Using Machine Learning to predict Income Classification

Training a Model to Allocate Stimulus Checkss

Business Problem

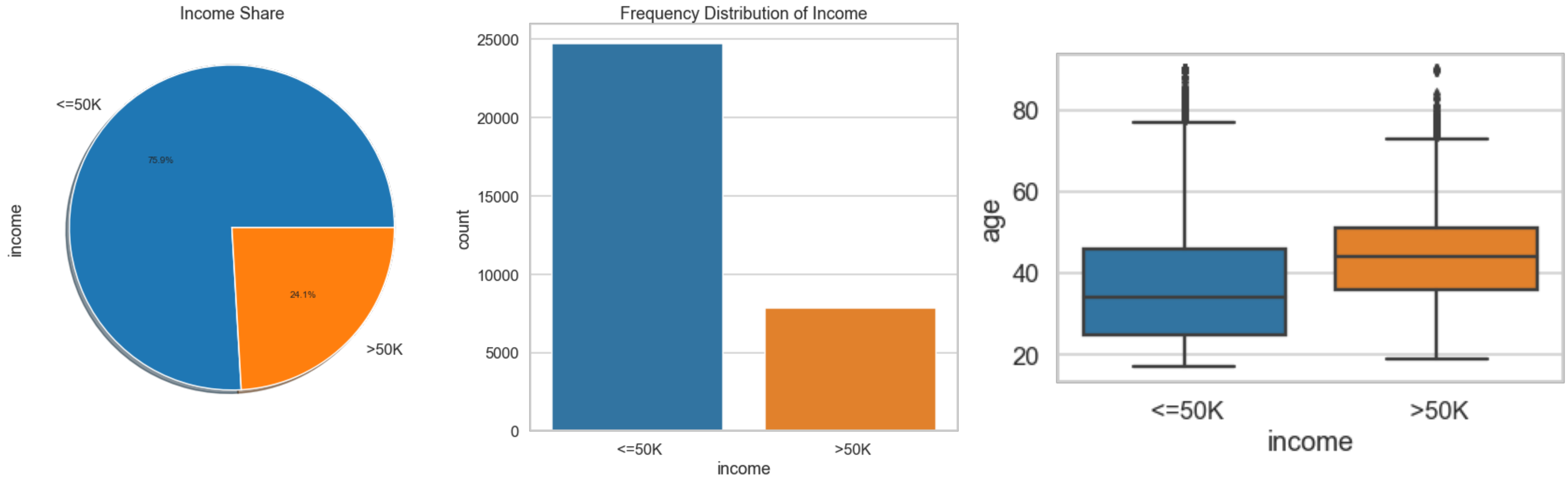
- Stimulus checks to address the bind COVID unemployment
- Machine learning algorithm to classify people who we should disperse stimulus under circumstances:
  - Taxes system backed up
  - Immediate job circumstances changes for many Americans
  - Use UCI data set to see if we can identify factors that ID income level so we can prioritize helping those in need
  - Minimize missing those in need (False positives → max precision)

Questions

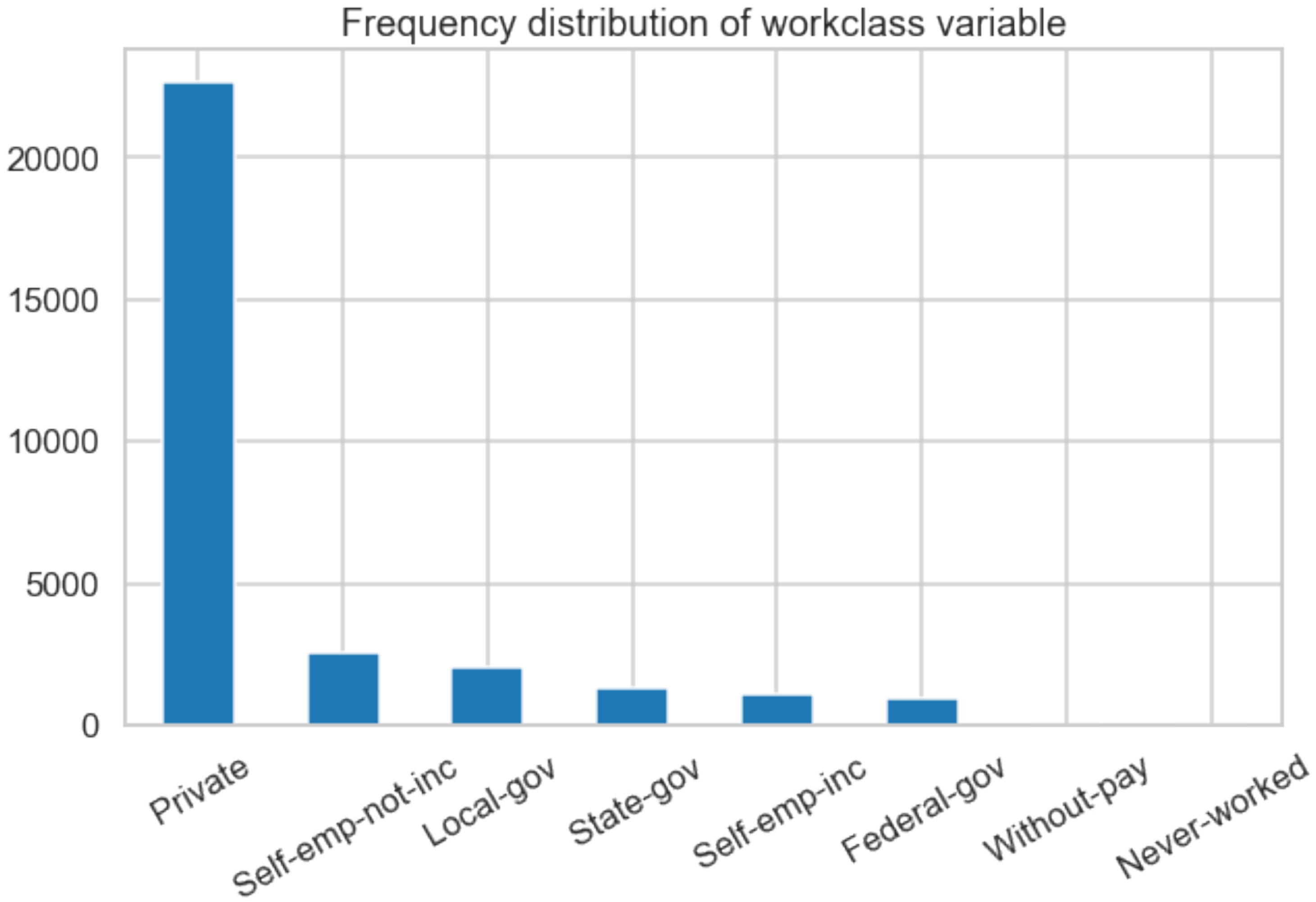
- What factors contribute most to telling us who makes above 50k, least at risk without a stimulus check? What types of people need the most support based on?
- What is the highest precision we can get to minimize the false positives, where we assume a family can make it without a check but can't?
- How do we minimize the overall cost / impact of the pandemic?
- Is there an industry focus we should have? What industries should we focus on?

Consequences of Getting this Wrong

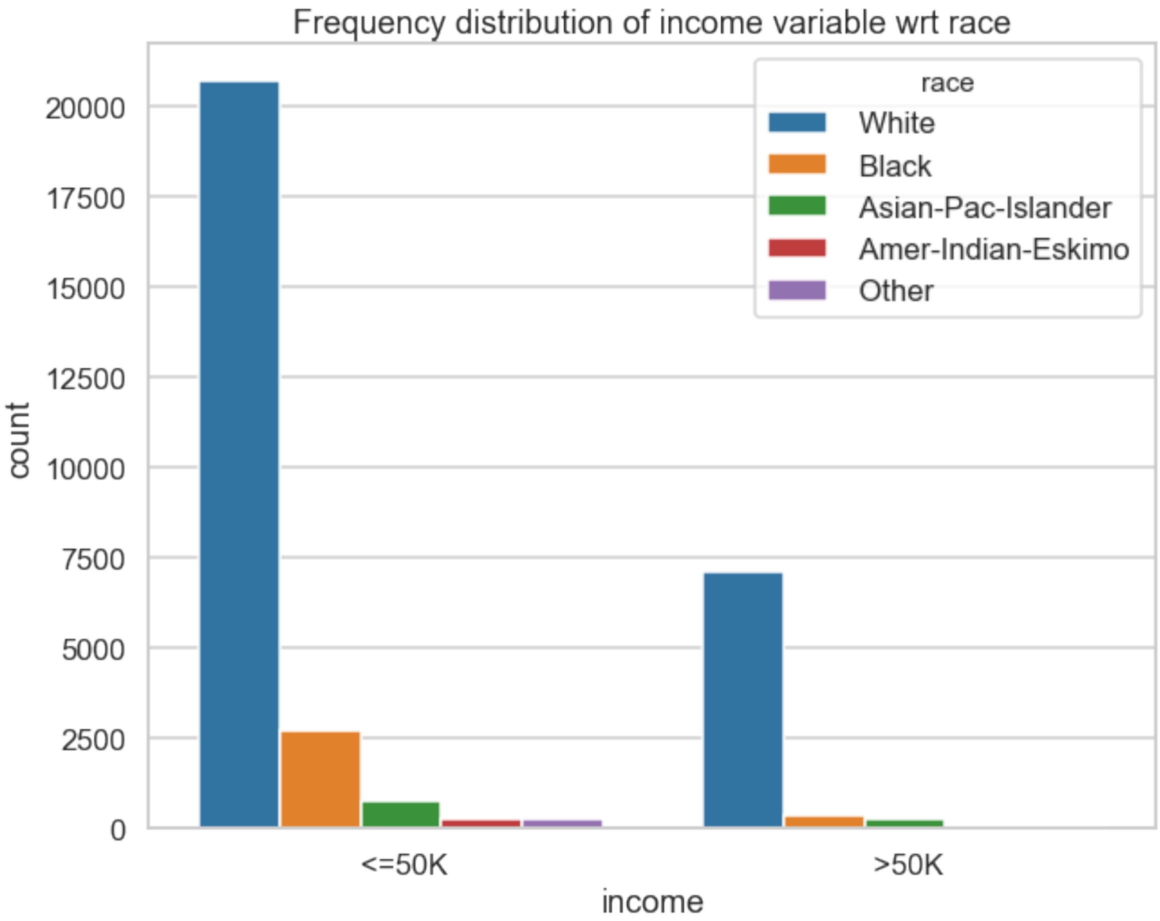
- False Negative → If we’re wrong about who makes under 50k
  - Send check [less resources to allocate to others and support future programs]
  - Citizens may use that check frivolously, but contribute back to the economy
- **False Positive** → If we’re wrong about who makes over over 50k [we’re wrong]
  - No check → puts citizens at risk of homeless (cost more), lost job
  - We have more resources to reallocate, but the time lost doesn’t help you pay your bills
  - Enact policies to alleviate people who we’ve missed

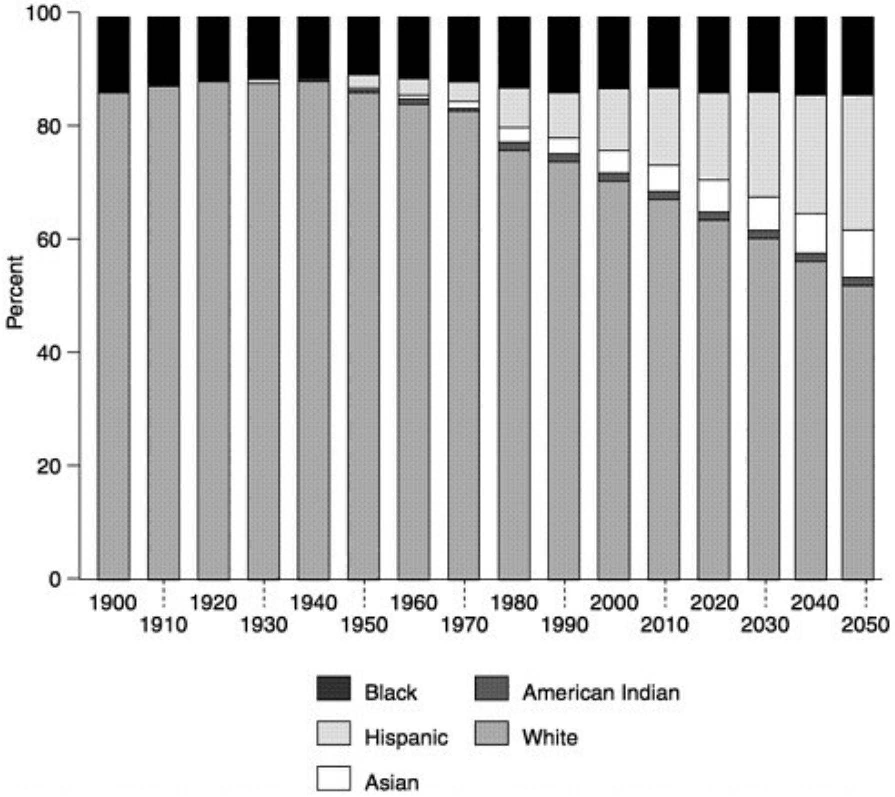


The percentage of the population in dire need of the check coincides with age



A focus on private industry gives the highest opportunity to address the 75% of the population that makes less than 50K





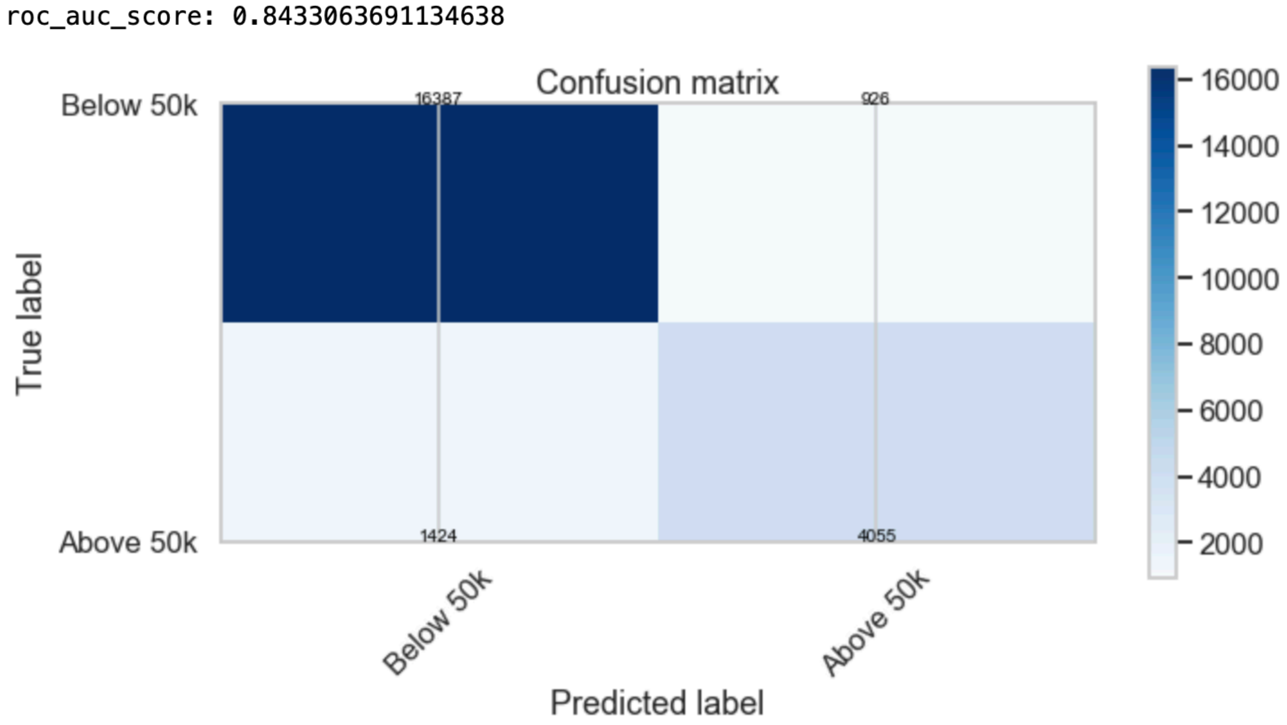
UCI Data Set this is based may not accurately represent our population today, but some of the characteristics it takes into account might still be relevant

Model Performance

Model	Precision	Accuracy	F1 Score
Random Forest (0.73)	0.99	0.98	0.97
KNN (0.85)	0.92	0.90	0.92
Decision Tree (0.63)	1.0	1.0	1.0

All Models listed can have variable cross-Validation rates x→ evaluator of overfit/underfitting of mode

- Can be improved with better feature engineering
- Wouldn't trust score < 0.7



```
roc_auc_score: 0.8433063691134638
```

Conclusions

- Industry or workplace was not a pivotal characteristic in predicting >50k according to model
  - Top 3 factors are AGE, MARRIED, EDUCATION
- Highest precision of our models thus far is 0.9
- Minimize overall cost by lowering the initial amount deployed, but iterating over multiple rounds to make sure people are taken care of

Future Recommendations

- Feed next rounds of data into model to train, collect information relevant to industries of relevant today → clearer
- Bin wage classes differently to assign different stimulus amounts based on wage classes
    - Alternate analysis can also be done with respect to wage class → different factors might distinguish those in the 150k and up class vs 75k-90k → zipcode could segment the data → identify risk of being able to weather pandemic (vaccine selection)
- Though not done in this study might be interesting to explore wage disparities by community / culture → if funds dispense at a state level, could prioritize support based on tiered approach [zipcode]

Thank you, what outstanding
Questions do you have?

- Takeaways: …confidence in people who receive it not needing it will feed the economy engine, in addition to those needing the support not becoming homeless which is inherently more expensive than the initial stimulus hit, and longer to recover from
- Not going to use race