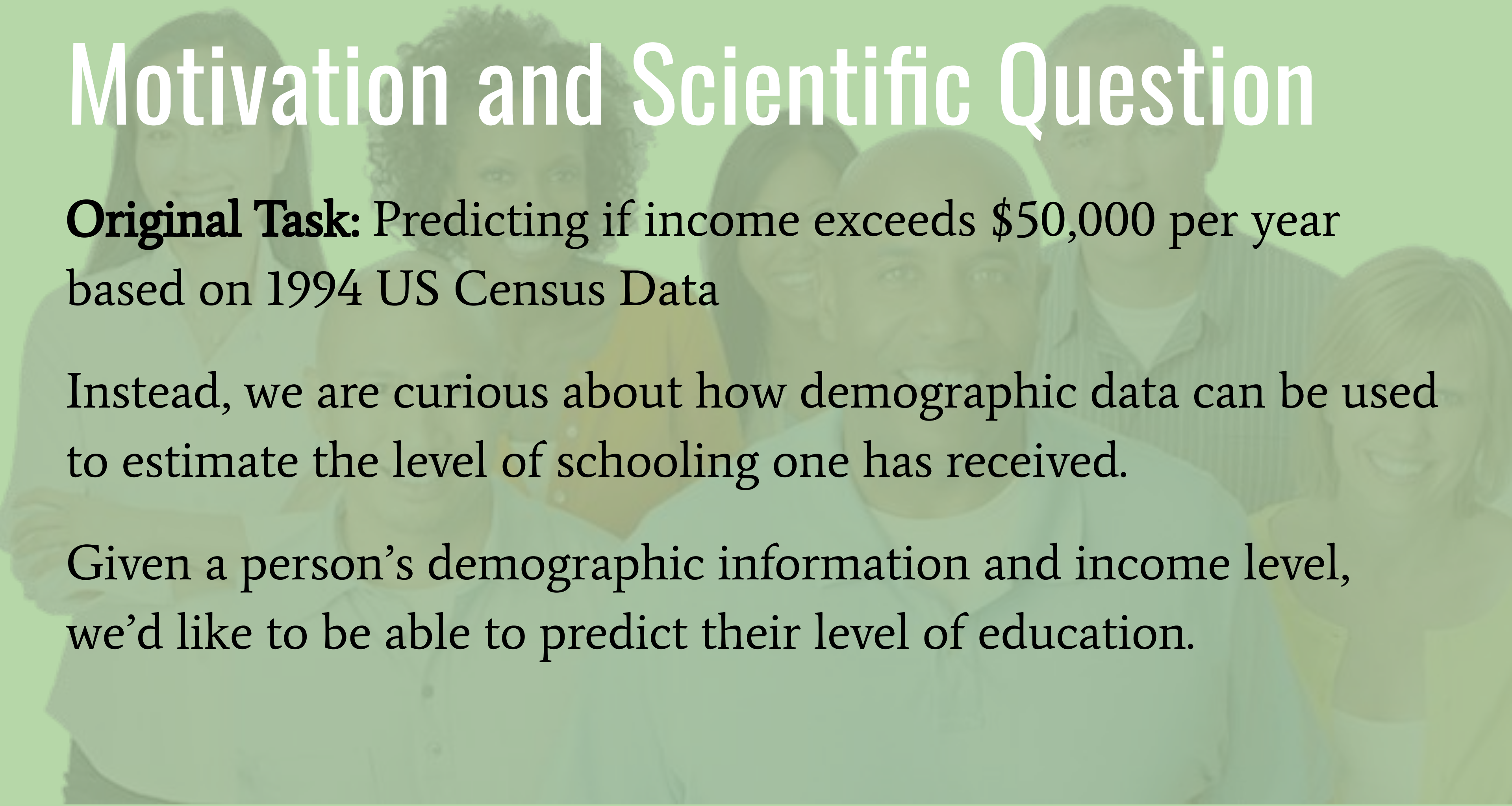
Census Data to Predict Levels of Education

•••

May 16th, 2019

Kenny Gwon, Raymond Liu, Richard Muniu

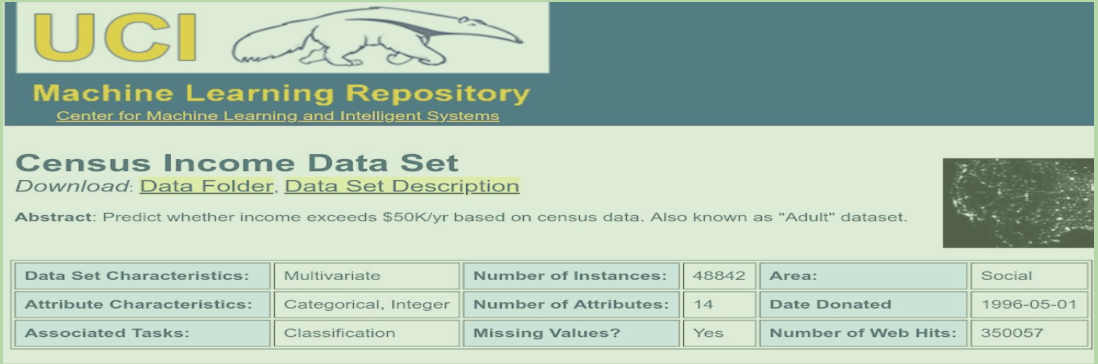
Motivation and Scientific Question



Original Task: Predicting if income exceeds $50,000 per year based on 1994 US Census Data

Instead, we are curious about how demographic data can be used to estimate the level of schooling one has received.

Given a person’s demographic information and income level, we’d like to be able to predict their level of education.


UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Census Income Data Set

Download: Data Folder, Data Set Description

Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Adult" dataset.

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	350057

Data

- UCI, extracted from 1994 Census database.
- 48,842 instances, mix of continuous and discrete features (train=32,561, test=16,281) containing some unknown features
- Randomly shuffled and split train data into train and test sets
- Labels
  - Multi-Class: No HS, Some HS, HS grad, Some College, College Grad, Masters, Doctorate
  - Binary: No College vs College

Data Attributes

- Age
- Work Class
- Education: This is what we're interested in predicting
- Marital Status
- Occupation
- Relationship
- Sex
- Hours Per Week
- Native Country
- Income

Methods

SVM

- Data preprocessing
    - 1 hot encoding for discrete features
    - Continuous features left alone
    - “?” = 0
- Used One vs. Rest classification for multi-class SVM
- Tuned hyper parameters
    - Limited in what we can tune with LinearSVC

Decision Trees

- Convert continous features to discrete
    - Caveat: Convert discrete features to binary features
- Arbitrary bin sizes
- Off the shelf implementation

Naive Bayes

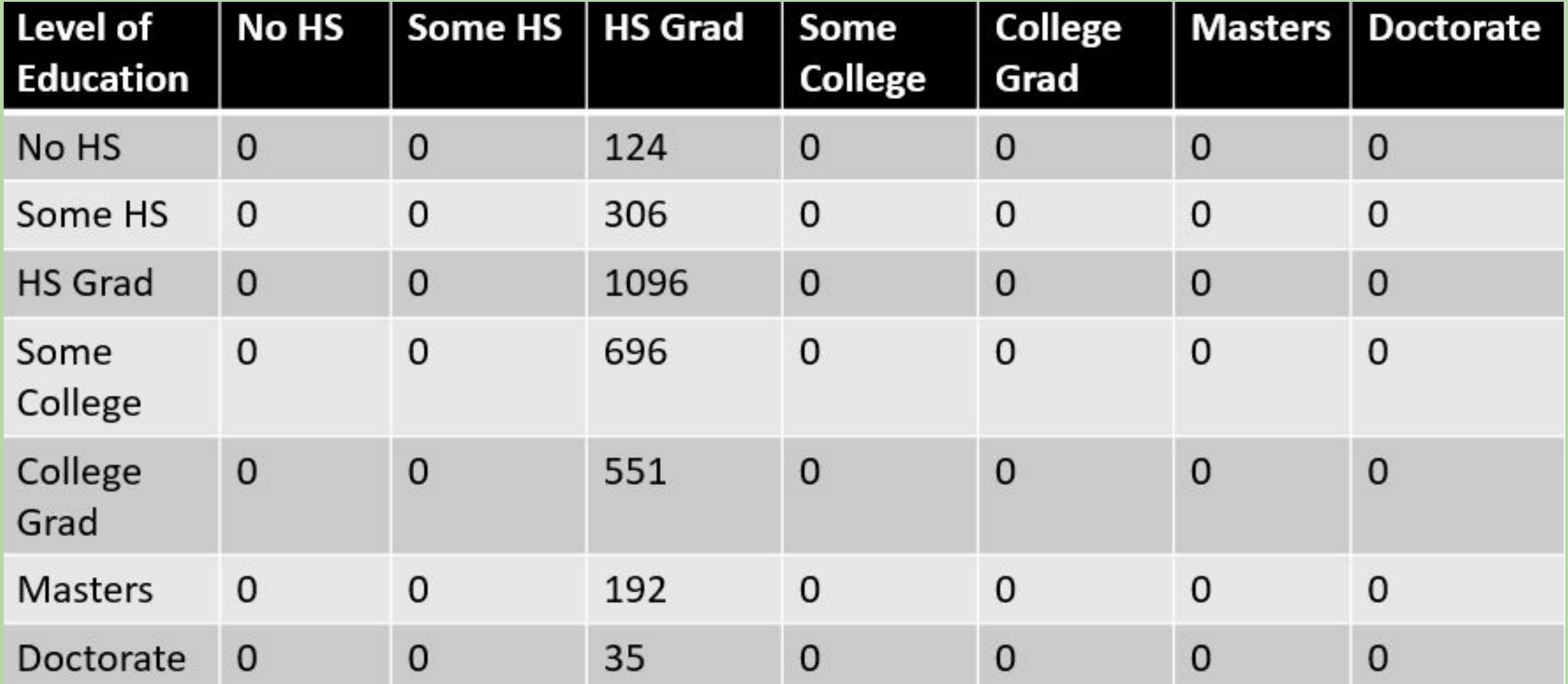
- Created an “unknown” value for each feature
- Convert continuous features to discrete
- Off the shelf implementation

Results and Interpretations

Most Frequent Class

- Identifies the most frequent class in train data
- Labels all test data as MFC
- Used as a baseline

Accuracy Score: 36.5%



Level of Education	No HS	Some HS	HS Grad	Some College	College Grad	Masters	Doctorate
No HS	0	0	124	0	0	0	0
Some HS	0	0	306	0	0	0	0
HS Grad	0	0	1096	0	0	0	0
Some College	0	0	696	0	0	0	0
College Grad	0	0	551	0	0	0	0
Masters	0	0	192	0	0	0	0
Doctorate	0	0	35	0	0	0	0

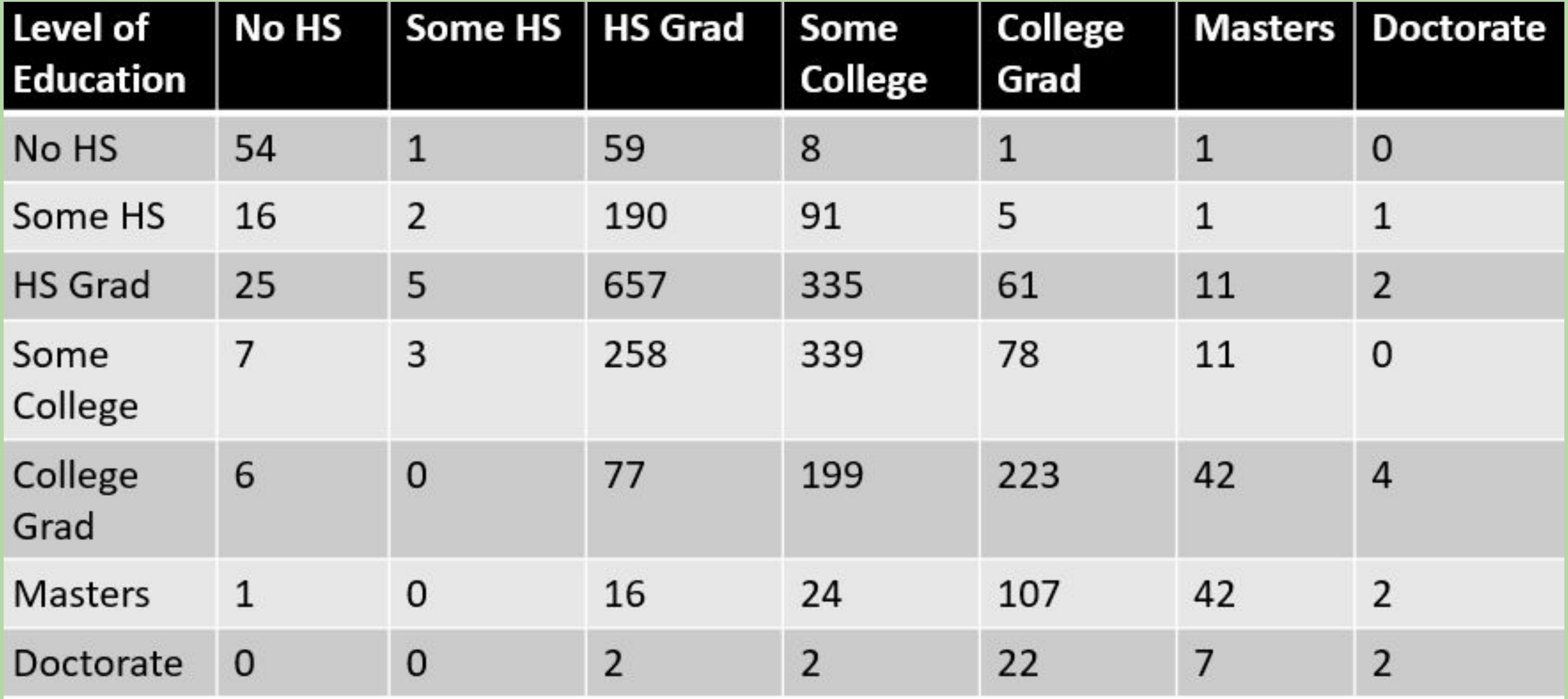
Accuracy Score: 36.5%

Level of Education	No HS	Some HS	HS Grad	Some College	College Grad	Masters	Doctorate
No HS	0	0	124	0	0	0	0
Some HS	0	0	306	0	0	0	0
HS Grad	0	0	1096	0	0	0	0
Some College	0	0	696	0	0	0	0
College Grad	0	0	551	0	0	0	0
Masters	0	0	192	0	0	0	0
Doctorate	0	0	35	0	0	0	0

SVM

- Achieved Highest Results!

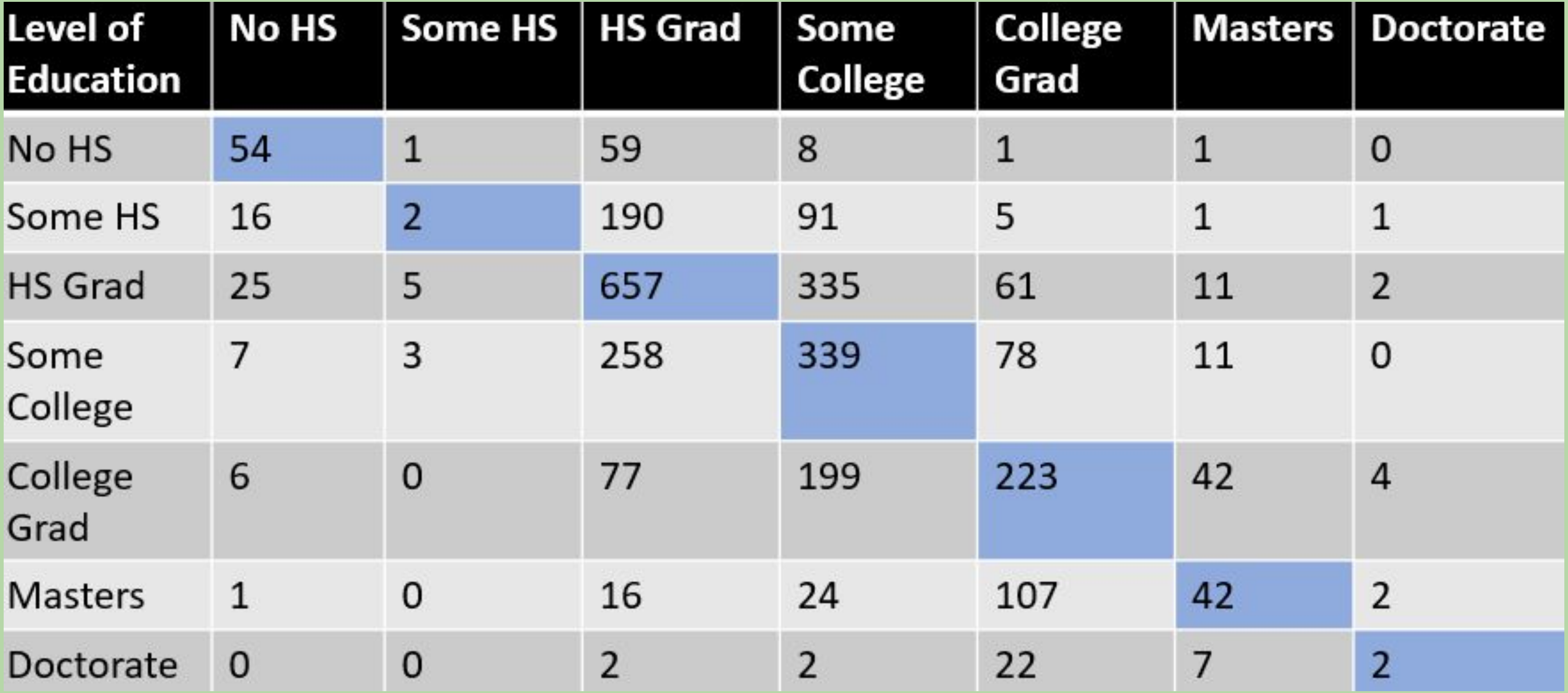
Accuracy Score: 44.0%



Level of Education	No HS	Some HS	HS Grad	Some College	College Grad	Masters	Doctorate
No HS	54	1	59	8	1	1	0
Some HS	16	2	190	91	5	1	1
HS Grad	25	5	657	335	61	11	2
Some College	7	3	258	339	78	11	0
College Grad	6	0	77	199	223	42	4
Masters	1	0	16	24	107	42	2
Doctorate	0	0	2	2	22	7	2

Highest Score: 46.3%

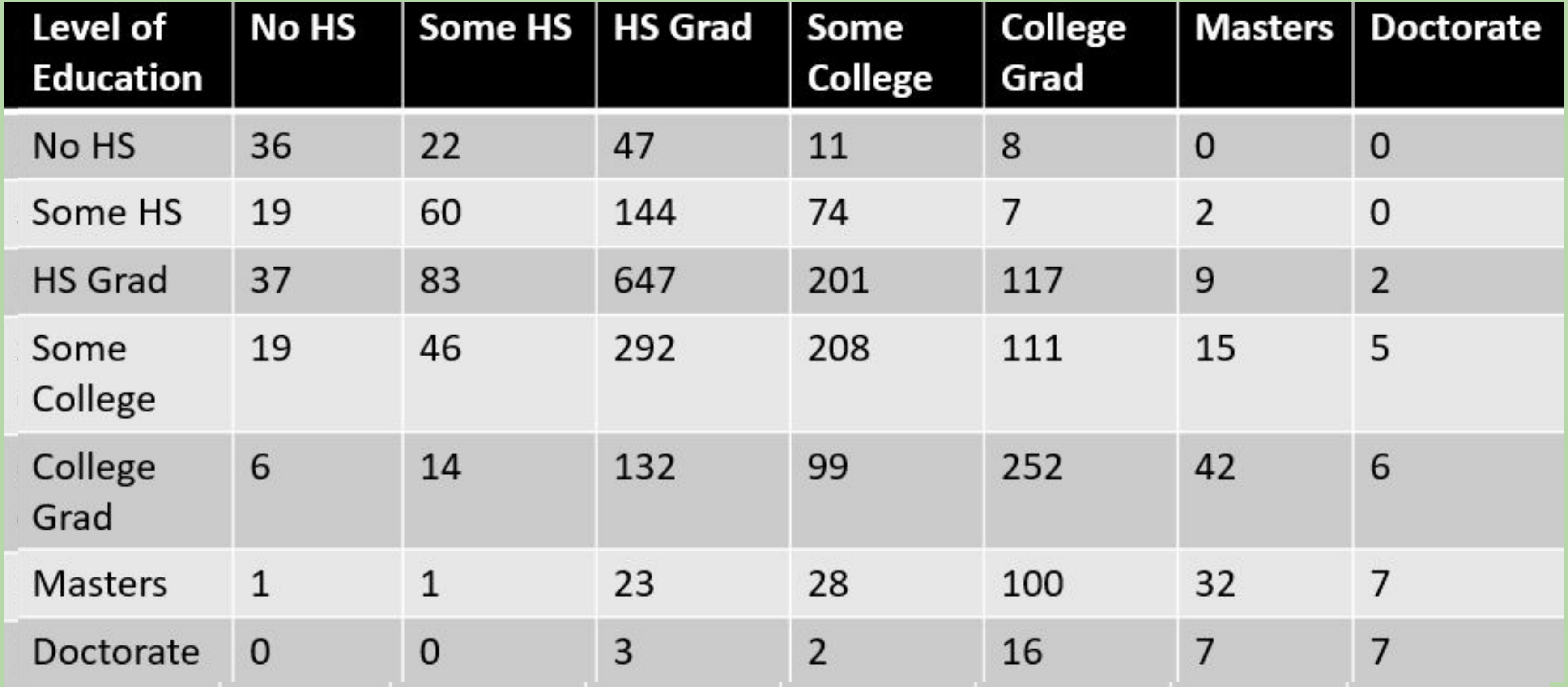
Accuracy Score: 44.0%



Level of Education	No HS	Some HS	HS Grad	Some College	College Grad	Masters	Doctorate
No HS	54	1	59	8	1	1	0
Some HS	16	2	190	91	5	1	1
HS Grad	25	5	657	335	61	11	2
Some College	7	3	258	339	78	11	0
College Grad	6	0	77	199	223	42	4
Masters	1	0	16	24	107	42	2
Doctorate	0	0	2	2	22	7	2

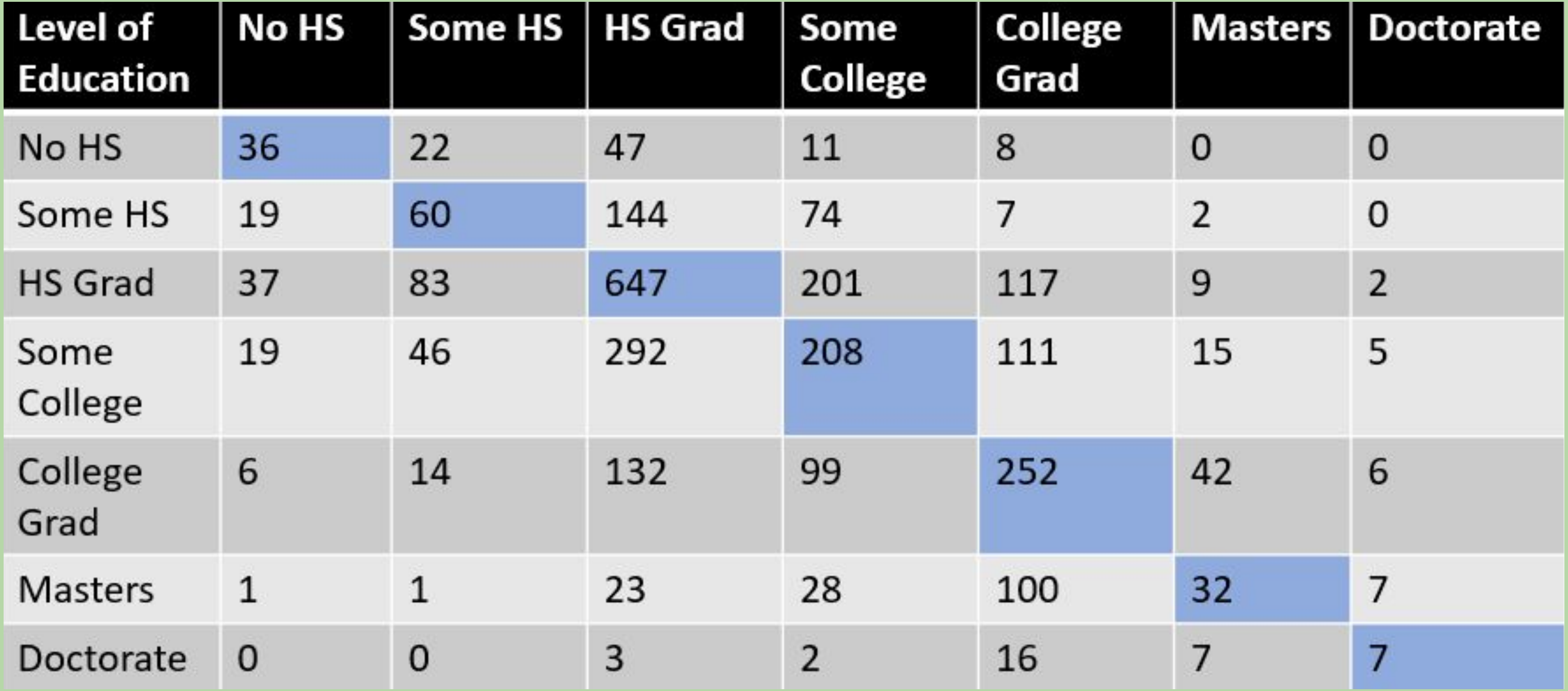
Decision Trees

Accuracy Score: 41.4%



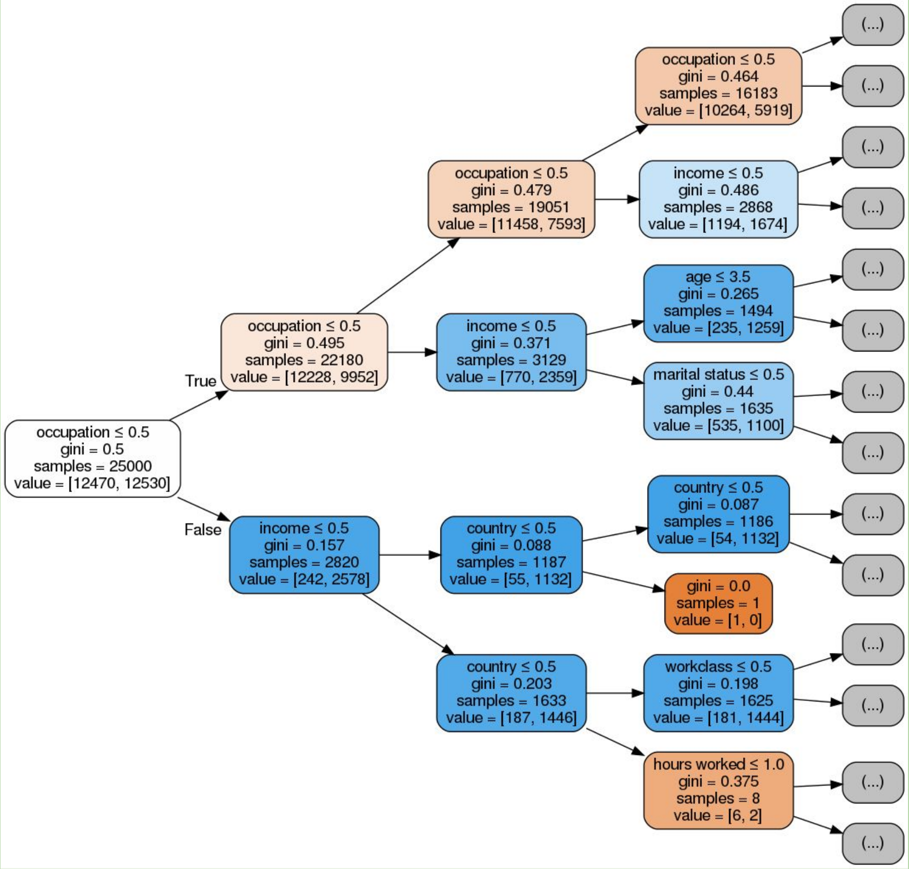
Level of Education	No HS	Some HS	HS Grad	Some College	College Grad	Masters	Doctorate
No HS	36	22	47	11	8	0	0
Some HS	19	60	144	74	7	2	0
HS Grad	37	83	647	201	117	9	2
Some College	19	46	292	208	111	15	5
College Grad	6	14	132	99	252	42	6
Masters	1	1	23	28	100	32	7
Doctorate	0	0	3	2	16	7	7

Accuracy Score: 41.4%



Level of Education	No HS	Some HS	HS Grad	Some College	College Grad	Masters	Doctorate
No HS	36	22	47	11	8	0	0
Some HS	19	60	144	74	7	2	0
HS Grad	37	83	647	201	117	9	2
Some College	19	46	292	208	111	15	5
College Grad	6	14	132	99	252	42	6
Masters	1	1	23	28	100	32	7
Doctorate	0	0	3	2	16	7	7

Tree Visual!

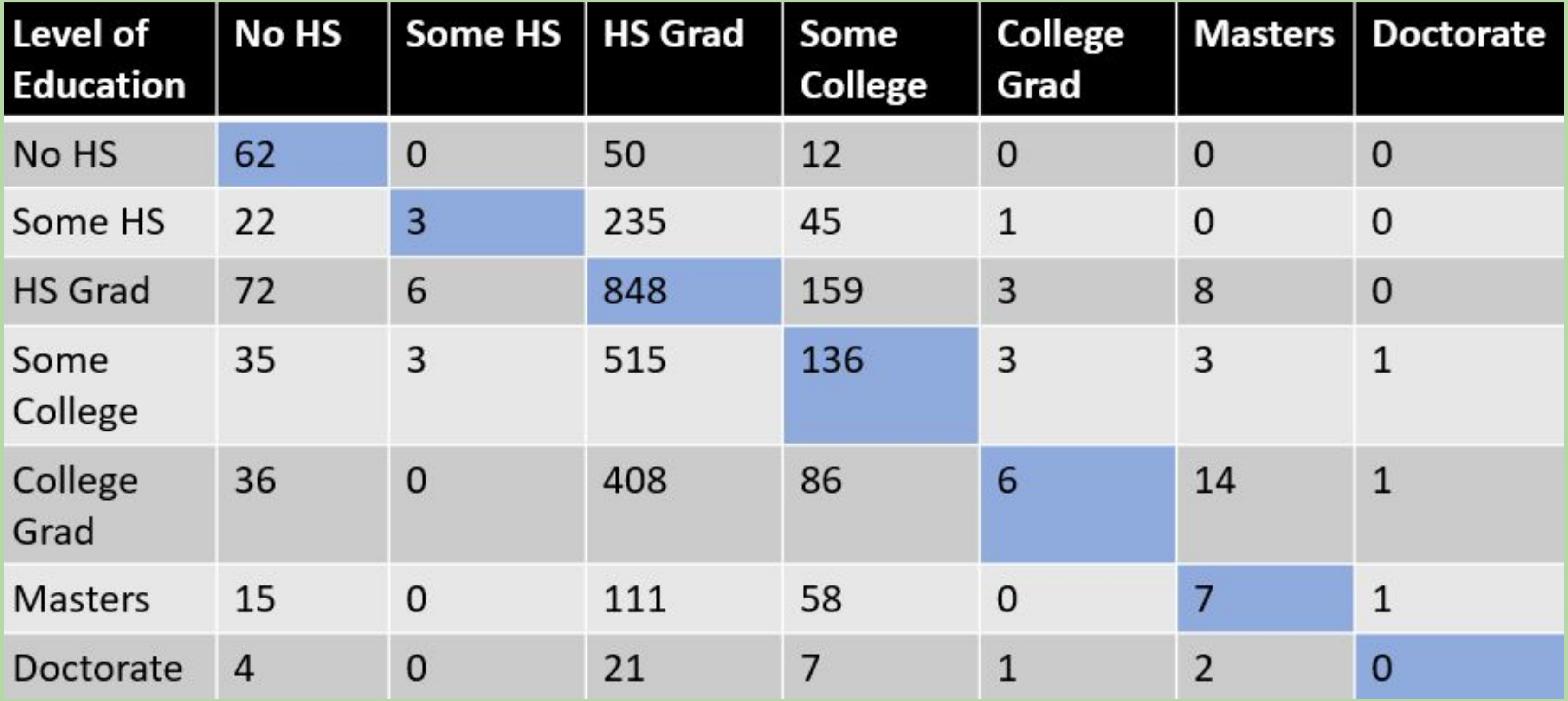

occupation ≤ 0.5 gini = 0.5 samples = 25000 value = [12470, 12530]
True
False
occupation ≤ 0.5 gini = 0.495 samples = 22180 value = [12228, 9952]
income ≤ 0.5 gini = 0.157 samples = 2820 value = [242, 2578]
occupation ≤ 0.5 gini = 0.479 samples = 19051 value = [11458, 7593]
income ≤ 0.5 gini = 0.371 samples = 3129 value = [770, 2359]
country ≤ 0.5 gini = 0.088 samples = 1187 value = [55, 1132]
country ≤ 0.5 gini = 0.203 samples = 1633 value = [187, 1446]
occupation ≤ 0.5 gini = 0.464 samples = 16183 value = [10264, 5919]
income ≤ 0.5 gini = 0.486 samples = 2868 value = [1194, 1674]
age ≤ 3.5 gini = 0.265 samples = 1494 value = [235, 1259]
marital status ≤ 0.5 gini = 0.44 samples = 1635 value = [535, 1100]
country ≤ 0.5 gini = 0.087 samples = 1186 value = [54, 1132]
gini = 0.0 samples = 1 value = [1, 0]
workclass ≤ 0.5 gini = 0.198 samples = 1625 value = [181, 1444]
hours worked ≤ 1.0 gini = 0.375 samples = 8 value = [6, 2]
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

Naive Bayes

Accuracy Score: 35.4%

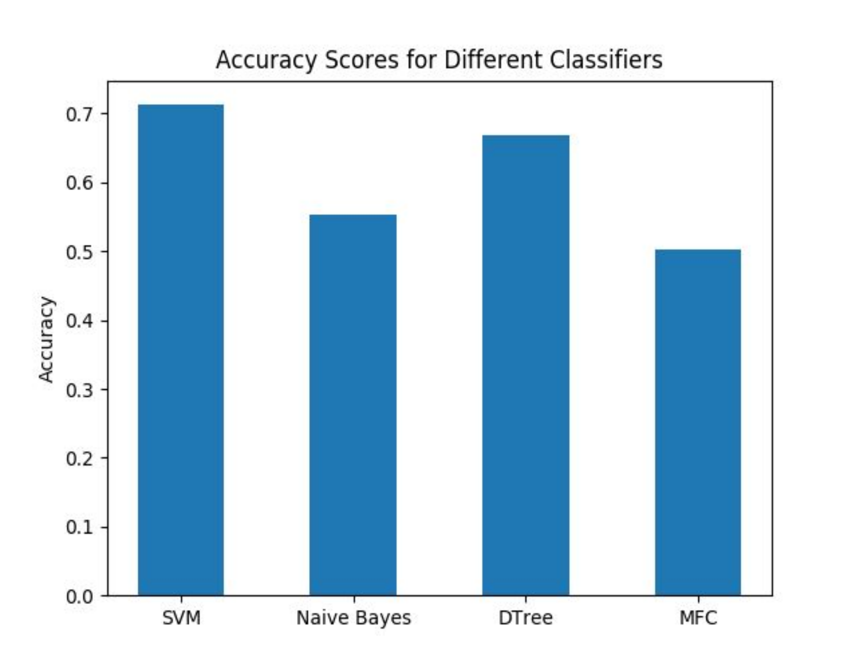
Level of Education	No HS	Some HS	HS Grad	Some College	College Grad	Masters	Doctorate
No HS	62	0	50	12	0	0	0
Some HS	22	3	235	45	1	0	0
HS Grad	72	6	848	159	3	8	0
Some College	35	3	515	136	3	3	1
College Grad	36	0	408	86	6	14	1
Masters	15	0	111	58	0	7	1
Doctorate	4	0	21	7	1	2	0

Accuracy Score: 35.4%

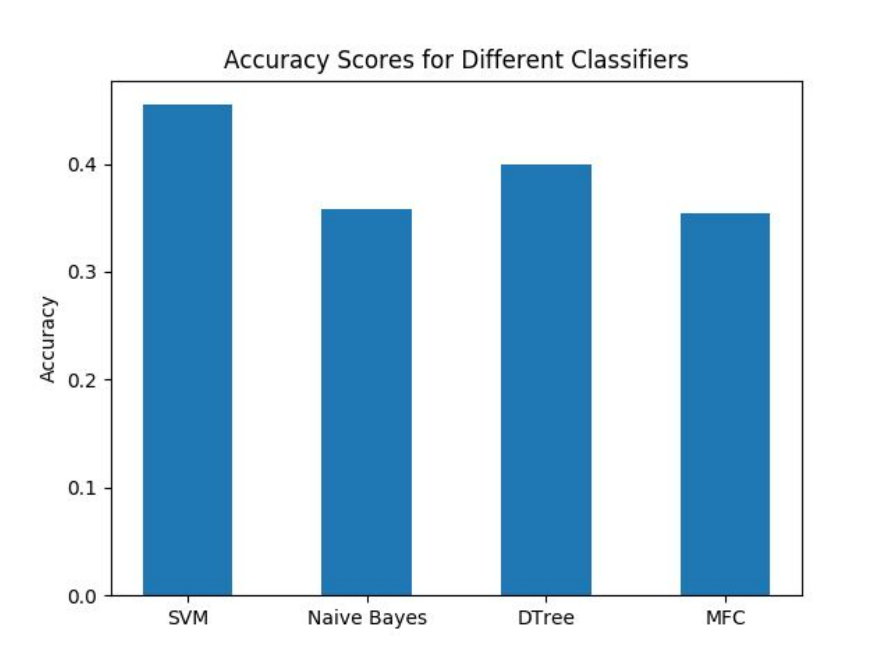


Level of Education	No HS	Some HS	HS Grad	Some College	College Grad	Masters	Doctorate
No HS	62	0	50	12	0	0	0
Some HS	22	3	235	45	1	0	0
HS Grad	72	6	848	159	3	8	0
Some College	35	3	515	136	3	3	1
College Grad	36	0	408	86	6	14	1
Masters	15	0	111	58	0	7	1
Doctorate	4	0	21	7	1	2	0

Interpretation: Overall Score Comparison



Binary Classification Task



Multiclass Classification Task

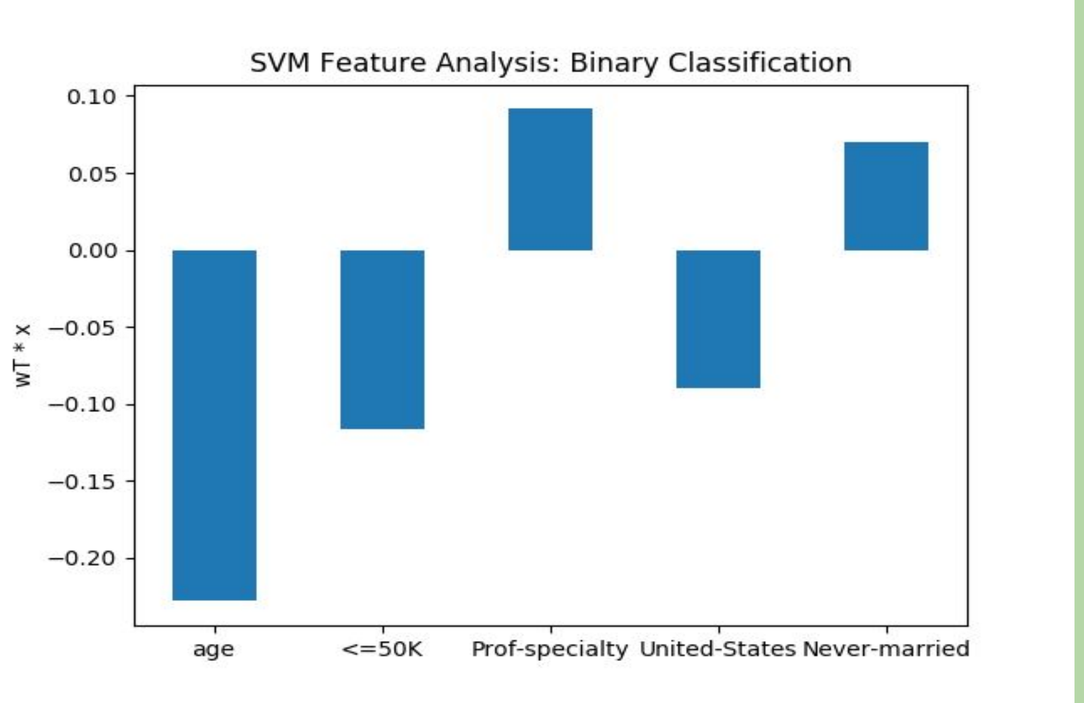
Feature Analysis: Multi-Class SVM

- Using LinearSVC, we train *n* models, where *n* is the number of classes.
- coef_ (ie. the matrix of coefficients) therefore has shape [n_class, n_features]
- Feature analysis is complex and is done for each model in relation to the rest

Feature Analysis: Binary SVM Classification

Here were the most impactful features:

- age : -0.228226
- <=50K : -0.116456
- Prof-specialty : 0.091492
- United-States : -0.089730
- Never-married : 0.069535



Conclusions and Future Work

Conclusion

- SVM model produced best results
- Income, occupation, age are the most important features (according to our Tree and SVM analysis)

Future Considerations

- Account for collinear independent variables
- View problem as regression rather than classification
- Smarter ways to create bin sizes

The Team

Kenny Gwon

Raymond Liu

Richard Muniu

Thoughts and Questions