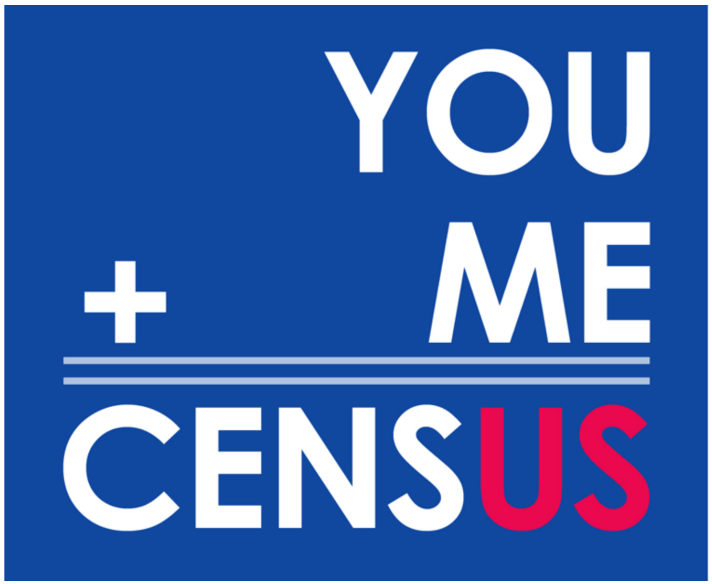
Who Makes the Benjamins?

A COMPARISON OF 3 MACHINE LEARNING MODELS TO PREDICT SALARY

TIM FRAHME

What's the 411?

- **FACT:** We all want to be rich
- **QUESTION:** How do I get rich, or die trying?
- **ANSWER:** Lets look at the traits of rich people
- Data was a subset 1994 Census Data (Kaggle)
- Predict who makes more than $50k
  - AKA -> I can afford to eat at Olive Garden instead of McDonalds



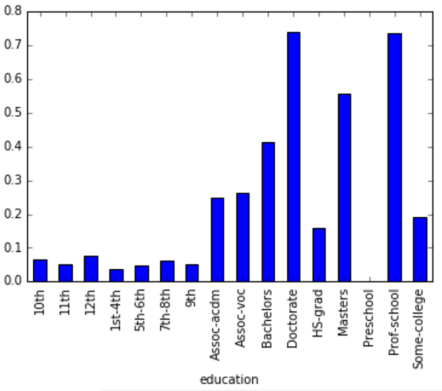
What are we working with?

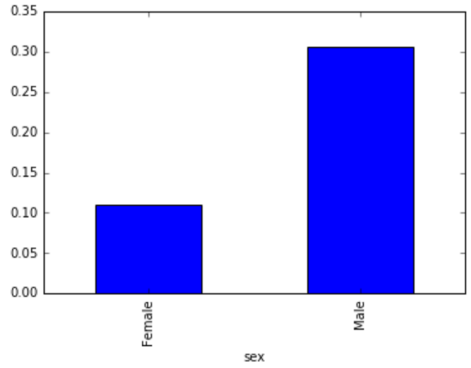
Base Case	
Total Records	32,561
People > 50k	7,841
Percentage	0.2408

```
Data columns (total 15 columns):
age                32561 non-null int64
workclass          32561 non-null object
fnlwgt             32561 non-null int64
education          32561 non-null object
education_num      32561 non-null int64
marital_status     32561 non-null object
occupation         32561 non-null object
relationship       32561 non-null object
race               32561 non-null object
sex                32561 non-null object
capital_gain       32561 non-null int64
capital_loss       32561 non-null int64
hours              32561 non-null int64
native_country     32561 non-null object
income             32561 non-null object
```

Let's go Exploring!

- Graphed the attributes against Salary
- Created Dummy Variables
- Looked at Correlation Matrix






```
census.corr()
```

	age	fnlwgt	education_num	capital_gain	capital_loss	hours	outcome	workclass_Federal-gov
age	1.000000	-0.076646	0.036527	0.077674	0.057775	0.068756	0.234037	0.051227
fnlwgt	-0.076646	1.000000	-0.043195	0.000432	-0.010252	-0.018768	-0.009463	-0.007525
education_num	0.036527	-0.043195	1.000000	0.122630	0.079923	0.148123	0.335154	0.060518
capital_gain	0.077674	0.000432	0.122630	1.000000	-0.031615	0.078409	0.223329	-0.005768
capital_loss	0.057775	-0.010252	0.079923	-0.031615	1.000000	0.054256	0.150526	0.010798
hours	0.068756	-0.018768	0.148123	0.078409	0.054256	1.000000	0.229689	0.013293
outcome	0.234037	-0.009463	0.335154	0.223329	0.150526	0.229689	1.000000	0.059372

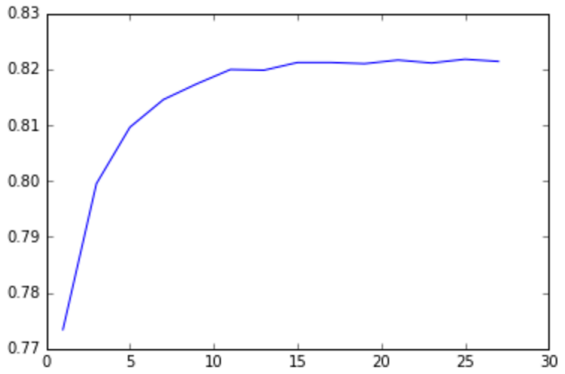
KNN

- Tested a number of different combinations of features
- 5 fold Cross Validation and Grid Search
- Best:
  - Null: .2408
  - Accuracy: .8211
  - Roc/Auc: .8280

```
knn = KNeighborsClassifier(15)
cross_val_score(knn, X, y, cv=5, scoring='accuracy').mean()
```

```
0.82119720585289446
```

```
[<matplotlib.lines.Line2D at 0x17b411390>]
```




```
cross_val_score(knn, X, y, cv=5, scoring='roc_auc').mean()
```

```
0.82804620583833233
```

Logistic Regression/Naive Bayes

- Tested Logistic Regression and Naive Bayes
- Null: .2408
- Logistic Regression
  - Accuracy: .8289
  - Roc/Auc: .8770
- Naive Bayes
  - Accuracy: .8017
  - Roc/Auc: .8371

```
#Logistic Regression
```

```
logreg = LogisticRegression()
cross_val_score(logreg, X, y, cv=5, scoring='accuracy').mean()
```

```
0.82896698570351268
```

```
cross_val_score(logreg, X, y, cv=5, scoring='roc_auc').mean()
```

```
0.87706519962715623
```

```
#Naive Bayes
```

```
nb = MultinomialNB()
cross_val_score(nb, X, y, cv=5, scoring='accuracy').mean()
```

```
0.80172587163605125
```

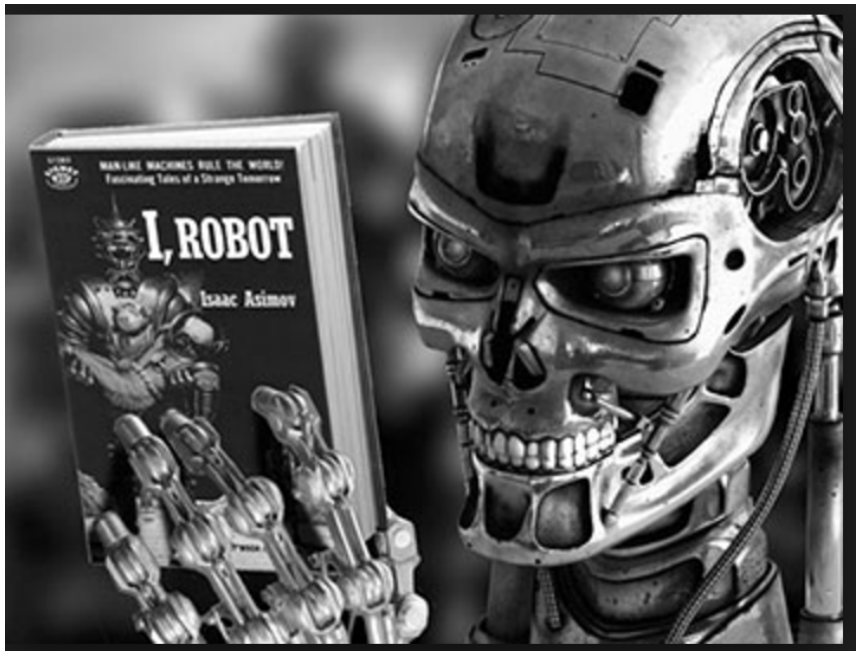
```
cross_val_score(nb, X, y, cv=5, scoring='roc_auc').mean()
```

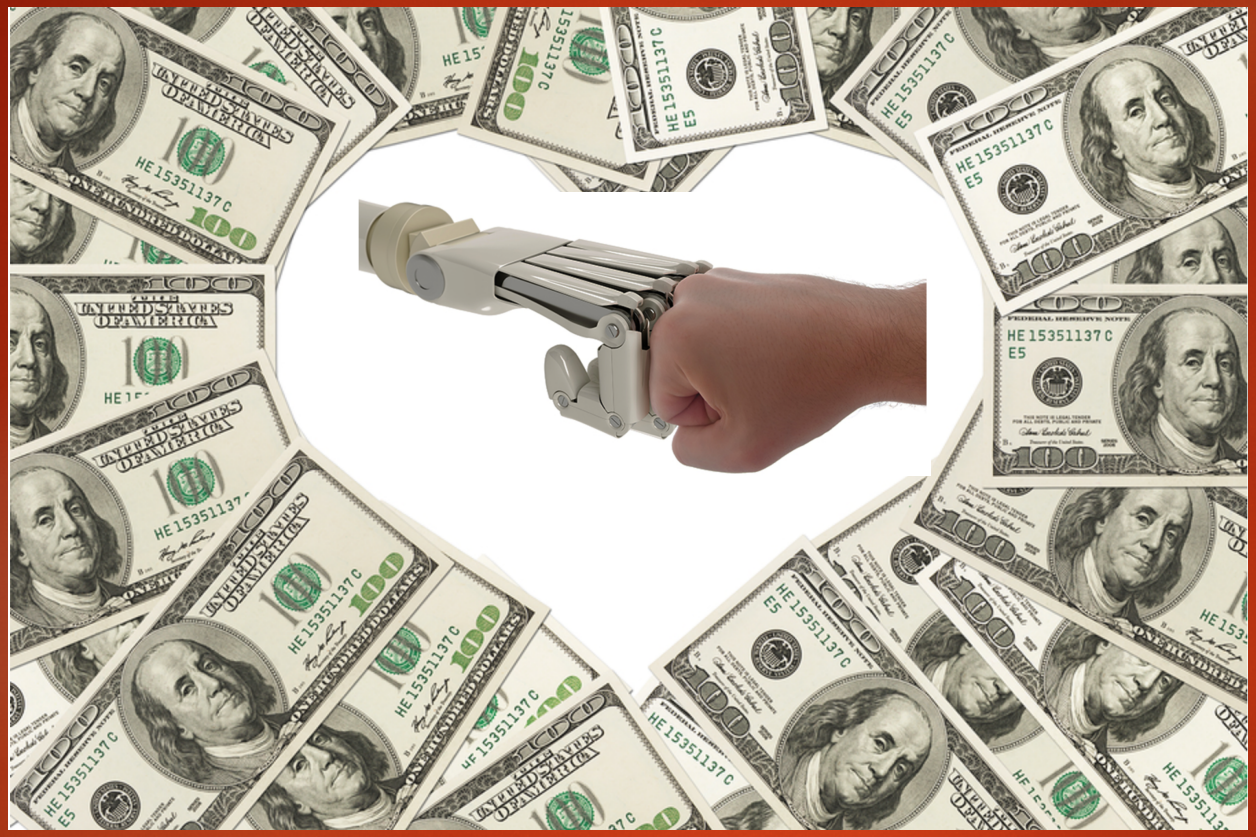
```
0.8371324333027429
```

What Model Learned the Best?

Features: Age, Education, Workclass, Marital Status, Occupation, Race, Sex				
	Knn	Logistic Regression	Naive Bayes	Base Case
CV	5	5	5	
Knn Neighbors	15			
Accuracy	0.82119	0.82896	0.80172	0.2408
Roc Auc	0.86	0.87706	0.83713	

Features: Age, Education, Workclass, Marital Status, Occupation, Race, Sex, Hours, Capital Gains				
	Knn	Logistic Regression	Naive Bayes	Base Case
CV	5	5	5	
Knn Neighbors	10			
Accuracy	0.78815	0.82101	0.7789	0.2408
Roc Auc	0.75409	0.88638	0.5859	





Thank You!

Appendix

iPython Notebook
https://github.com/tjfrahme/sfdat26-frahme/blob/master/Final/Census.ipynb

Github Folder
https://github.com/tjfrahme/sfdat26-frahme/tree/master/Final