# Project 4
# Hackathon - Am I getting paid enough?

By Gouri, Casey, Mac

## Problem Statement

Supervised binary classification problem where we have to predict if the wages are above 50,000K within 8 hours

## Project Constraint:

che csv

GOOD
THIS DOESNT EXIST
FAST
CHEAP

## Data:

- 6,513 - observations
- 14 - features

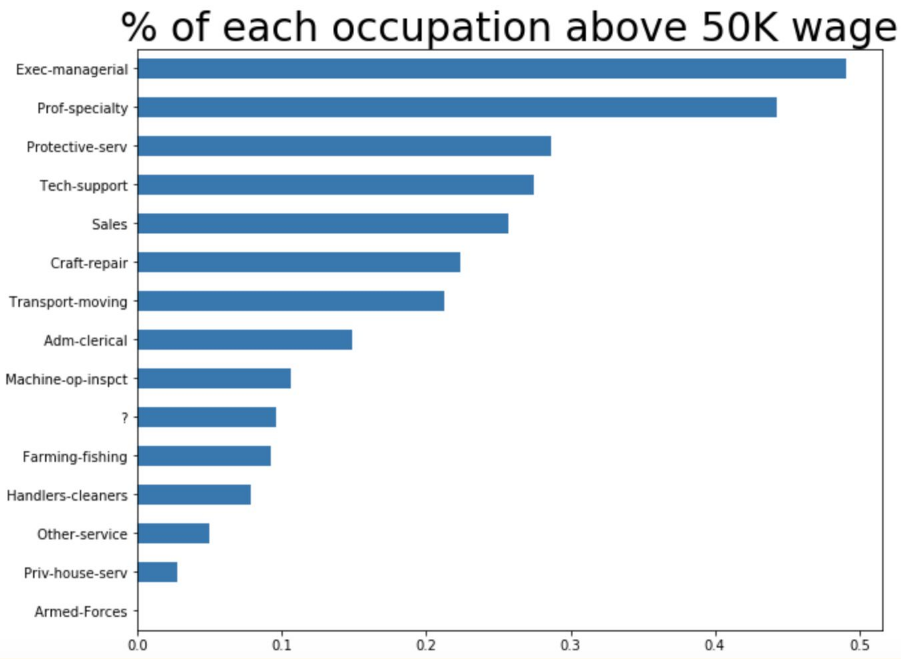
## Test Data:

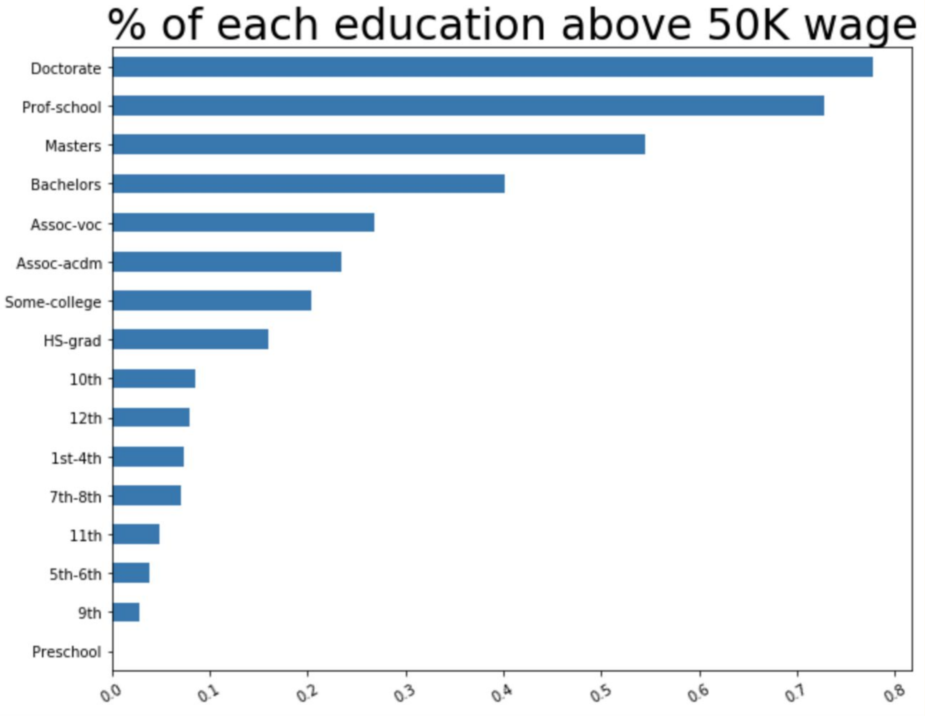
- 16,281 observations and 14 features

## Additional Features:

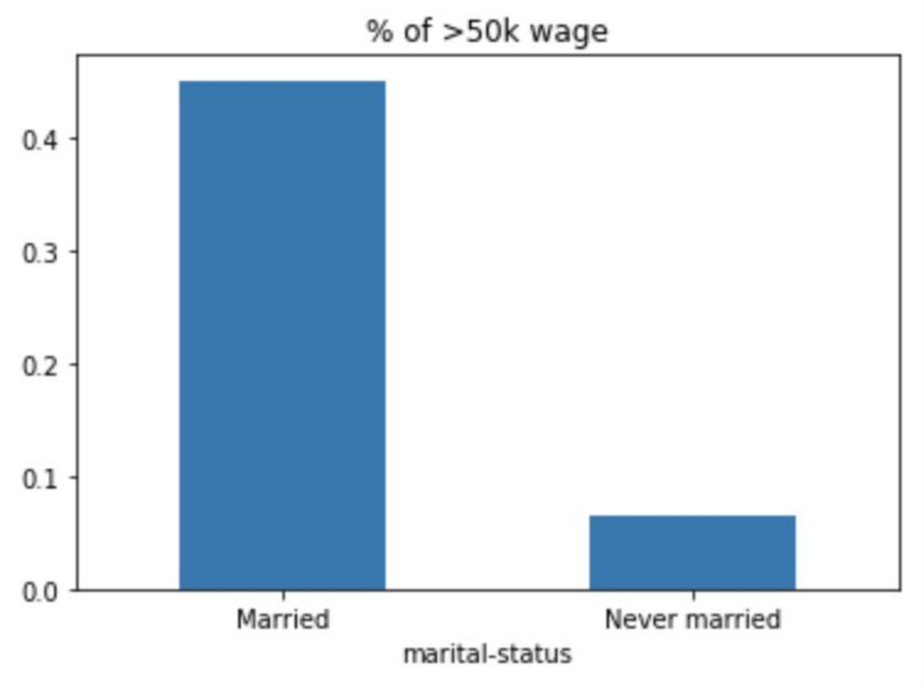
- Dummy Fields
- No of yrs of exp
- Bin age
- Bin hours per week

# Early EDA


% of each occupation above 50K wage


% of each education above 50K wage

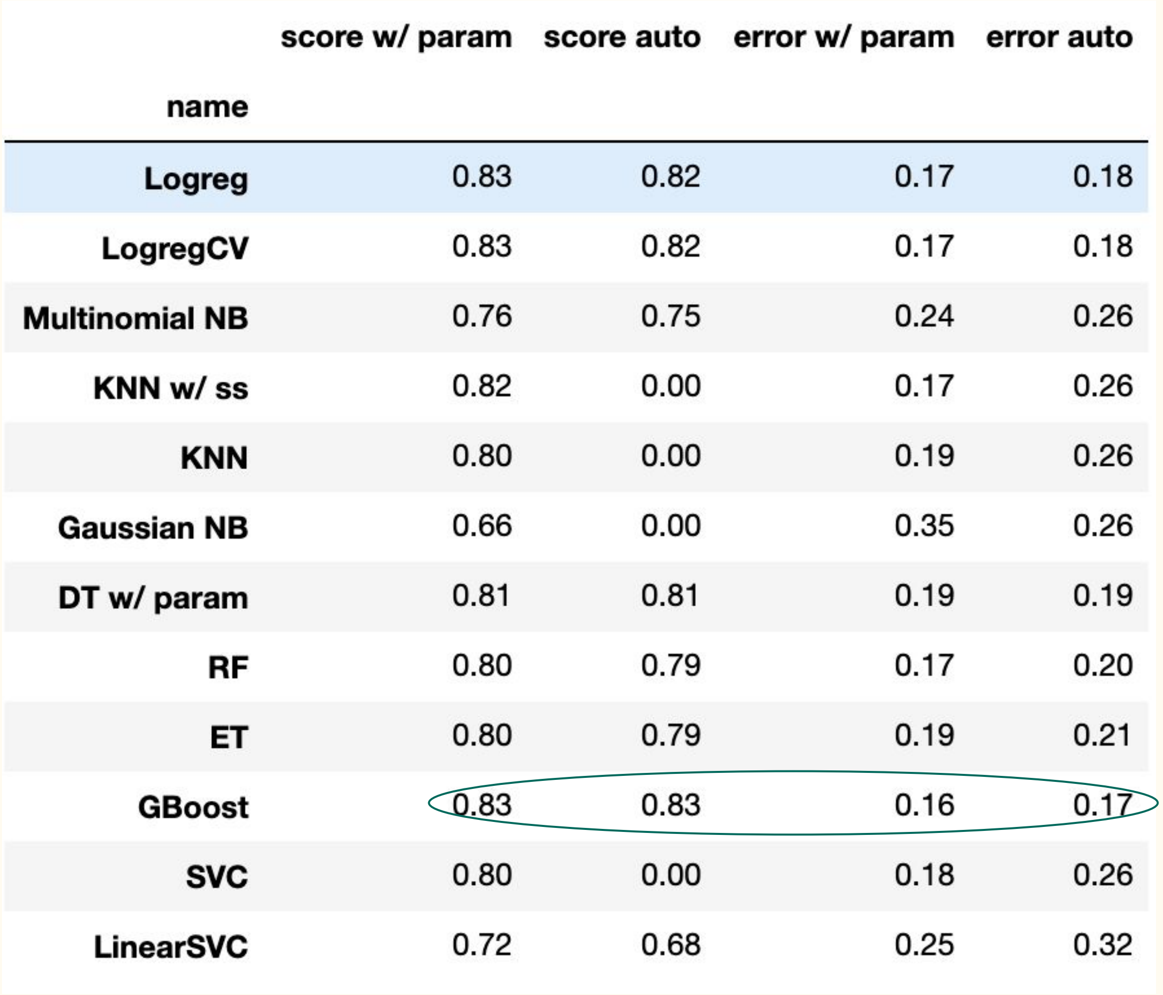
# Early EDA...


% of >50k wage
0.4
0.3
0.2
0.1
0.0
Married
Never married
marital-status

# Accuracy
# Error

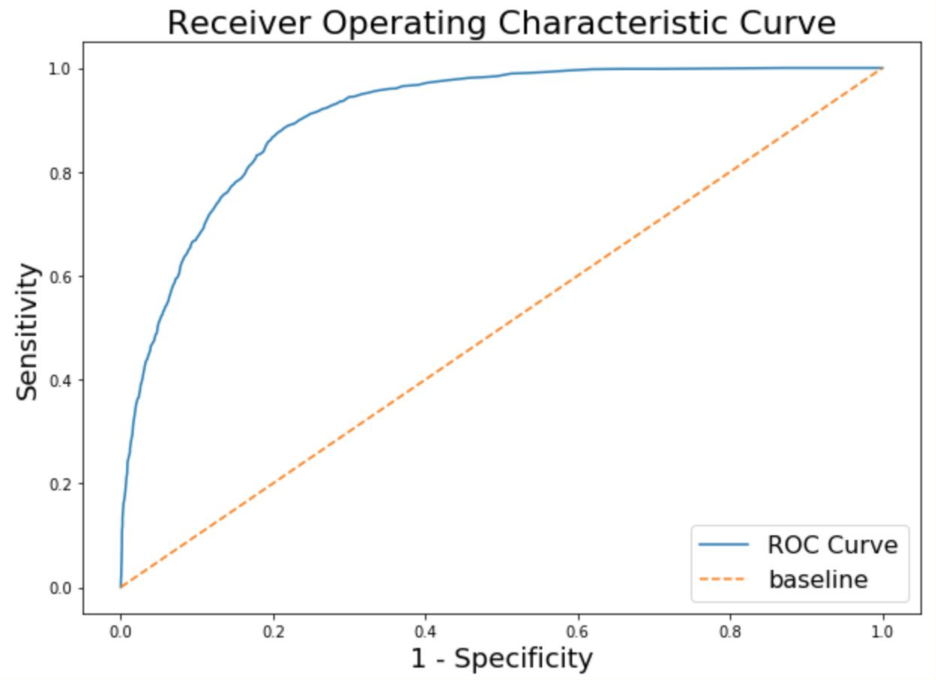
Baseline accuracy:

75%

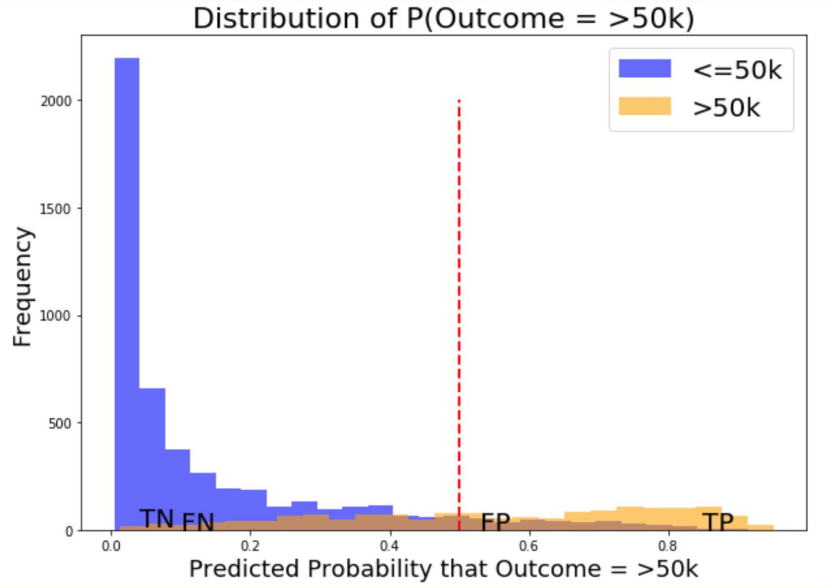


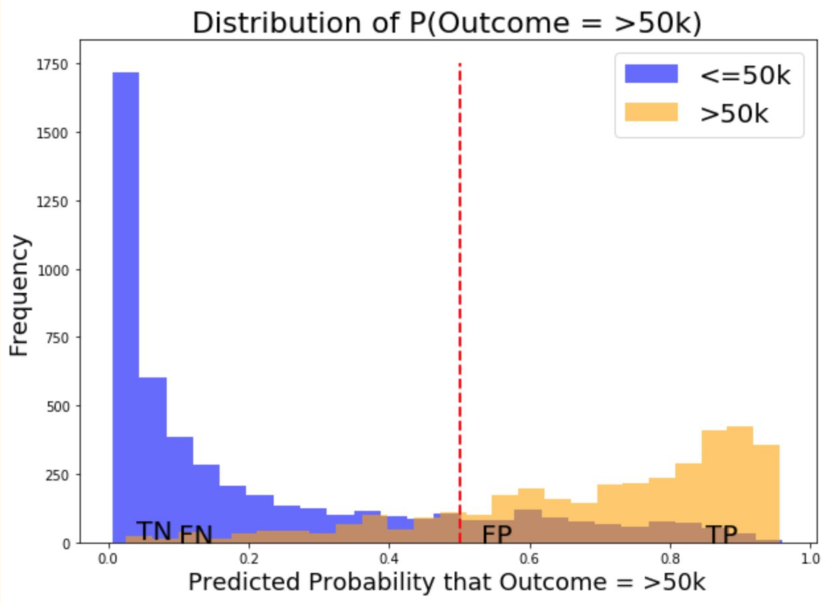
| | score w/ param | score auto | error w/ param | error auto |
|---|---|---|---|---|
| **name** | | | | |
| **Logreg** | 0.83 | 0.82 | 0.17 | 0.18 |
| **LogregCV** | 0.83 | 0.82 | 0.17 | 0.18 |
| **Multinomial NB** | 0.76 | 0.75 | 0.24 | 0.26 |
| **KNN w/ ss** | 0.82 | 0.00 | 0.17 | 0.26 |
| **KNN** | 0.80 | 0.00 | 0.19 | 0.26 |
| **Gaussian NB** | 0.66 | 0.00 | 0.35 | 0.26 |
| **DT w/ param** | 0.81 | 0.81 | 0.19 | 0.19 |
| **RF** | 0.80 | 0.79 | 0.17 | 0.20 |
| **ET** | 0.80 | 0.79 | 0.19 | 0.21 |
| **GBoost** | 0.83 | 0.83 | 0.16 | 0.17 |
| **SVC** | 0.80 | 0.00 | 0.18 | 0.26 |
| **LinearSVC** | 0.72 | 0.68 | 0.25 | 0.32 |

# ROC AUC


Receiver Operating Characteristic Curve
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve
baseline

# Given data vs bootstrapped data :balanced class





# Best Model

**Classifier :** GradientBoostingClassifier(max_depth=3, n_estimators=100, learning_rate=0.1)

- **Features :** Age, hours per week, marital-status, education num, sex , workclass, country
- **Bin Fields** - Age, hours per week
- **Dummy fields** - marital-status,sex , workclass and country
- **Ignored Fields:** fnlwgt, education, capital-gain, capital-loss

**Predicting the test data (16,281) for submission:**

- 13111 <= 50k
- 3170 >50k