Classification Prediction of Income

Author: Navnoor S. Kahlon

Source: kaggle.com

Stakeholder: A non-profit organization surveying individuals from various walks of life. They want to understand the population and see how can they serve them.

Problem:

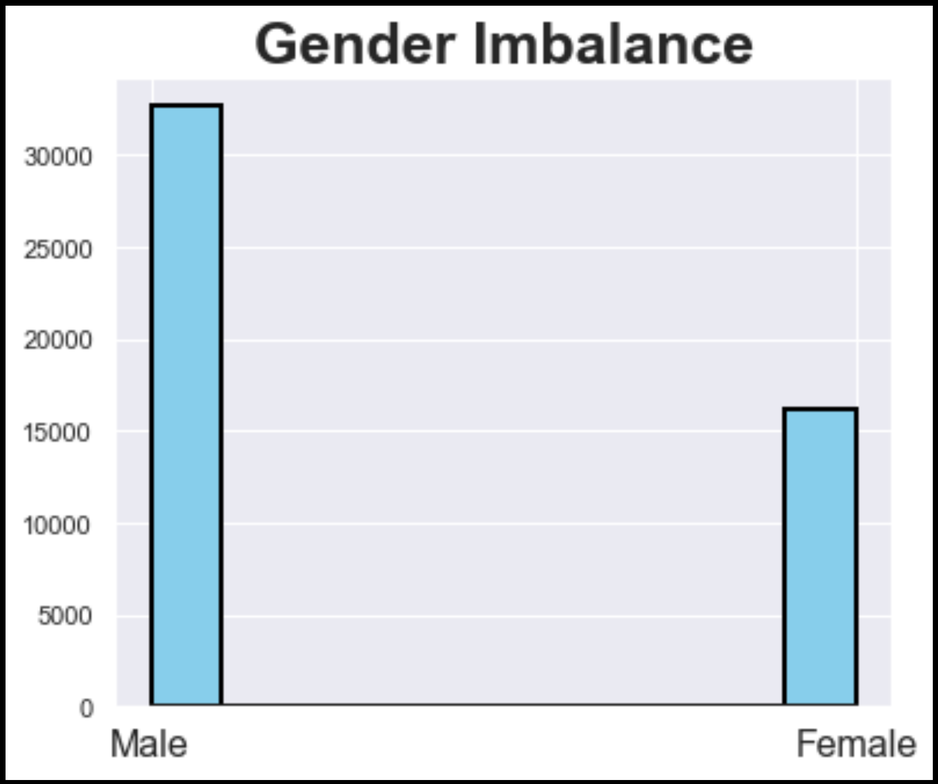
- The company wants to create a model that predicts what income bracket an individual falls into.
- Are there any patterns we can see among the income types?

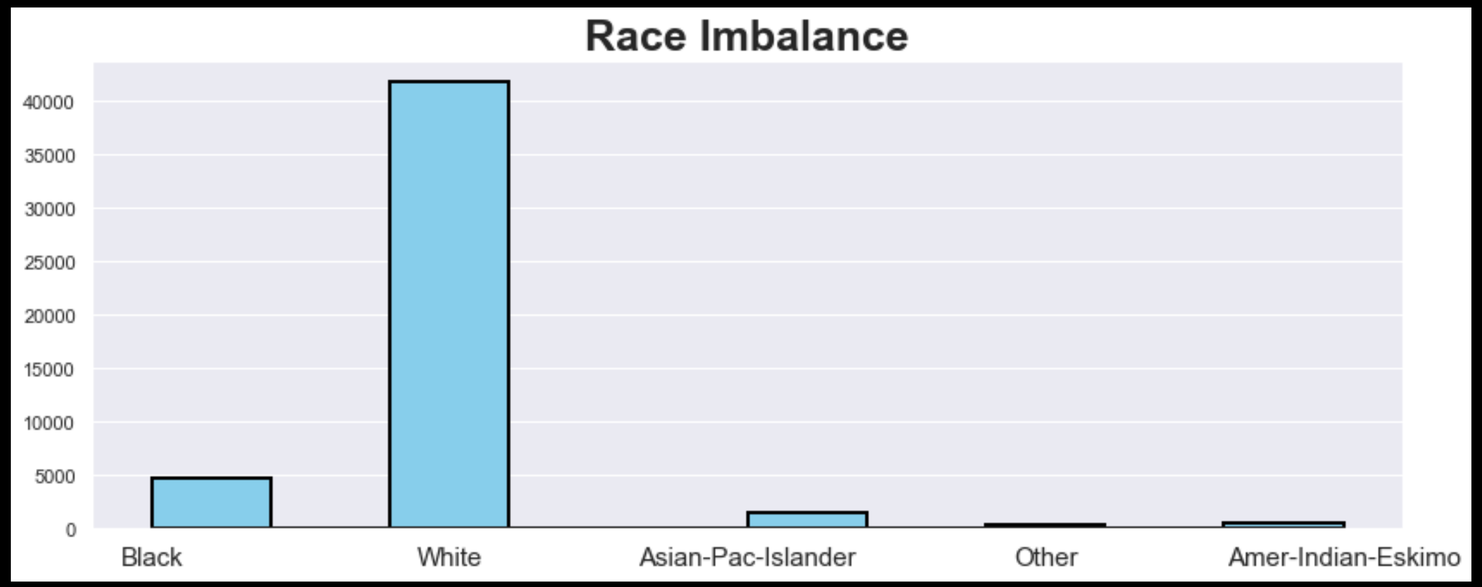
Brief intro to the Dataset

- Dataset covered multiple aspects of a person’s life.
- Such as race, education, workplace, marital status, capital gain and much more.
- Every race is represented to some point.
- Every types of job & every marital status is explored.
- Education levels spanned from 11th grade all the way up to Doctorate.
- Individuals from over 41 countries were surveyed.

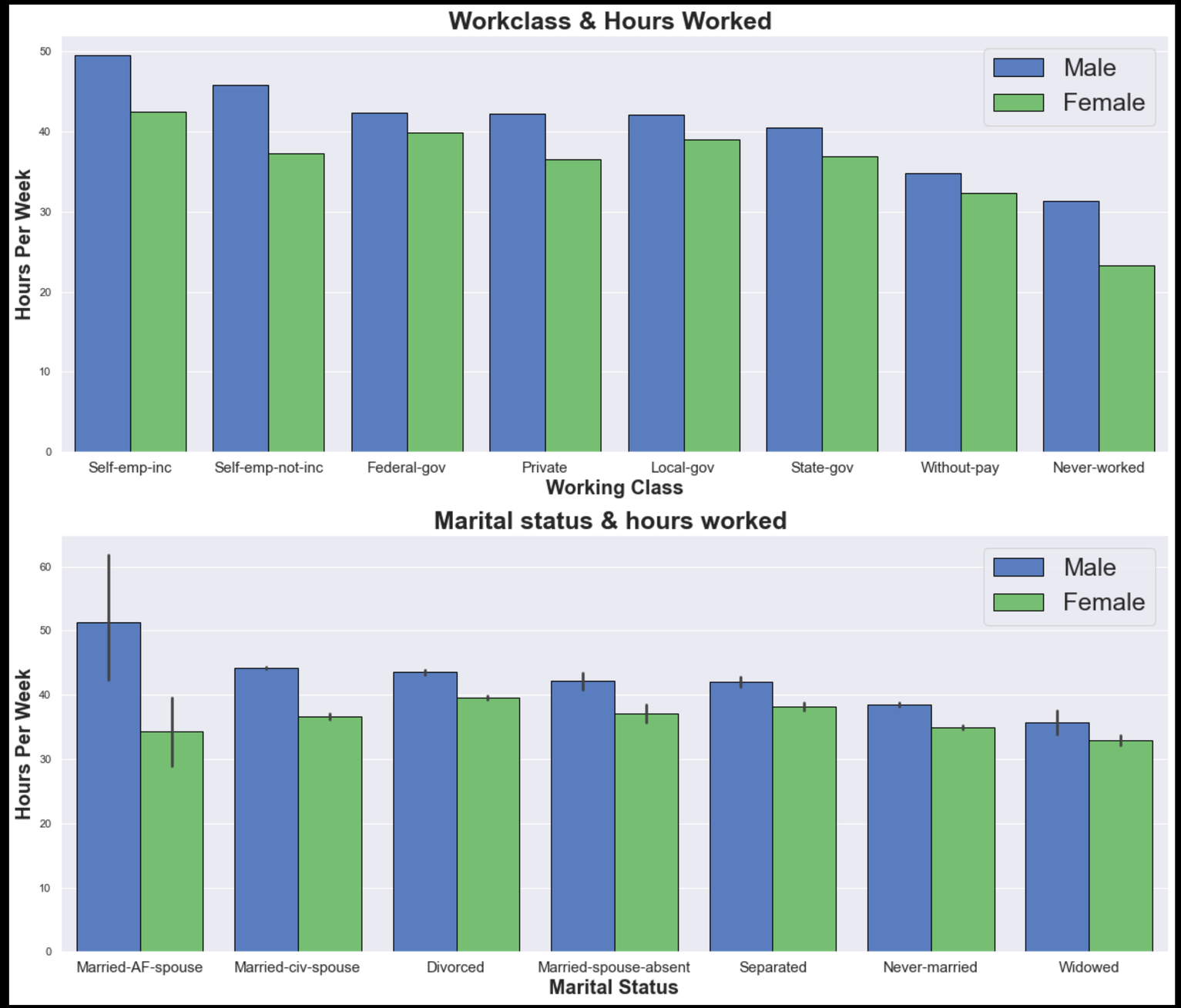
The Imbalances

- The White race is overwhelmingly represented in the dataset.
- There were half as many women in the dataset as men.
- Reason: Majority of individuals surveyed are from the U.S.A.





- Men made more money over women in every type of work place.
- Men in every martial status made more money than women as well.
- Usual Mistake: The survey does not count the number of hours an individual spends taking care of their house and family.



Strengths and limitations of the model

Strengths:

- The wide span of features that the dataset has, gives us the ability to make a strong model.
- When it comes to predicting the Below Middle class category the model is getting a score upwards of 94.
- The False Positive is at 0.064. Since majority of the population is in the Below Middle Class it is good to know that the model can predict them with a large percentage.

Limitation:

- Some of the unexplained features made it harder to use them to the best of there ability. Such as fnlwgt and educational-num.
- The model is not as strong in predicting individuals who make more than $50K a year.
- The False Negative has a score of 43 which is more than what we would like to see.

Final recommendations based on analysis.

- Need to understand why fewer women were surveyed, even in the western countries.
- The clustering of the data found that an individual who worked 50 hours a week instead of 40 hours had exponential financial gains over the others. A few individuals had immense capital gains and most had none. Hence, survey should only focus on those who the non-profit wants to serve.
- The survey needs more individuals surveyed from other countries than the US.