Techniques in Predicting Income

Jan C. Bierowiec

Pre-Processing

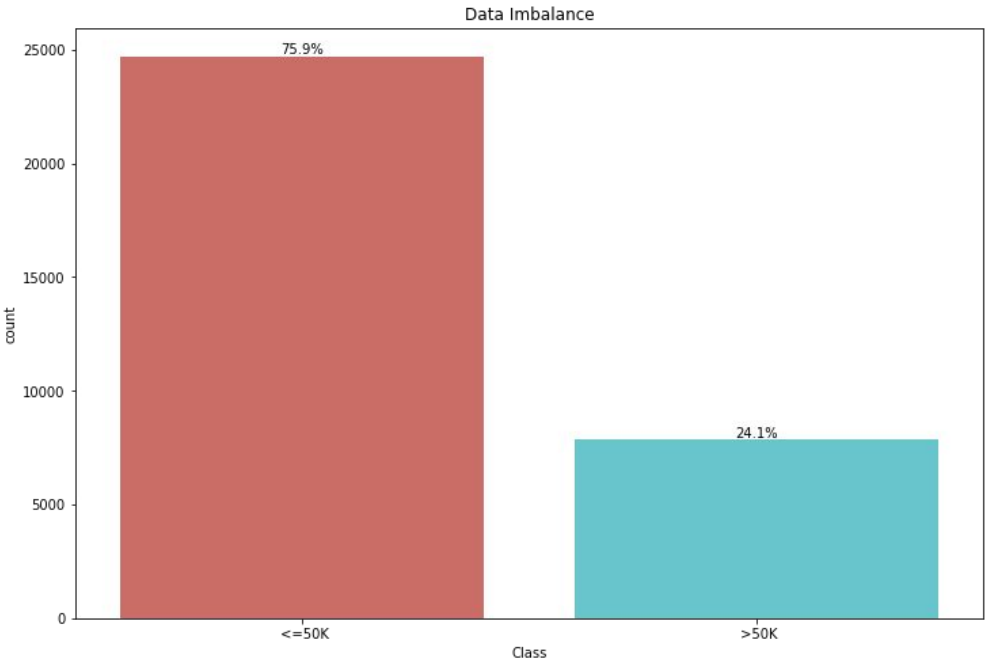
- Encoding
- Imbalanced Dataset
- Missing Values

Encoding

- 14 variables
  - 6 continuous - age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week
  - 8 discontinuous - workclass, education, marital status, occupation, relationship, race, sex, native country
- Data encoded using dummy encoding
  - k-1 binary variables

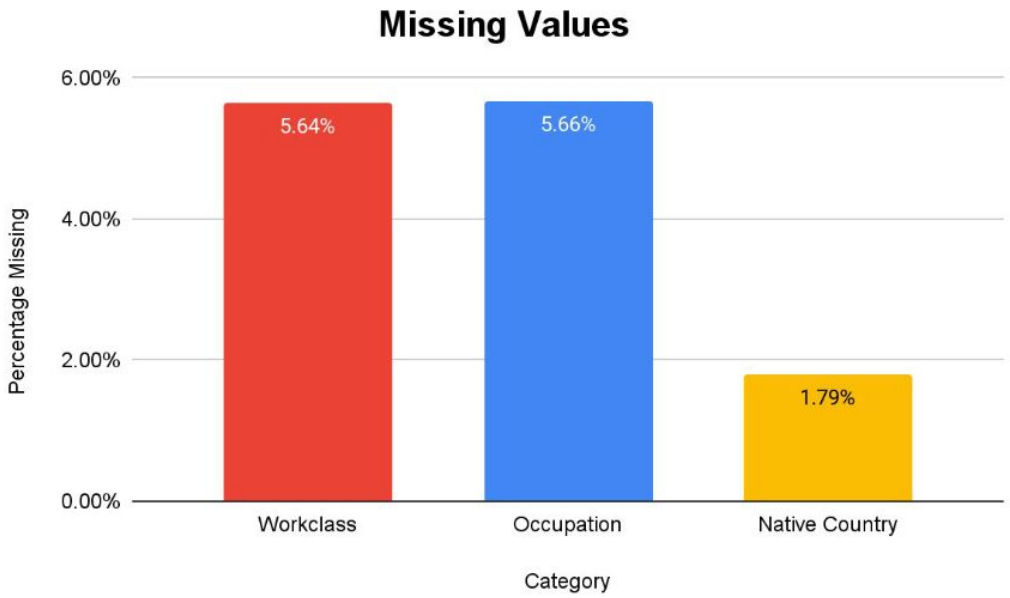
Imbalanced Dataset

- 75.9% <=50,000
- 24.1% >50,000
- Bagging method
    - Select random variables in training set and replace them with either 0, 1, >1 copies



Missing Values

- Three categories with missing values:
  - Workclass
  - Occupation
  - Native-Country
- Two methods chosen and compared:
  - KNN based imputation
  - Random Forest based imputation



Algorithm

- Ensemble
- Imputation Results
- Normalization Results

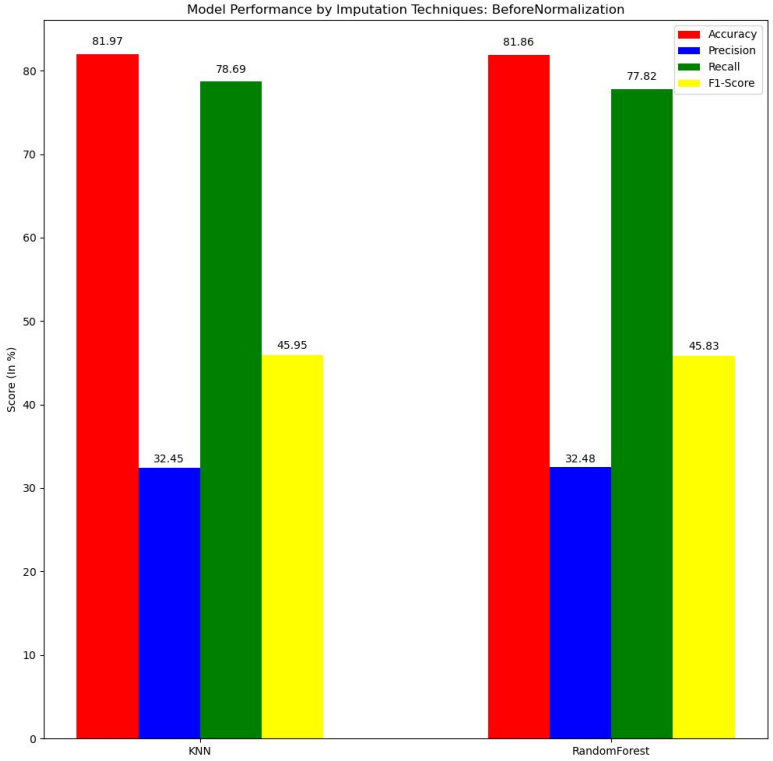
Ensemble Learning

- Ensemble classifier built from three supervised learning models:
  - KNN
  - Naive Bayes
  - Random Forest
- Ensemble learning is less biased towards any one individual learning algorithm.
- Z-score normalization results compared with non-normalization results

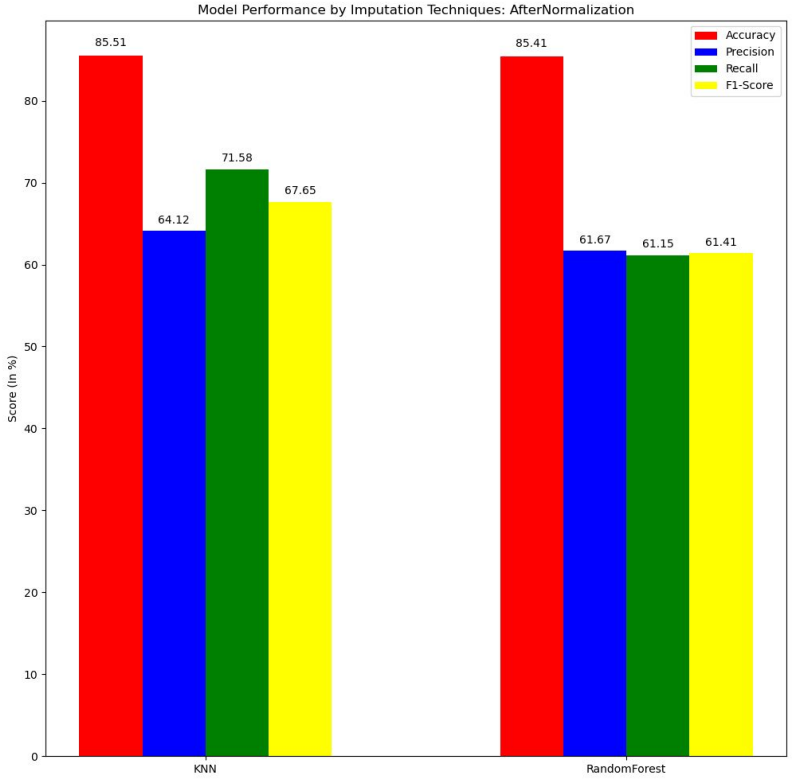
Imputation

- Model Based Approach :- Decrease variance , Decrease bias
  - KNN
    - Grower’s similarity coefficient to find the K nearest neighbor
    - Aggregation of the K values to impute the missing values
    - Aggregation is based on most frequent occurrences.
  - Random Forest
    - Build random forest for each individual missing value feature
    - Generalize ensembles of D-trees through bagging
    - Performs iterations to reduce *Out-Of-Bag* error

Imputation Methods Without Normalization



Imputation Methods with Normalization



Results Comparison

Method	Accuracy	Precision	Recall	F1-Score
KNN	81.97%	32.45%	78.69%	45.95%
Random Forest	81.86%	32.48%	77.82%	45.83%
Normalized KNN	85.51%	64.12%	71.58%	67.65%
Normalized Random Forest	85.41%	61.67%	61.15%	61.41%

Future Steps

- Parameter Optimization
  - Are there “best” parameters for the algorithms in each ensemble classifier for each data set
- Feature Selection
  - Determine if removing attributes improves predictive capabilities of learning algorithm
  - Try the Wrapper Method of Feature Selection
- Find out if there is a more optimal algorithm
- Try implementing a Neural Network

Sources

- Pantanowitz, Adam, and Tshilidzi Marwala. "Evaluating the impact of missing data imputation through the use of the random forest algorithm." *arXiv preprint arXiv:0812.2412* (2008).
- Ganganwar, Vaishali. "An overview of classification algorithms for imbalanced datasets." *International Journal of Emerging Technology and Advanced Engineering* 2.4 (2012): 42-47.
- Stekhoven, Daniel J. "Using the missForest package." *R package* (2011): 1-11.
- Kowarik, Alexander, and Matthias Templ. "Imputation with the R Package VIM." *Journal of statistical software* 74 (2016): 1-16.
- https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb