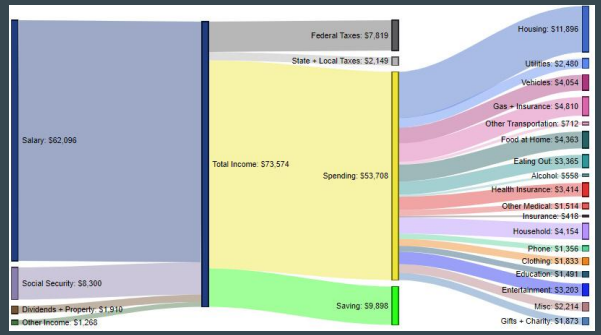
U.S. Census Insight: Predicting Adult Income

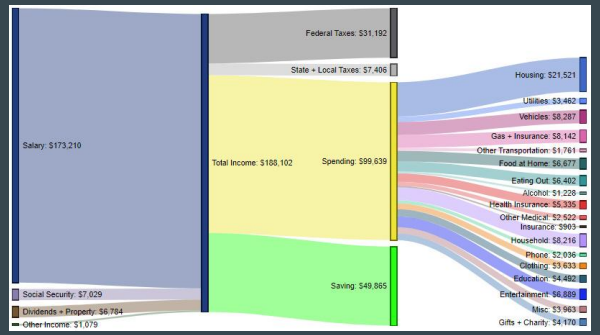
•••

February 19, 2021

Proposal

- Americans income-to-expense ratio varies by income level.
- What factors will allow us to accurately predict the annual income of adults in the US?
- If we can predict income, we can predict expenditure.

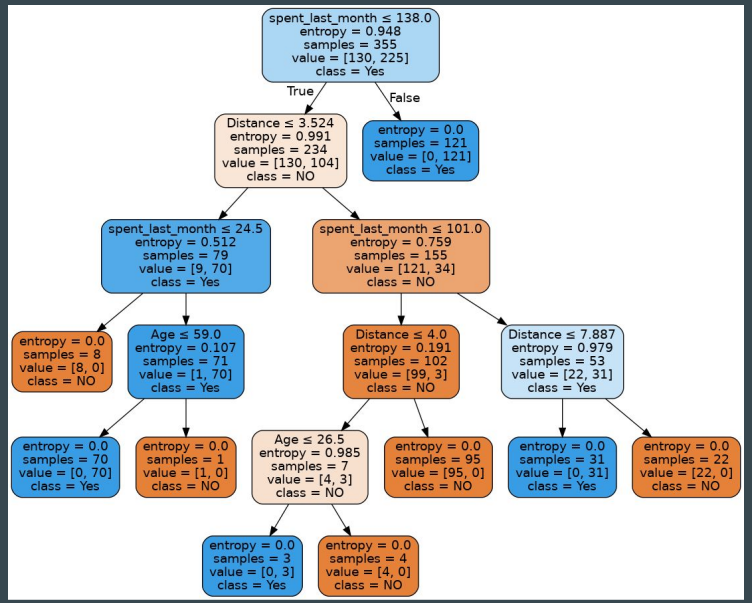




*Desjardins, https://www.visualcapitalist.com/how-americans-make-spend-money/

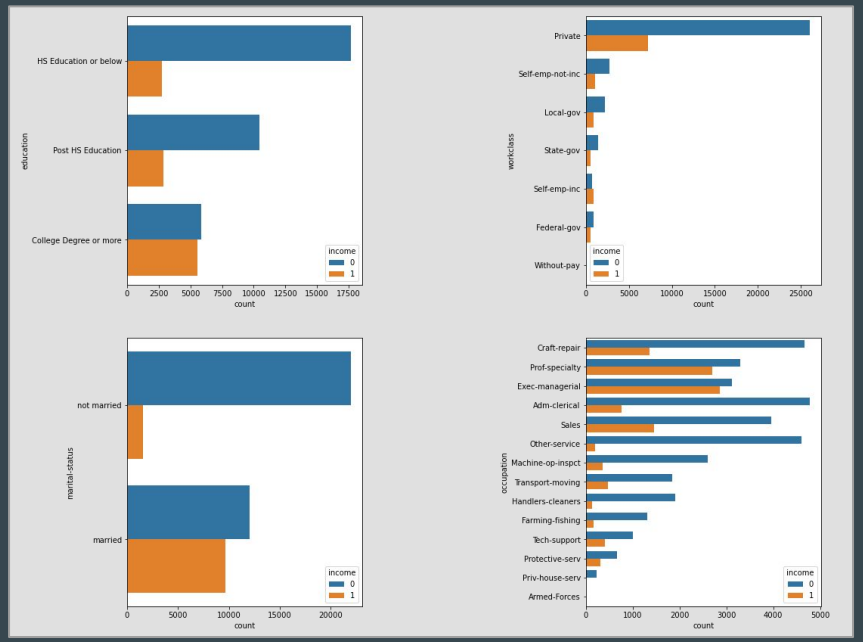
The Methodology

- Clean the data to be categorical in nature.
- Create a categorical model with a RandomForest algorithm.
- Train a model to correctly categorize the income level of adults.



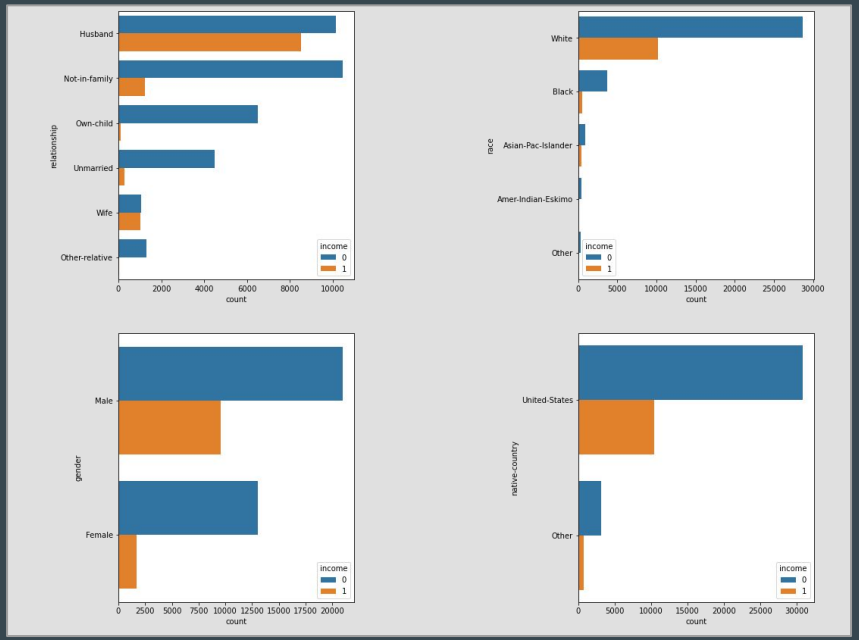
Cleaning

- Binary Income:
  - 0 assigned <= $50k/year,
    1 assigned >$50k/year
- Education cleaned into 3 categories
  - Little difference and representation in lower/higher education extremes
- Marital status simplified to binary
  - Little variation in income via technical status (ex: “widowed” vs. “never married” have similar income ratios)



Cleaning (cont.)

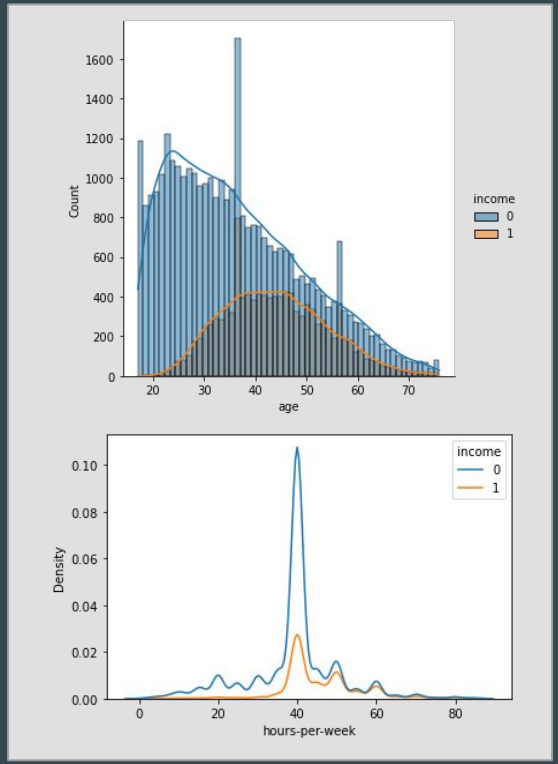
- Race & native country -
  Not enough data representation
  to be reliable in categorical model.
- Gender and relationship show
  strong trends in relation to income.
  - Relationship is in regards to others
    in the home.



Cleaning (cont.)

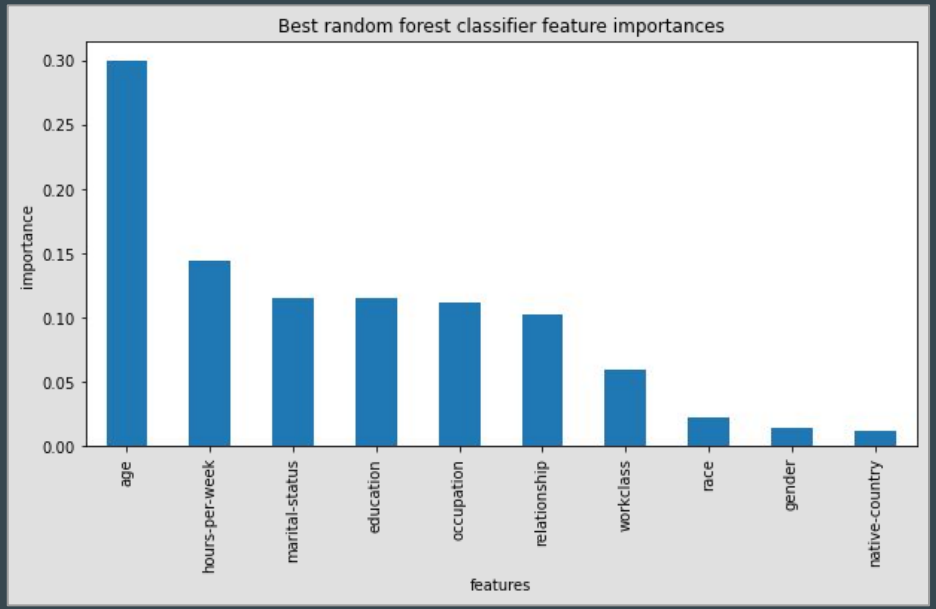
- Age shows strong correlation to income values; with a tail to the outlier ages of 70+.
    - Removed the 1% outliers

- Hours-per-week is focused in on an average of 40; as expected.
    - < 40 hours/week not conducive with making >$50k?



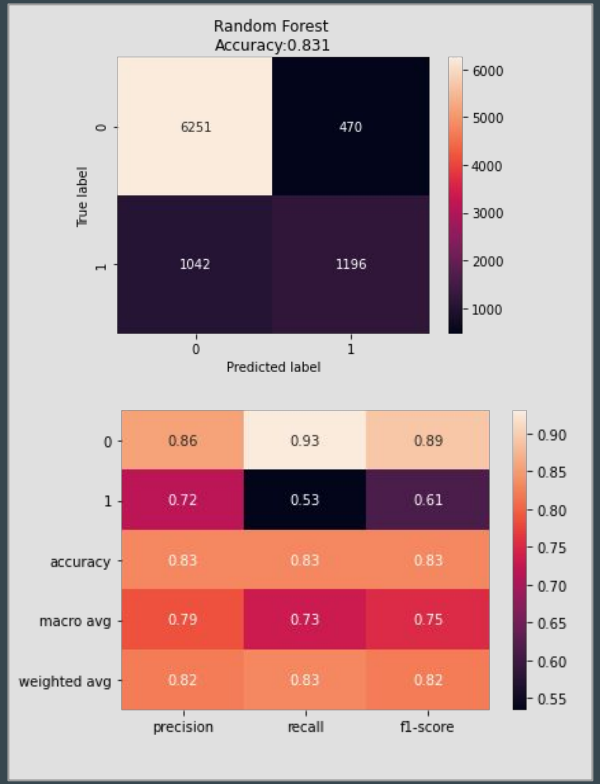
Feature Identification

- No feature has a higher score than 30% correlation to income.
- As predicted, race had little value in predicting income from this dataset.
- Top 5 features:
  - Age ~ 30%
  - Hours-per-week ~ 15%
  - Marital Status ~12%
  - Education ~ 12%
  - Occupation ~12%



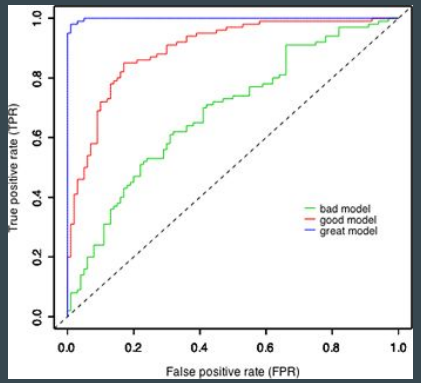
Results of Random Forest Model

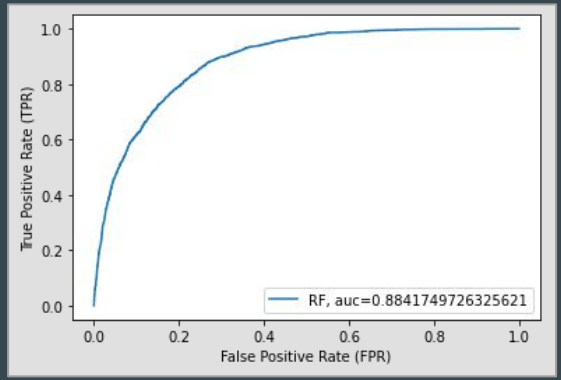
- Trained on 80% of data, tested on 20%.
  - Training accuracy: 85%
  - Testing Accuracy: 83.1%
- Recall or True Positive Rate
  - (TP/(TP+FN): 83%, better w/ <= $50k data
- Precision
  - (TP/(FP+TN): 82%, better w/ <= $50k data
- F1-score
  - (2*P*R/(P+R): 82%
  - It is a harmonic mean of precision and recall
- Accuracy
  - ((TP+TN)/(N+P)): Overall ~83%
  - Percentage of total items classified correctly



Results of Random Forest Model

- Great? Good? Bad?!
- ROC-AUC Score:
  - Likelihood of randomly choosing a positive case & negative case where the positive case outranks the negative case according to the classifier.
  - Our model's score: 88.4%; pretty good.





Conclusions

- We can accurately predict the income level of an adult in the US with an error of $< 12\%$.
  - Allows us to predict expenses, savings, and taxes.
  - Enables tailored marketing to demographics according to expenses.
- For the future:
  - Need more representative data of minorities & immigrants
  - Consider adding more income tiers, changing from binary classification to a new model.