Adults Dataset Prediction Model

By

Mugagga Innocent

About the Dataset

Source:

- UCI Machine Learning Repository

Background:

- Extraction was done by Barry Becker from the 1994 Census database

Stakeholder:

- The Government

Data Scope

Target:

- Prediction task is to determine whether a person makes over 50K a year

Category:

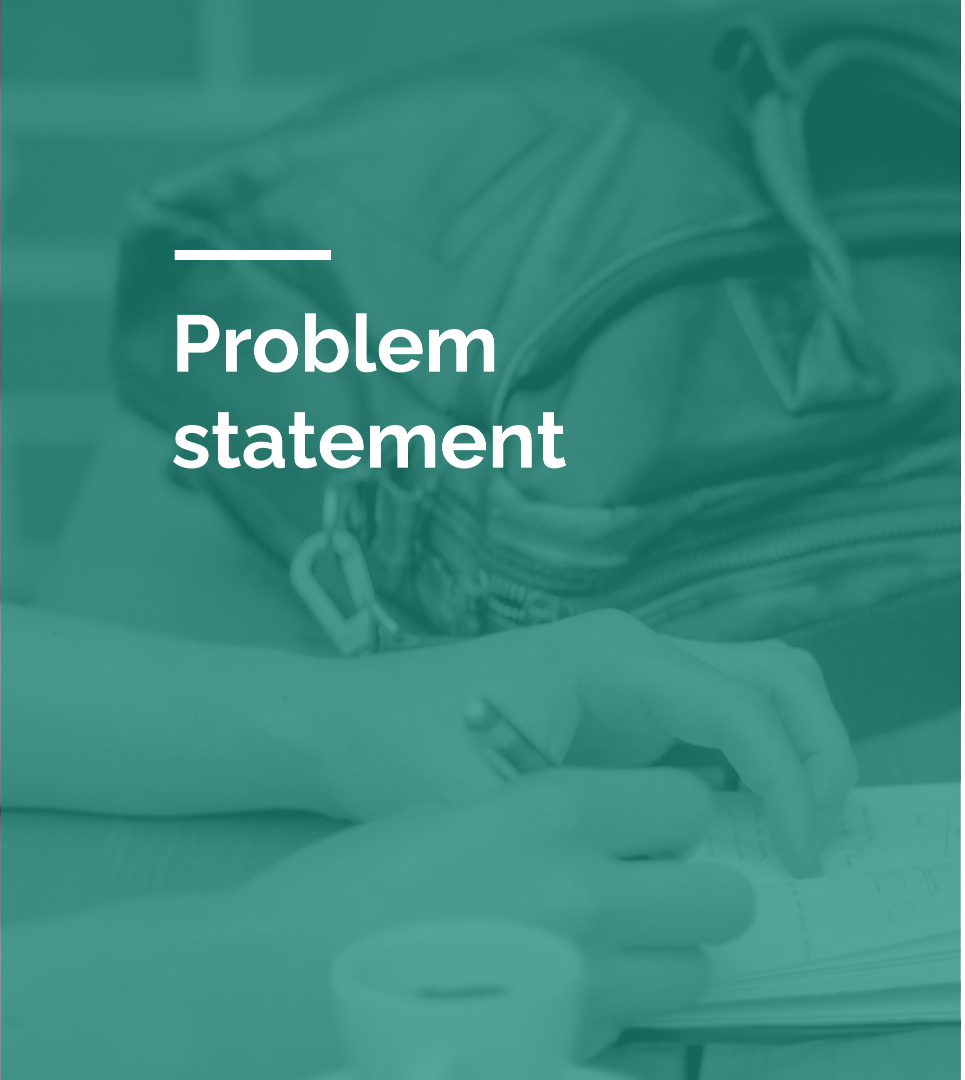
- Classification Problem

Features:

- 14 Features

Samplesize:

- 32561 Entries

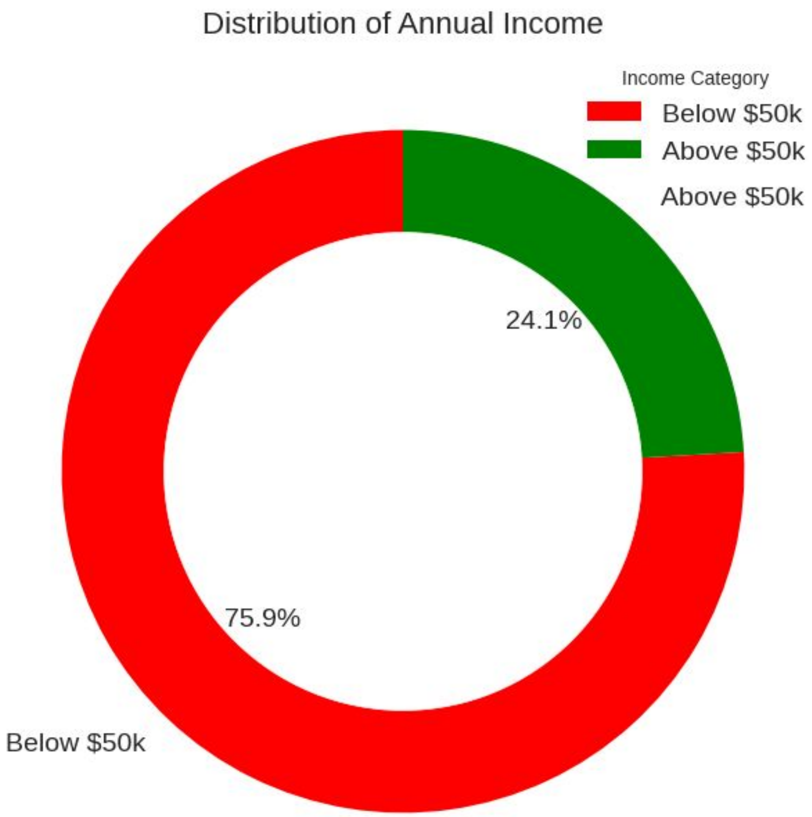


Problem statement

- To predict whether an individual's income exceeds $50,000 per year or not, based on demographic and employment-related information.

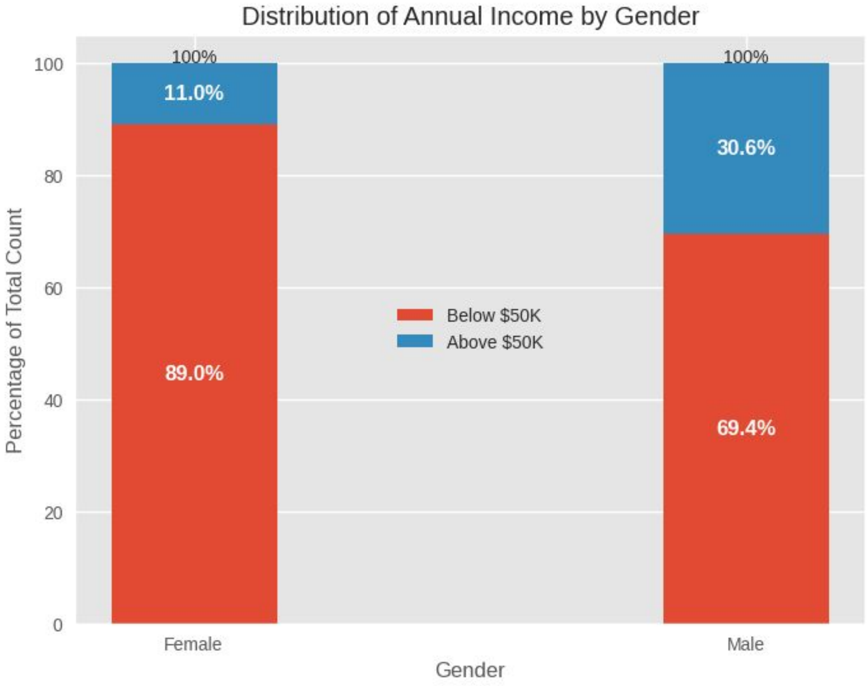
Distribution of Annual Income

- Imbalance of Incomes.
- 75.9% Percent of the population below the $50K Income Level



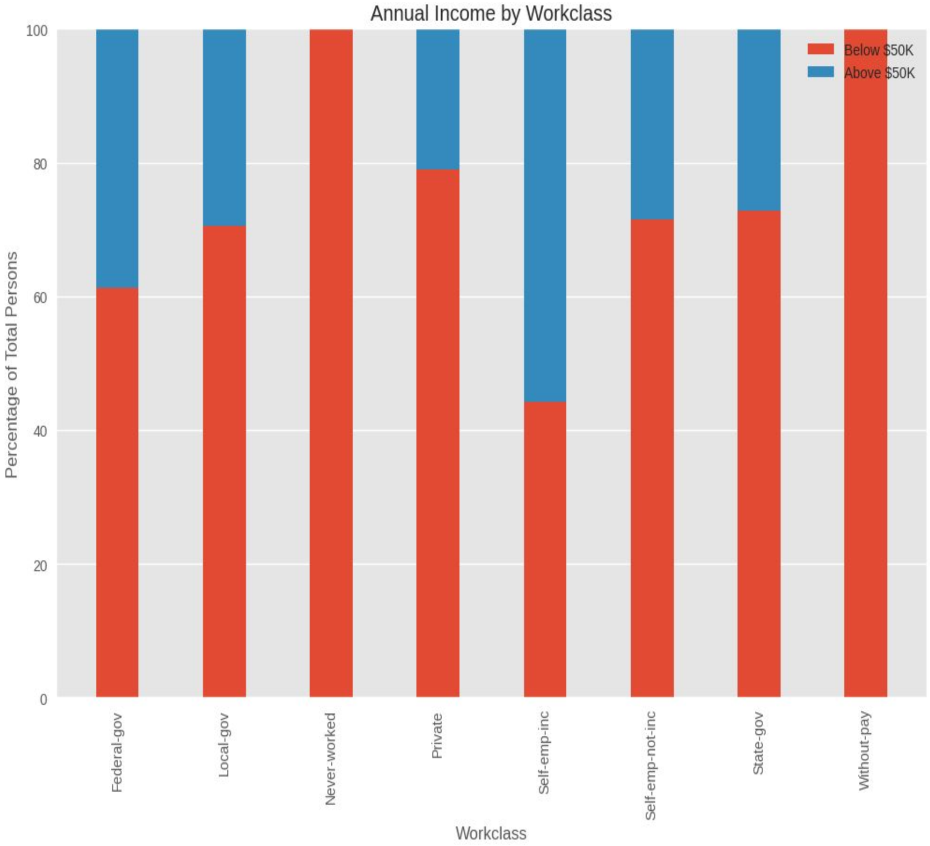
Distribution of Annual Income by Gender

- Males - 66.9%.
- Females - 33.1%
- More Females than Males are Below the baseline income.
- 89% of the Females are Below $50k Annual Income.
- The number of Males Below $50k are twice that of those Above $50k.
- In those Above_$50K, the Males are 19.6% Higher than The Females



Distribution of Annual Income by Work Class

- Only persons in the Self-emp-inc Workclass are Above_$50K to greater percentage as compared to those Below_$50K.
- All other work classes have more persons with the annual Income Below_$50K more than those with the Annual Income Above_$50K.



Model Strengths and Metrics

Precision:

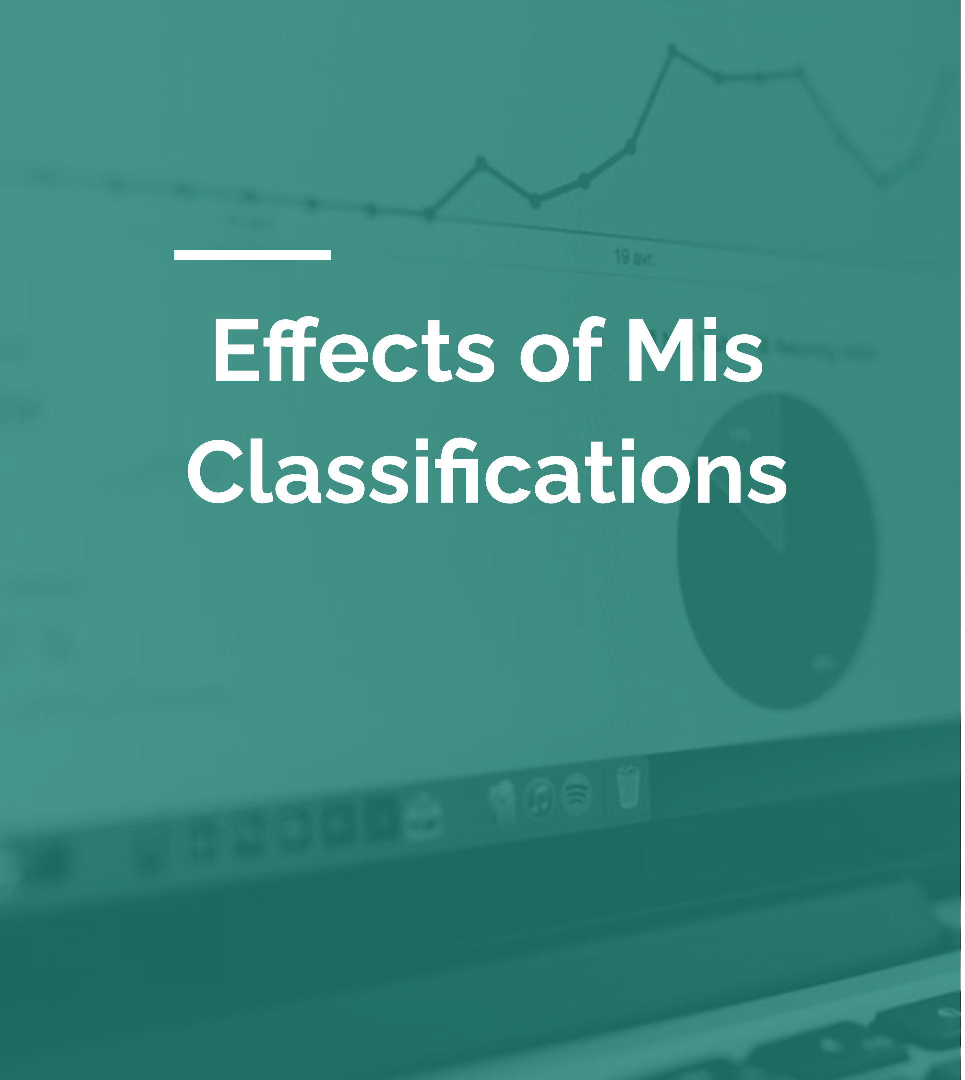
- Weighted Avg: 85.3%

Recall:

- Weighted Avg: 85.9%

F1 Score:

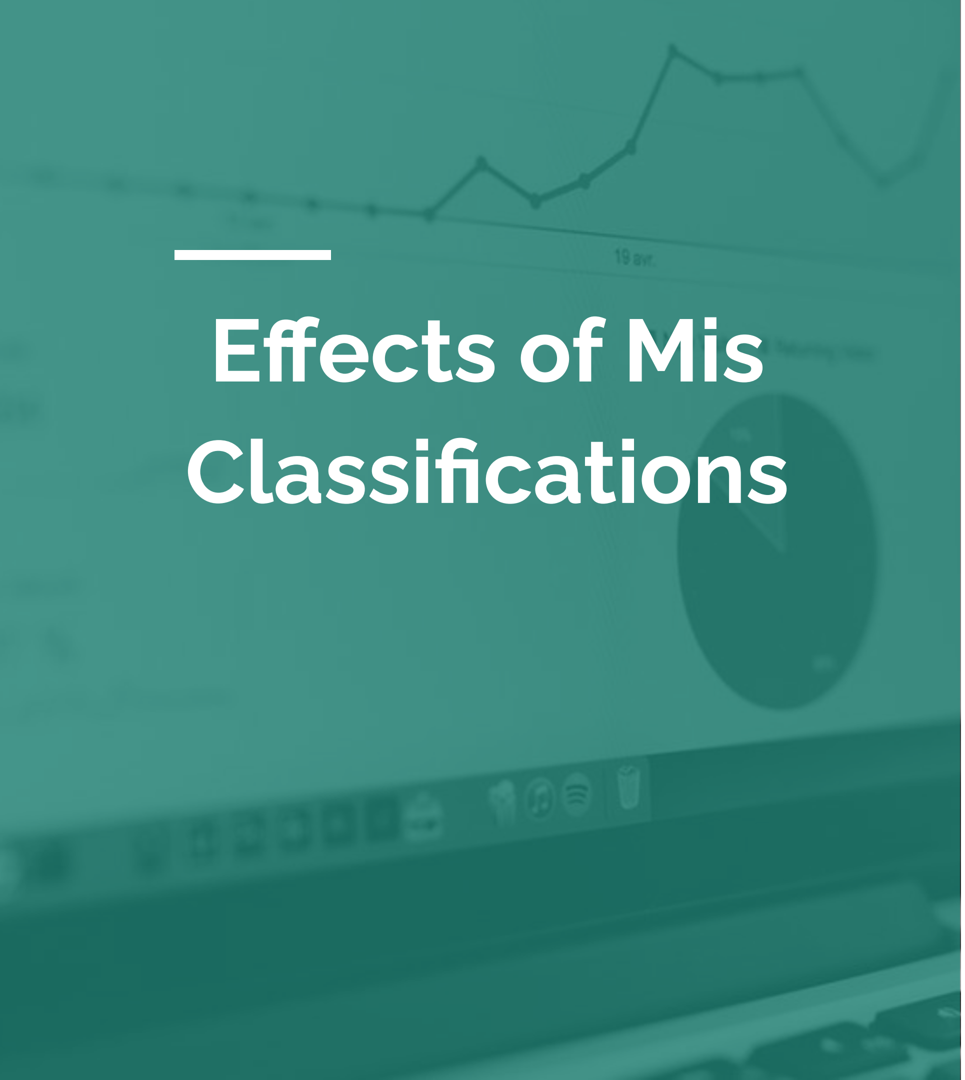
- Model's precision and recall for a particular class summarized
- Weighted F1-Score of 85.3% achieved
- High F1 score indicates high precision and high recall
- Model is effective in accurately classifying new data.
- New data can be classified identified while avoiding false positives



Effects of Mis Classifications

False Positives:

- These occur when persons who are Classified by the Model as having income Above $50K but in reality they are Below $50K.
- This can lead to these individuals missing out on government relief/aid programs intended for those in need.
- This could result in dire consequences for their well-being.



Effects of Mis Classifications

False Negatives:

- These occur when individuals are classified as having income below $50K by the model, but in reality, they earn above $50K.
- This can result in these individuals being subjected to receiving national aid/relief, which they do not require, leading to the misallocation of resources.
- Therefore, it's crucial to adjust the model's parameters to ensure that such misclassifications are minimized.
- Thereby increasing its accuracy and reducing wastage of government resources.

Model Recommendations

Model Performance:

- Inaccurate classifications can result in two types of errors.
- These can have significant consequences for individuals and the government as indicated above.
- To avoid such misclassifications, it's essential to fine-tune the model's parameters to increase its precision.
- It's also crucial to adjust the model's parameters to ensure that such misclassifications are minimized, thereby increasing its accuracy and reducing wastage of government resources.