DS-UA 202, Responsible Data Science, Spring 2022 Course Project: Nutritional Labels for Automated Decision Systems

Elaine Shan, ys3719

Grace Yang, gy654

Background information

Credit Scoring algorithm

- make predictions about loan applicants' probability of default
  - the probability that someone will experience financial distress in the following two years.

- Stakeholders:
  - banks
  - borrowers

Trade offs

- predictive accuracy vs. fairness to subgroups
- Sensitive data protection vs. accuracy
- Transparency vs. its potentiality for exploitation
- Fitting to the static data vs. the dynamic nature of people and economy
- Predicting 2-year financial distress vs. predicting long-term financial distress

Input and output

Input

- Training vs testing:

  Training dataset contains 150,000 entries

  Testing dataset contains 101503 entries

- Predictors:Array of input features manifesting the person’s credit history and financial capabilities.

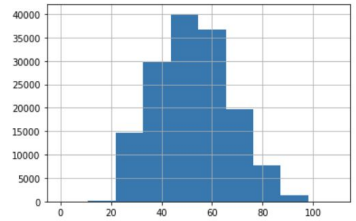
- Label: Binary Label indicating if the person experienced 2-years of serious delinquency.

Predictors

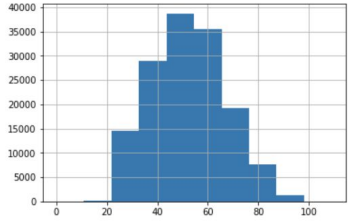
- 1. The total balance on credit cards and personal lines of credit real estate and no installment debt like car loans divided by the sum of credit limits
- 2. Age of borrower in years
- 3. The number of times borrower has been 30-59 days past due but no worse in the last 2 years
- 4. Monthly debt payments, alimony, and living costs divided by monthly gross income
- 5. Monthly income
- 6. The number of open loans and lines of credit
- 7. The number of times a borrower has been 90 days or more past due
- 8. The number of mortgage and real estate loans including home equity lines of credit
- 9. The number of times borrower has been 60-89 days past due but no worse in the last 2 years
- 10. The number of dependents in the family excluding themselves

data profiling

- Distribution of person experienced 90 days past due delinquency or worse
- distribution of age in the original dataset vs the distribution of age in the post-processing dataset
- Correlation heatmap between 11 features.

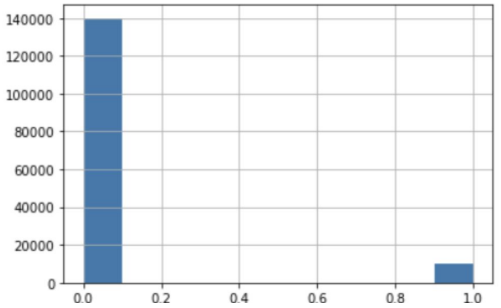


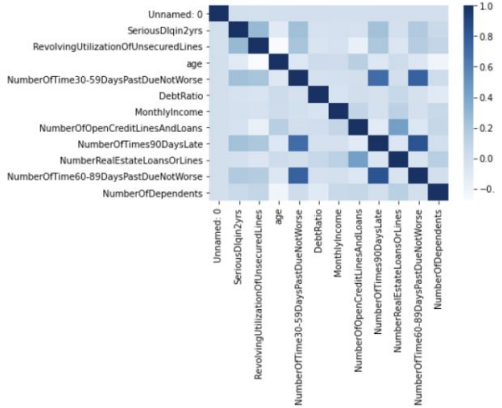
The distribution of age in the original dataset



The distribution of age in the pro-processed dataset

0 139974
1 10026





Output

- Binary label SeriousDlqin2yrs,
    - indicates whether a person is predicted to experience 90 days past due delinquency or worse based on the provided predictors
- Split the training data into training data and test data and ran the simple random forest model
    - 42904 people receive a label of 0
        - meaning they will not experience 90 days past due delinquency or worse
    - 903 receives a label of 1
        - meaning they will experience 90 days past due delinquency of worce
    - accuracy of 0.93 and an AUC of 0.58

Implementation and validation

Implementation

- Preprocessing
  - The ADS produces four datasets according to the different modes of modification mentioned above.
- Run various scikit models on various datasets
  - The best model is the random forest with a depth of 9 and 16 estimator
  - When we use the data set that removes the RUUL outliers, it gives us an AUC of 0.8662.

Validation

- k-fold cross-validation on the training dataset
- Compare the AUC of different models
  - one with the highest AUC score has best performance in meeting its stated goal
- A glimpse of the performance of the ADS on a subset of training data
  - Accuracy: 0.9377496747095213
  - AUC: 0.5865963364026058

Outcome

ADS performance

- A glimpse of the performance of the model:
  Accuracy: 0.937749674709521 3
  AUC: 0.586596336402605 8

- Accuracy is insufficient. We will analyze the accuracy of the ADS by comparing its performance on young people vs old people (cutoff age = 55).
  - unprivileged class: age<55 vs. the privileged class: age>= 55

Fairness measures

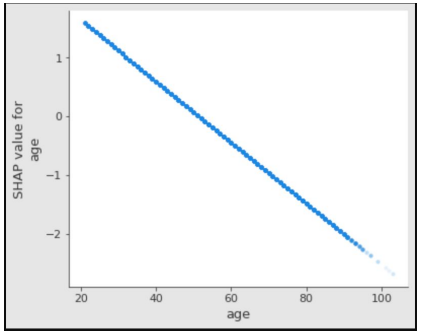
- with age<55 as unprivileged, age>= 55 as the privileged class
- We calculate **mean_difference** and **disparate_impact**
- **disparate impact**
  - The original training dataset has a disparate impact of 0.945, and the predicted training dataset has a slightly improved disparate impact of 0.977.
- **mean_difference**
  - The original training dataset has a disparate impact of -0.053, and the predicted training dataset has a slightly improved disparate impact of -0.023
- model mitigate bias slightly
- Compare the true positive rate and false positive rate by subgroups.
- The similar false-positive
  - comparable probability of being marked as having a low risk of delinquency.

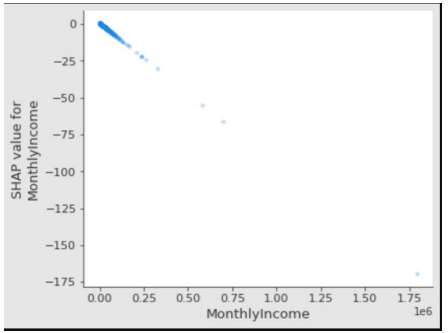
- false-negative rate for the unprivileged groups is 0.16 times the false-negative rate for the privileged groups.
  - The model will be slightly more likely to overestimate the default risk of people with an age smaller than 55 than of people with an age greater than 55

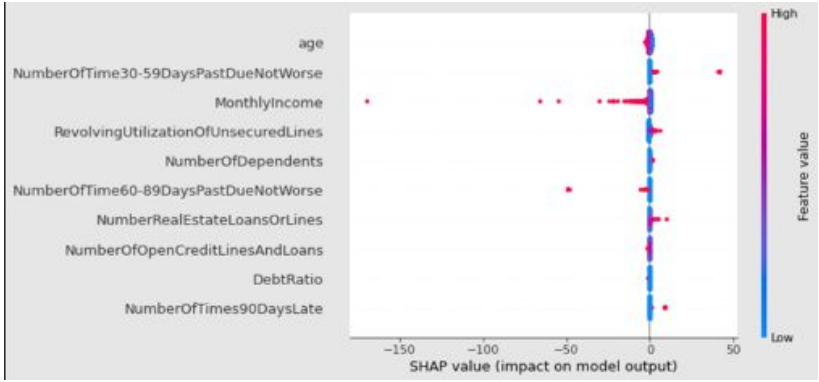
The model is overall fair across different age groups

Transparency of the ADS-SHAP

- Summary plot (features' contribution to the outcome)
- Older age corresponds to lower risk of financial distress.
- Greater income corresponds to lower risk of financial distress.
- Examine specific entries







Conclusion

Appropriate Data

- include many positive features
- does not contains sensitive attributes
- highly correlated
  - need to filter through redundant information
- anonymized
  - potentially vulnerable to linkage attack

Robust, accurate, and fair?

- We analyze its accuracy on people from different age groups and draw the conclusion that the model is overall robust and distributes the positive outcome relatively fairly based on similar FPR and FNR across subgroups, mean difference approaching 0, and disparate impact ratio approaching 1.
- Stakeholders

  Banks may find the optimization of accuracy appropriate for banks by reducing the risk of credit default.
  Young credit applicants will find the optimization of FPR and FNR appropriate such that they will not be treated disparately and denied a loan based on their age Ceteris paribus.

- Future work on data privacy

Recommendations and future improvements

- Since our ADS is fair as it does not contain any sensitive features and has a reasonable score on the fairness metrics , we will say that it will be comfortable to deploy this ADS to the public.

- However, before deploying the ADS to the public sector, we need to ensure that the ADS is robust enough to resist linkage attacks and be protective of people's privacy.
    - We recommend transforming the original dataset into DP synthetic data containing randomly generated data that will prevent others from recognizing each individual.
    - We recommend ADS follow the ethnic principle of respecting individuals and get informed consent from the data provider.

Thank you