With the dataset, Decision tree performed better and the graphs are summarized below:



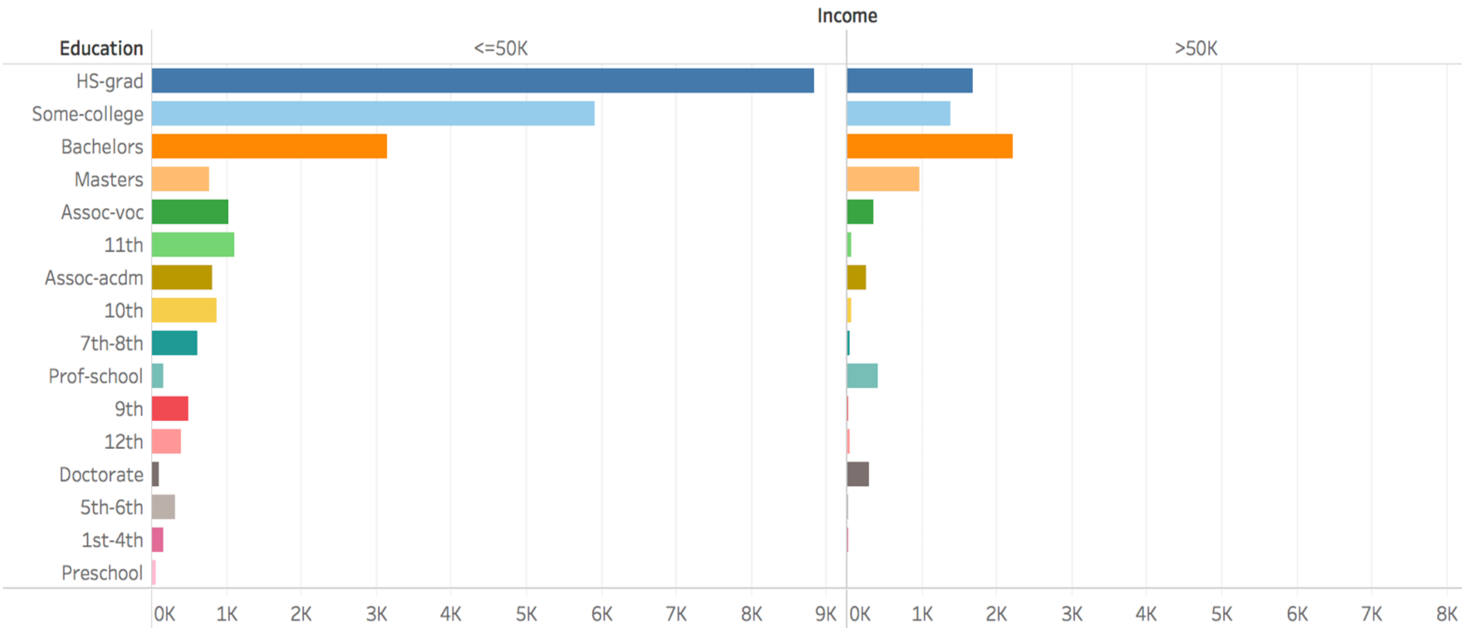




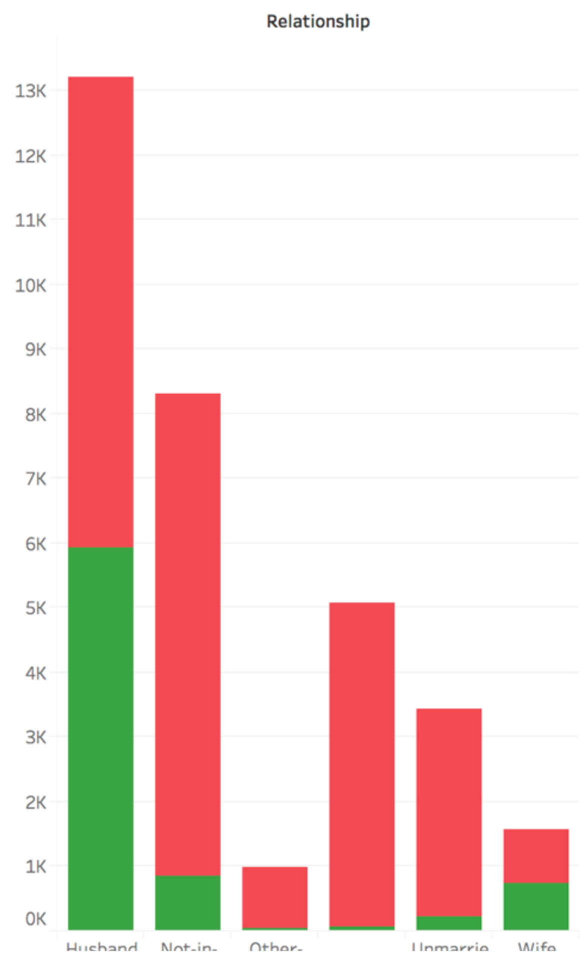
8. Key Insights

- **People with higher degrees tend to earn more than 50k**

As the level of education increases, number of people earning > 50k increases



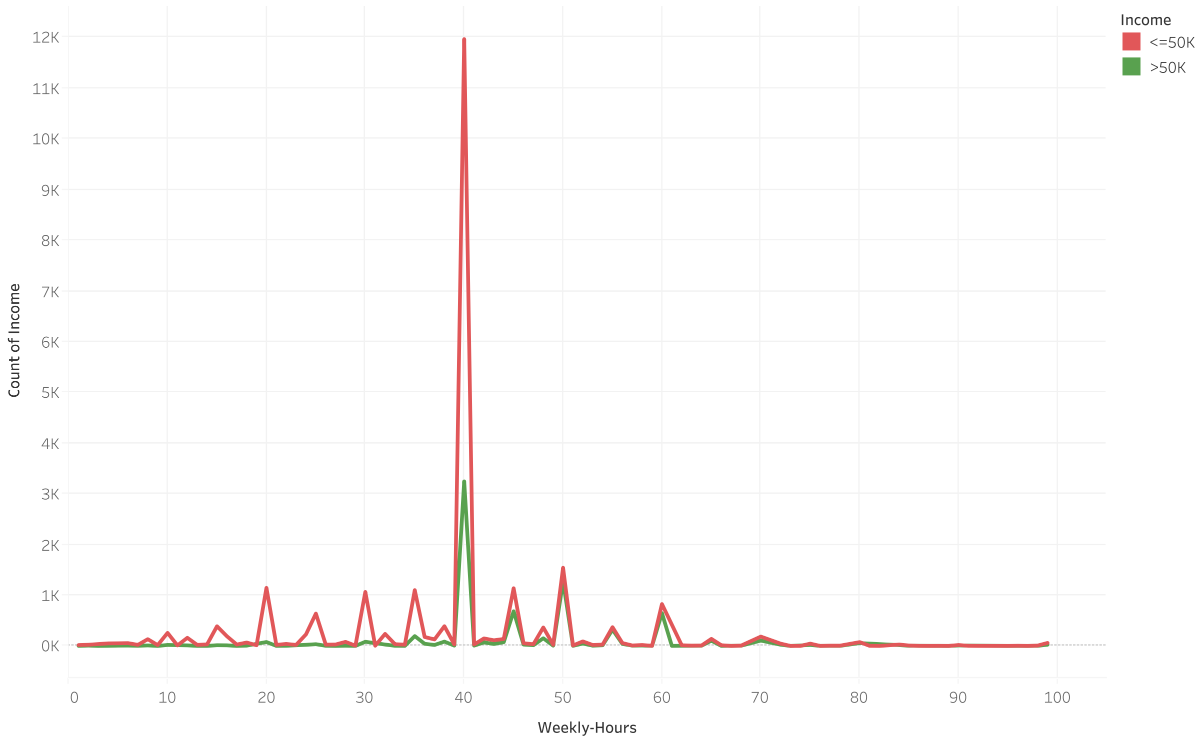
- **People in a stable relationship usually earned more than 50k**



- **Working class has 6% of the records missing and the bulk of the population fall into private sector**

workclass	type:integer class:factor	1. ?	1: 1836 (5.6%)	32561/32561 (100.0%)
		2. Federal-gov	2: 960 (2.9%)	
		3. Local-gov	3: 2093 (6.4%)	
		4. Never-worked	4: 7 (0%)	
		5. Private	5: 22696 (69.7%)	
		6. Self-emp-inc	6: 1116 (3.4%)	
		7. Self-emp-not-inc	7: 2541 (7.8%)	
		8. State-gov	8: 1298 (4%)	
		9. Without-pay	9: 14 (0%)	

- **People who put in close to 40 hours earned more than 50k**



9. Challenges and Opportunities

- With the given time, implementing a classification algorithm from scratch was not feasible
- With more time, a good work will be to compare all classification algorithms and plot the accuracy of them

10. References

[1] https://www.tableau.com/learn/training
[2] https://archive.ics.uci.edu/ml/datasets/Adult
[3] https://www.r-bloggers.com/classification-trees-using-the-rpart-function/