Census Income Analysis

By Emma Shi, Matthew Sebastian, Kathryn Le

Introduction

- [1994 Census Income Data Set]
- Over 30,000 people
- 14 input features originally
- Over 30,000 observations

	Age	Workclass	FNLWGT	Education	EducationNum	MaritalStatus	Occupation	Relationship	Race	Sex	CapitalGain	CapitalLoss	HoursPerWeek	NativeCountry	Income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

Machine Learning Question

Does a person given a set of characteristics make below or above 50k a year?

Features and Feature Engineering

Original Features:

- Age
- ~~Education~~
- Education Number
- Workclass
- Marital Status
- ~~Relationship~~
- Occupation
- Native Country
- ~~FNLWGT~~
- Sex
- Race
- ~~Capital Gain/Loss~~
- Hours Per Week

Engineering

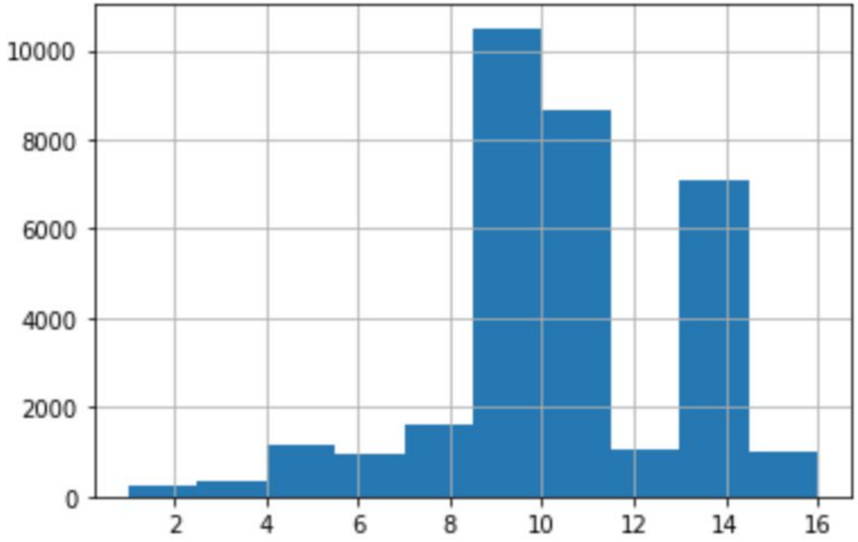
Engineered Features:

- Age_scaled (both standard and MinMax)
- EducationNum_scaled
- HoursPerWeek_scaled
- Sex_female/male (made binary)
- Race_white, black, other, etc. (binary)
- Oc_Cleaners, exec_managerial, etc.
- MS_married, divorced, etc.
- Native (native to US = 1)

EDA - Histograms

```
df.EducationNum.hist()
```

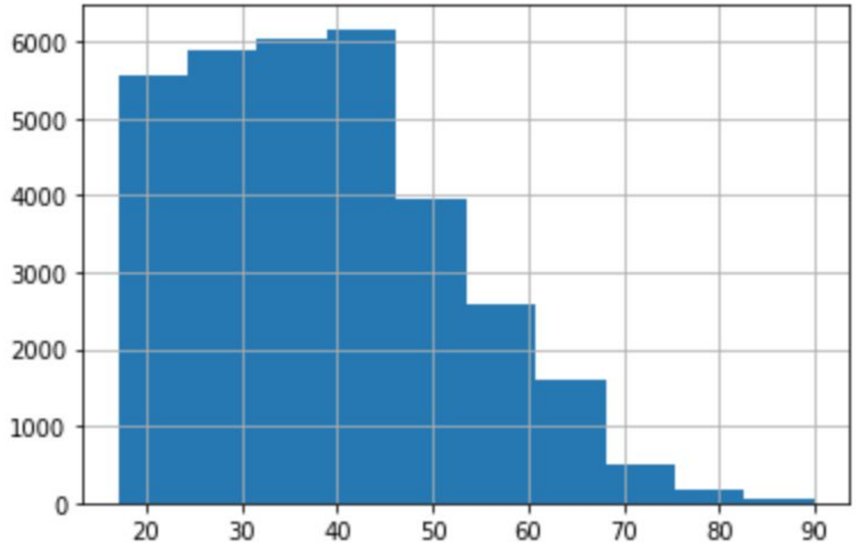
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f62e2
```



1) Histogram of people’s number of years of education

```
df.Age.hist()
```

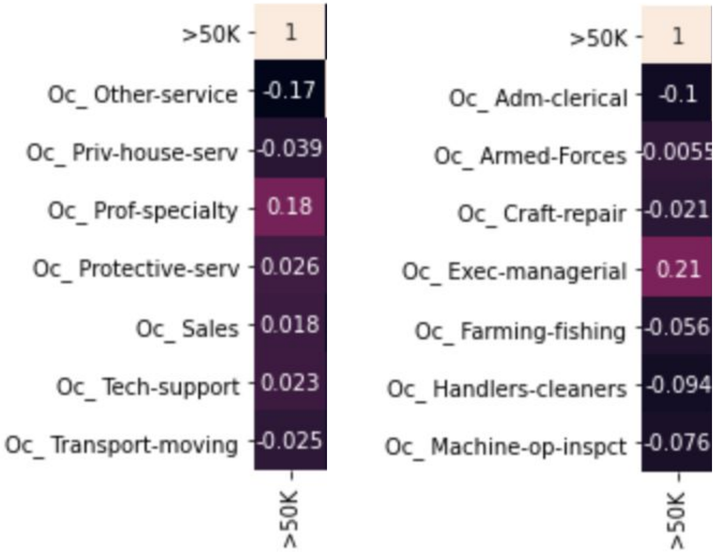
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f62
```



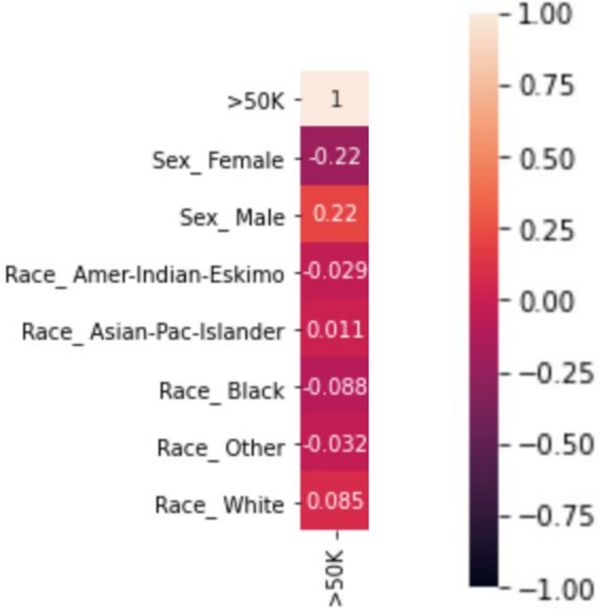
2) Histogram of people’s ages

EDA - Heatmaps

Age, Marital Status and income.



Occupation and Income



Race and Income

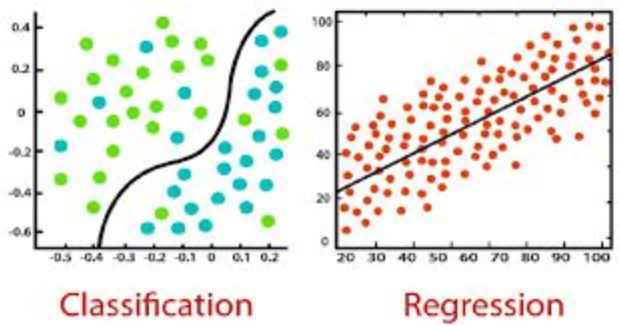
EDA Part 2 - Insights

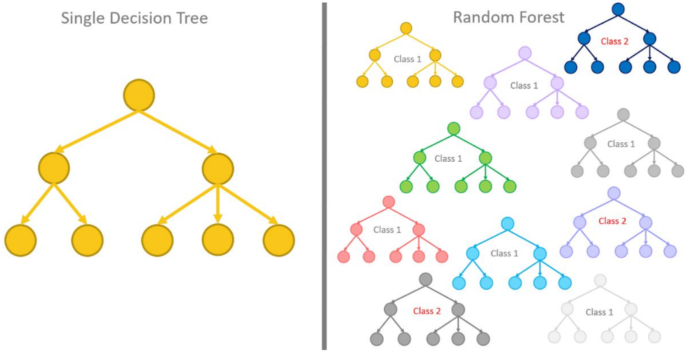
- Marital Status, Age, and Income have direct relationship
- Race has low direct correlation with income
- Race may have an indirect effect on income
- Sex did correlate with Income and affected model performance

>50K	0	1
EducationNum		
1	1.000000	0.000000
2	0.960265	0.039735
3	0.958333	0.041667
4	0.937163	0.062837
5	0.945055	0.054945
6	0.928049	0.071951
7	0.943702	0.056298
8	0.923077	0.076923
9	0.835671	0.164329
10	0.799940	0.200060
11	0.736802	0.263198
12	0.746032	0.253968
13	0.578509	0.421491
14	0.435771	0.564229
15	0.250923	0.749077
16	0.253333	0.746667

Models We Tried

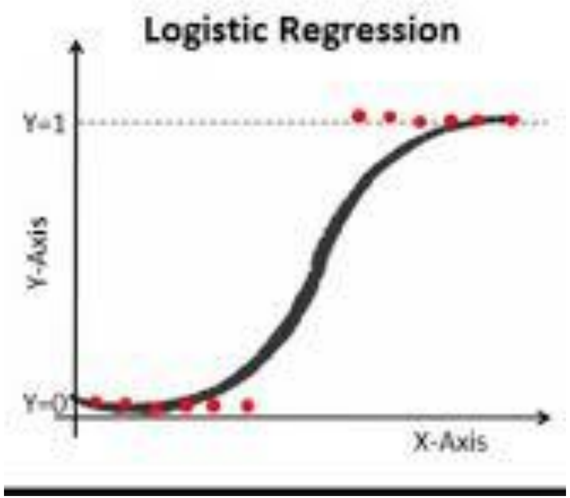
- Linear Regression
- Classification
  - K Nearest Classifier
  - Logistic Regression
- Clustering
  - K Means Clustering
- Bagging
  - Random Forest Classifier
- Boosting
  - AdaBoost Classifier
  - Gradient Boosting Classifier





Most Successful Model - LogR

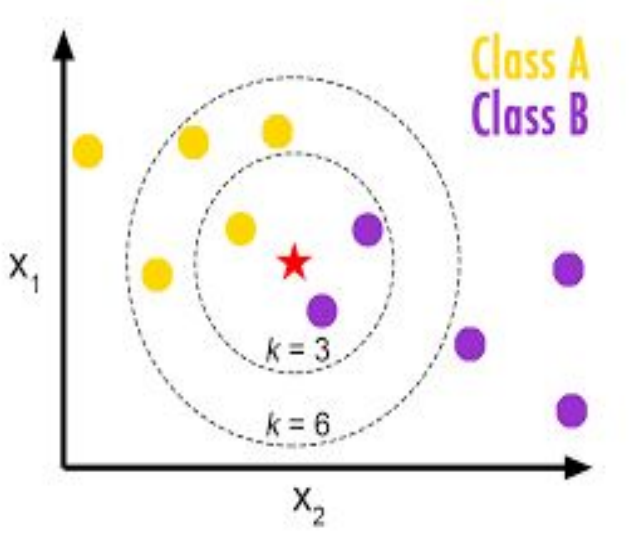
Logistic Regression with Sex, Occupation, Workclass, Marital Status, Hours Per Week, Age, and if “Native”, Education Number




```
Accuracy Score: 0.8289569657184537

Recall Score: 0.7371560091942069

Precision Score: 0.779115130313960б

F1 Score: 0.7536157338387605

KFold cross validation: 0.7528922335800193
```

How We Improved Our Model

- We used the Sequential Feature Selector (SFS) to determine which features were most impactful to the model, and removed unnecessary features
- Hypertuning parameters for Random Forest Classifier to get the best max_depth
- Hypertuning parameters for KNNs to get the optimal n_neighbors.



Audience Question - SFS

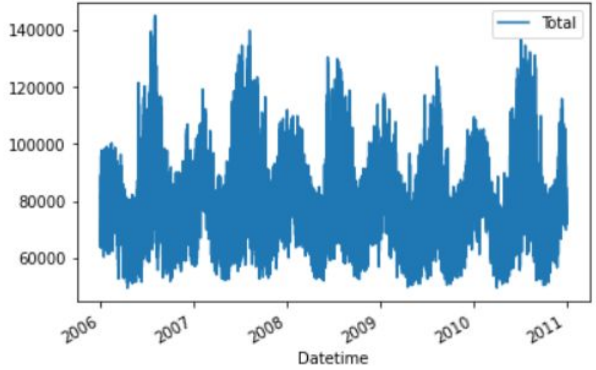
16: {'avg_score': 0.7285858500392628,
'ci_bound': 0.006488994865143978,
'cv_scores': array([0.72808874, 0.72444074, 0.7237961 , 0.73791665, 0.72868702]),
'feature_idx': (6, 7, 9, 10, 12, 13, 14, 15, 16, 22, 23, 24, 25, 31, 32, 35),
'feature_names': ('Race_ Other', 'Race_ White', 'Oc_ Armed-Forces', 'Oc_ Craft-repair', 'Oc_ Farming-fishing', 'Oc_ Handlers-cleaners', 'Oc_ Machine-op-inspct', 'Oc_ Other-service', 'Oc_ Priv-house-serv', 'MS_ Divorced', 'MS_ Married-AF-spouse', 'MS_ Married-civ-spouse', 'MS_ Married-spouse-absent', 'WC_ Never-worked', 'WC_ Private', 'WC_ State-gov')

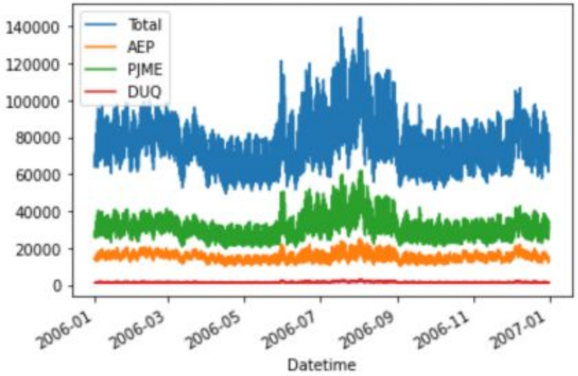
Based on our model and our interpretation, race did not seem to affect the model significantly and sex did yet race did make it onto the list of best features while sex did not. This is similar to the other features as well such as HoursPerWeek and Education Number. How might we explain this? What does this tell us about machine learning?

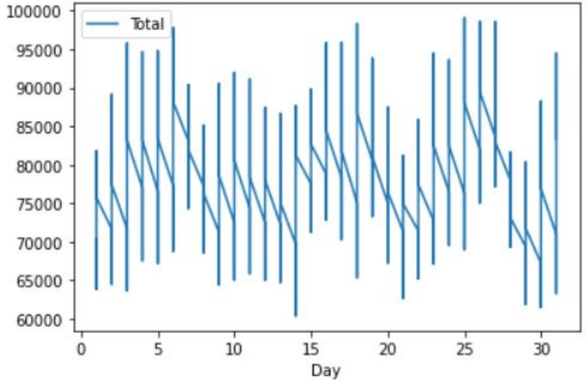
Conclusion - Real World Significance

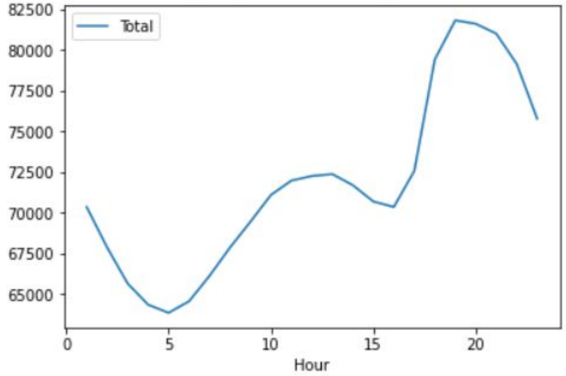
- ★ Can be used by employers to determine income
- ★ Can be used by people looking for jobs to see their future approximate income
- ★ Brings awareness to income inequality
  - ○ Racial or sexual discrimination
- ★ We can see what factors affect income and how much someone makes as well as patterns that we didn't see before

Power Consumption Introduction









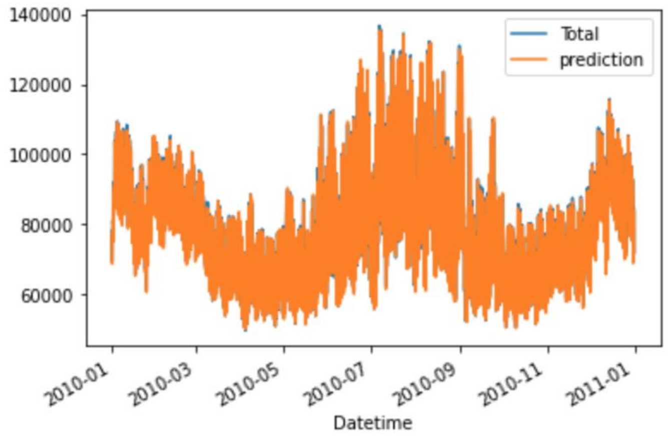
https://www.kaggle.com/robikscube/hourly-energy-consumption?select=DEOK_hourly.csv

Power Consumption Findings

- People tend to use consume more energy during the summer than in the winter
- We used the RandomForestRegressor and got a mean squared error of ~480 and Linear Regression with ~900 error.
- We made new features such as year, month, day, the the difference of energy consumption between each hour.

```
df1_holdout.plot.line(y=['Total','prediction'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0b98a
```



THANK YOU!

~ Group 3