Classifying Income Levels

Zakaria Zerhouni
Emmanuel Akindele
Steven Markoe

Dataset

- Observations taken from 1994 census data
- Demographic and financial information of 30,000+ individuals

https://archive.ics.uci.edu/ml/datasets/adult

Given only financial and demographic information, can we classify workers as having a ***salary greater than $50,000?***

Cleaning

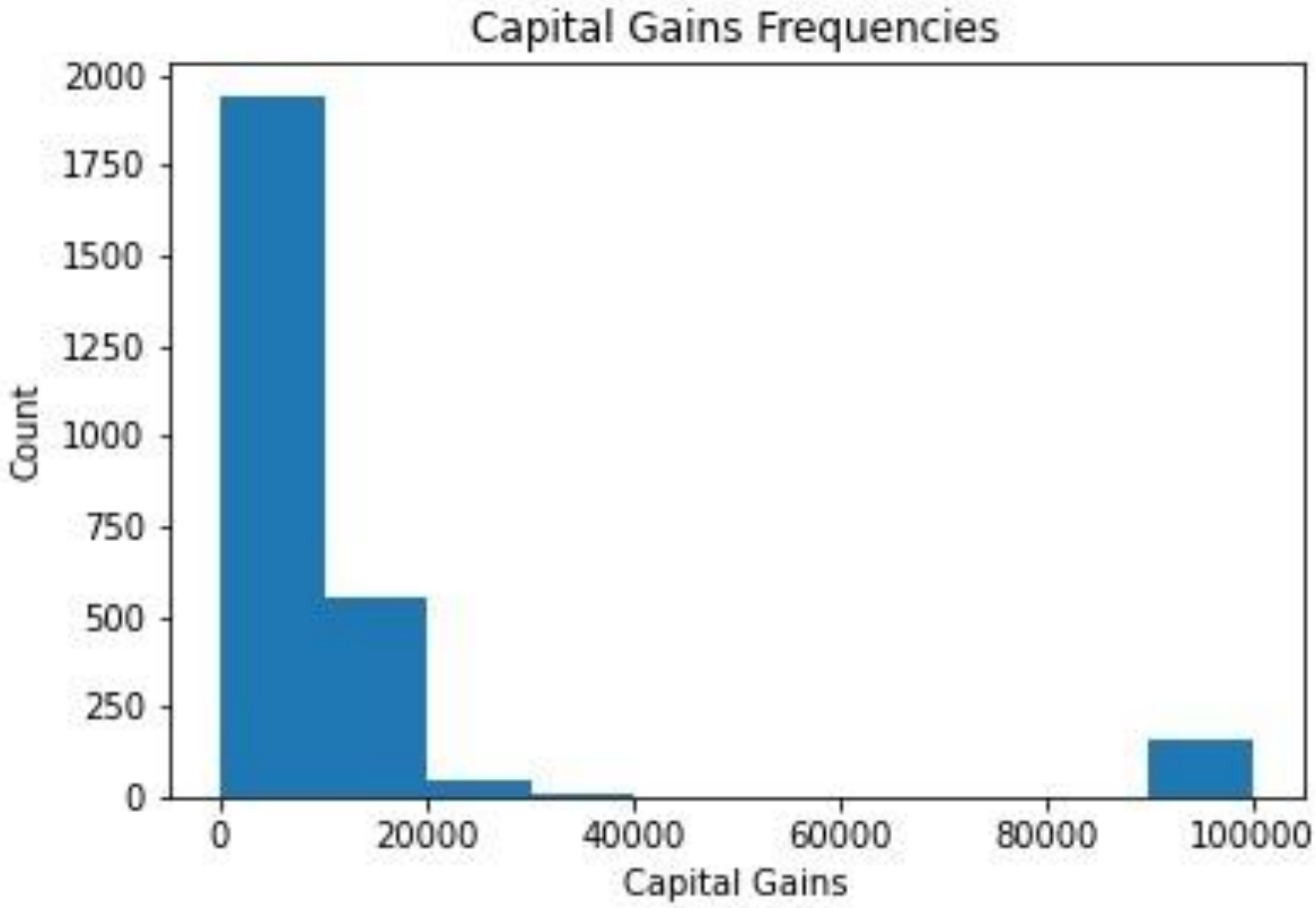
Dropped incomplete observations*

Dummified columns

Dropped ‘fnlwgt’*

Feature Engineering

- Split columns into manageable features
  - Age
  - Market Participation
  - Hours Worked
  - Immigrant
  - etc.


Capital Gains Frequencies
2000
1750
1500
1250
1000
750
500
250
0
Count
0
20000
40000
60000
80000
100000
Capital Gains

Baseline Accuracy → 75.9%

Modeling: Hyper Parameter Selection

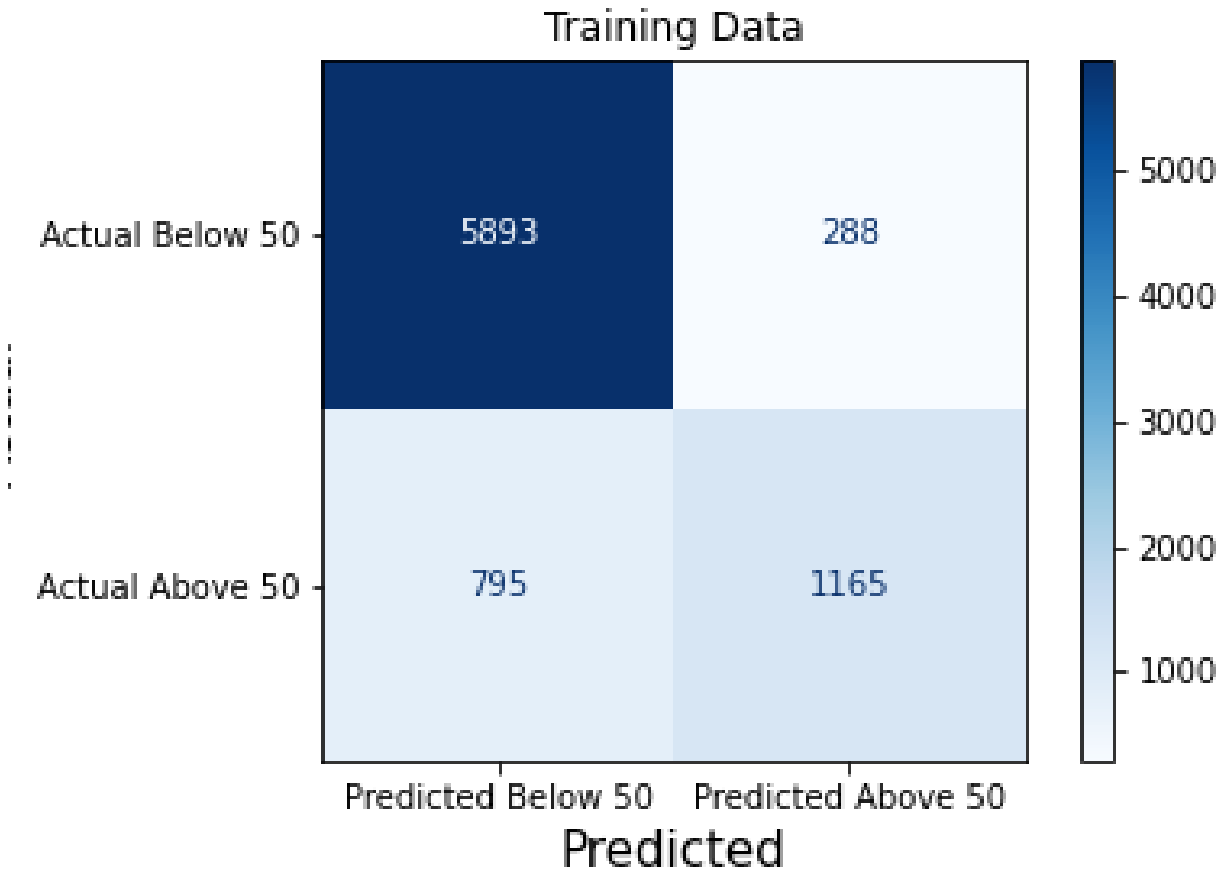
- Out of Bag Samples:
  - Send the sample that didn't match to the next iteration
- Max Features of 75:
  - Some features were dropped as not contributing to higher accuracy
- Warm Start:
  - Existing fitted attributes are used
- Entropy used to split:
  - Guages the disorder using a logarithm based metric
- Cost-Complexity Pruning Alpha of 0.001:
  - Helped to avoid initial overfitting
- Max Samples of 0.3:
  - 30% of the data as the max bootstrap sample size gave a small increase in performance

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Confusion Matrix

Most missclassification was of those above $50,000 who were predicted to be below $50,000.

These were mainly of classification married with a civilian and with capital gains below 200 (raw value).



Conclusions

- Random Forest can beat the baseline accuracy
- Less false positives, more false negatives
- Small improvements from feature engineering

Final score of 86.7%

Next Steps

- Feature engineering, or getting more data to fix the misclassification
- We know where we are falling short, and now we can address it.