Predicting Wage Level from U.S. Census Data using Random Forest Machine Learning Model

Presented by

Team Algorithm:

Jessie Owens, Alaa Senjab, Eddie Reed

October 21, 2019

PROBLEM

Bank of America’s marketing department wants to target potential customers who earn over $50,000 annually with specific loan products using U.S Census data.

Data Collection and EDA

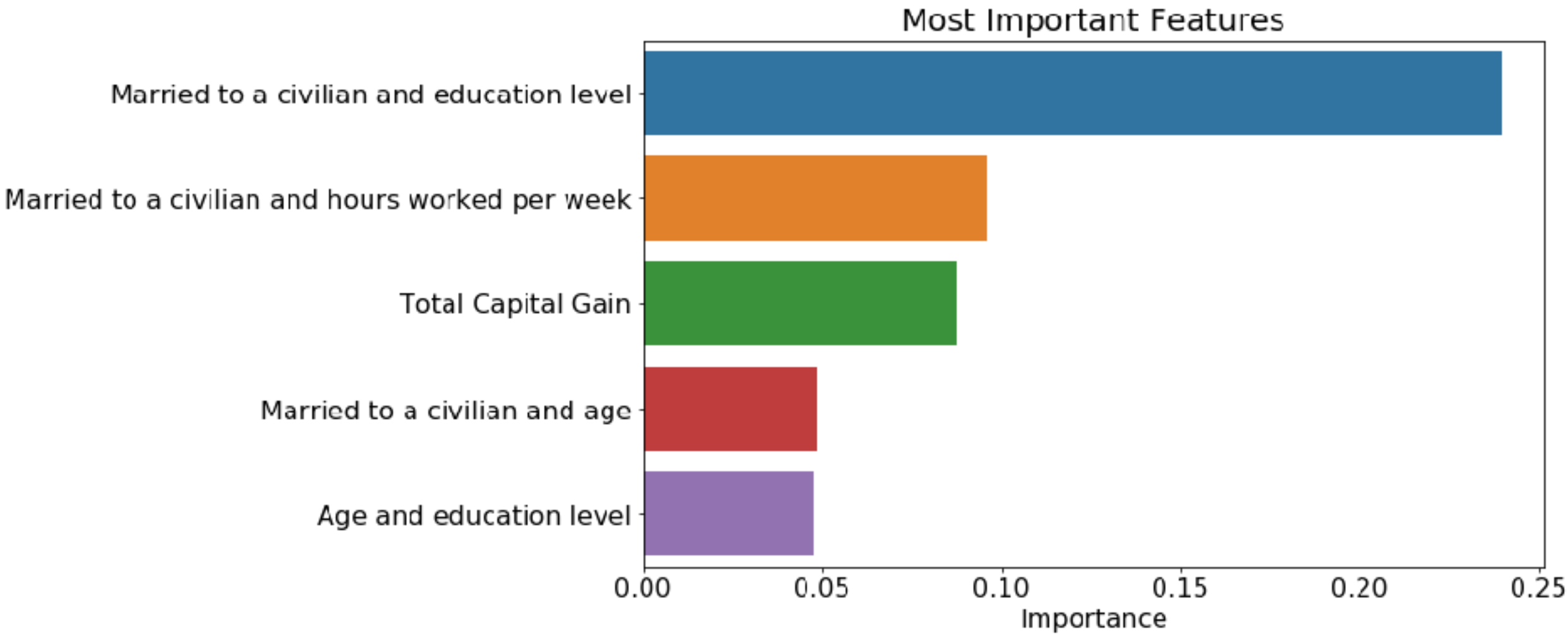
- Imported U.S. Census data
- Checked for missing values
- Imputed missing categories
- Stripped blank spaces in data

Preprocessing and Feature Engineering

Created dummy variables for all categorical features

Determined the top 5 features based on feature importance

Created feature interactions using Polynomial features on the top 5 features


Most Important Features
Married to a civilian and education level
Married to a civilian and hours worked per week
Total Capital Gain
Married to a civilian and age
Age and education level
0.00
0.05
0.10
0.15
0.20
0.25
Importance

Model Evaluation

Model used: Random Forest

Hyperparameters:
- max depth = 11
- max features = 60
- min samples leaf = 4
- n estimators = 30

Scores:
- Baseline: 75%
- Train: 88%
- Test: 86%
- Specificity: 94%

Conclusion

Random forest is a reliable and fast model that performed well in predicting an individual's wages.

Key Takeaways:

- Optimized for specificity with a score of 94%
- Demographic information introduces bias
- Factors affecting wage that are intuitive (age, education level)
- Factors affecting wage that are not intuitive (married to a civilian)